Robert Thomson · Christopher Dancy
Ayaz Hyder · Halil Bisgin (Eds.)

LNCS 10899

Social, Cultural, and Behavioral Modeling

11th International Conference, SBP-BRiMS 2018
Washington, DC, USA, July 10–13, 2018
Proceedings

Springer

Lecture Notes in Computer Science 10899

More information about this series at http://www.springer.com/series/7409

Robert Thomson · Christopher Dancy
Ayaz Hyder · Halil Bisgin (Eds.)

Social, Cultural, and Behavioral Modeling

11th International Conference, SBP-BRiMS 2018
Washington, DC, USA, July 10–13, 2018
Proceedings

Springer

Editors

Robert Thomson
United States Military Academy
West Point, NY
USA

Ayaz Hyder
The Ohio State University
Columbus, OH
USA

Christopher Dancy
Bucknell University
Lewisburg, PA
USA

Halil Bisgin
University of Michigan–Flint
Flint, MI
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-93371-9 ISBN 978-3-319-93372-6 (eBook)
https://doi.org/10.1007/978-3-319-93372-6

Library of Congress Control Number: 2018944433

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG
part of Springer Nature
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Improving the human condition requires understanding, forecasting, and impacting sociocultural behavior both in the digital and nondigital world. Increasing amounts of digital data, embedded sensors collecting human information, rapidly changing communication media, changes in legislation concerning digital rights and privacy, spread of 4G technology to third-world countries and so on are creating a new cyber-mediated world where the very precepts of why, when, and how people interact and make decisions are being called into question. For example, Uber took a deep understanding of human behavior vis-à-vis commuting, developed software to support this behavior, ended up saving human time (and so capital) and reducing stress, and thus indirectly created the opportunity for humans with more time and less stress to evolve new behaviors. Scientific and industrial pioneers in this area are relying on both social science and computer science to help make sense of and impact this new frontier. To be successful, a true merger of social science and computer science is needed. Solutions that rely only on the social science or only on the computer science are doomed to failure. For example, Anonymous developed an approach for identifying members of terror groups such as ISIS on the Twitter social media platform using state-of-the-art computational techniques. These accounts were then suspended. This was a purely technical solution. The response was that those individuals with suspended accounts just moved to new platforms, and resurfaced on Twitter under new IDs. In this case, failure to understand basic social behavior resulted in an ineffective solution.

The goal of this conference is to build this new community of social cyber scholars by bringing together and fostering interaction between members of the scientific, corporate, government, and military communities interested in understanding, forecasting, and impacting human sociocultural behavior. It is the charge of this community to build this new science, its theories, methods, and its scientific culture in a way that does not give priority to either social science or computer science, and to embrace change as the cornerstone of the community. Despite decades of work in this area, this new scientific field is still in its infancy. To meet this charge, to move this science to the next level, this community must meet the following three challenges: deep understanding, sociocognitive reasoning, and re-usable computational technology. Fortunately, as the papers in this volume illustrate, this community is poised to answer these challenges. But what does meeting these challenges entail?

Deep understanding refers to the ability to make operational decisions and theoretical arguments on the basis of an empirical-based deep and broad understanding of the complex sociocultural phenomena of interest. Today, although more data are available digitally than ever before, we are still plagued by anecdotal-based arguments. For example, in social media, despite the wealth of information available, most analysts focus on small samples, which are typically biased and cover only a small time period, and use that to explain all events and make future predictions. The analyst finds the magic tweet or the unusual tweeter and uses that to prove their point. Tools that can help the analyst to reason using more data or less biased data are not widely used, are often more complex than the average analyst wants to use or they take more time than the analyst wants to spend to generate results. Not only are more scalable technologies needed, but so too is a better understanding of the biases in the data and ways to overcome them, and a cultural change to not accept anecdotes as evidence.

Sociocognitive reasoning refers to the ability of individuals to make sense of the world and to interact with it in terms of groups and not just individuals. Today most social–behavioral models either focus on (1) strong cognitive models of individuals engaged in tasks and so model a small number of agents with high levels of cognitive accuracy but with little if any social context, or (2) light cognitive models and strong interaction models and so model massive numbers of agents with high levels of social realisms and little cognitive realism. In both cases, as realism is increased in the other dimension the scalability of the models fail, and their predictive accuracy on one of the two dimensions remains low. By contrast, as agent models are built where the agents are not just cognitive by socially cognitive, we find that the scalability increases and the predictive accuracy increases. Not only are agent models with sociocognitive reasoning capabilities needed, but so too is a better understanding of how individuals form and use these social cognitions.

More software solutions that support behavioral representation, modeling, data collection, bias identification, analysis, and visualization support human sociocultural behavioral modeling and prediction than ever before. However, this software is generally just piling up in giant black holes on the Web. Part of the problem is the fallacy of open source; the idea that if you just make code open source others will use it. By contrast, most of the tools and methods available in Git or R are only used by the developer, if that. Reasons for lack of use include lack of documentation, lack of interfaces, lack of interoperability with other tools, difficulty of linking to data, and increased demands on the analyst's time due to a lack of tool-chain and workflow optimization. Part of the problem is the "not-invented here" syndrome. For social scientists and computer scientists alike, it is simply more fun to build a quick and dirty tool for your own use than to study and learn tools built by others. And, part of the problem is the insensitivity of people from one scientific or corporate culture to the reward and demand structures of the other cultures that impact what information can or should be shared and when. A related problem is double standards in sharing, where universities are expected to share and companies are not, but increasingly universities are relying on that intellectual property as a source of funding just like other companies. While common standards and representations would help, a cultural shift from a focus on sharing to a focus on re-use is as or more critical for moving this area to the next scientific level.

In this volume, and in all the work presented at the SBP-BRiMS 2018 conference, you will see suggestions of how to address the challenges just described. SBP-BRiMS 2018 carried on the scholarly tradition of the past conferences out of which it has emerged like a phoenix: the Social Computing, Behavioral-Cultural Modeling, and Prediction (SBP) Conference and the Behavioral Representation in Modeling and Simulation (BRiMS) Society's conference. A total of 85 papers were submitted as regular track submissions. Of these, 18 were accepted as full papers for an acceptance rate of 21.2% and 27 were accepted as short papers for an acceptance rate of 52.9%.

Additionally, there were a large number of papers describing emergent ideas, late-breaking results. This is an international group with papers submitted with authors from many countries.

The conference has a strong multidisciplinary heritage. As the papers in this volume show, people, theories, methods, and data from a wide number of disciplines are represented including computer science, psychology, sociology, communication science, public health, bioinformatics, political science, and organizational science. Numerous types of computational methods are used that include, but not limited to, machine learning, language technology, social network analysis and visualization, agent-based simulation, and statistics.

This exciting program could not have been put together without the hard work of a number of dedicated and forward-thinking researchers serving as the Organizing Committee, listed on the following pages. Members of the Program Committee, the Scholarship Committee, publication, advertising and local arrangements chairs worked tirelessly to put together this event. They were supported by the government sponsors, the area chairs, and the reviewers. We thank them for their efforts on behalf of the community. In addition, we gratefully acknowledge the support of our sponsors – the Army Research Office (W911NF-17-1-0138), the Office of Naval Research (N00014-17-1-2461), and the National Science Foundation (IIS-1523458). Enjoy the proceedings and welcome to the community.

April 2018

Kathleen M. Carley
Nitin Agarwal

Organization

Conference Co-chairs

Kathleen M. Carley	Carnegie Mellon University, USA
Nitin Agarwal	University of Arkansas – Little Rock, USA

Program Co-chairs

Halil Bisgin	University of Michigan-Flint, USA
Christopher Dancy II	Bucknell University, USA
Ayaz Hyder	The Ohio State University, USA
Robert Thomson	United States Military Academy, USA

Advisory Committee

Fahmida N. Chowdhury	National Science Foundation, USA
Rebecca Goolsby	Office of Naval Research, USA
Stephen Marcus	National Institutes of Health, USA
Paul Tandy	Defense Threat Reduction Agency, USA
Edward T. Palazzolo	Army Research Office, USA

Advisory Committee Emeritus

Patricia Mabry	Indiana University, USA
John Lavery	Army Research Office, USA
Tisha Wiley	National Institutes of Health, USA

Scholarship and Sponsorship Committee

Nitin Agarwal	University of Arkansas – Little Rock, USA
Christopher Dancy II	Bucknell University, USA

Industry Sponsorship Committee

Jiliang Tang	Michigan State University, USA

Publicity Chair

Donald Adjeroh	West Virginia University, USA

Web Chair

Kiran Kumar Bandeli	University of Arkansas – Little Rock, USA

Local Area Coordination

David Broniatowski	The George Washington University, USA

Proceedings Chair

Robert Thomson	United States Military Academy, USA

Agenda Chair

Robert Thomson	United States Military Academy, USA

Journal Special Issue Chair

Kathleen M. Carley	Carnegie Mellon University, USA

Tutorial Chair

Kathleen M. Carley	Carnegie Mellon University, USA

Graduate Program Chair

Yu-Ru Lin	University of Pittsburgh, USA

Challenge Problem Committee

Kathleen M. Carley	Carnegie Mellon University, USA
Nitin Agarwal	University of Arkansas – Little Rock, USA
Sumeet Kumar	Massachusetts Institute of Technology, USA
Brandon Oselio	University of Michigan, USA
Justin Sampson	Arizona State University, USA

BRiMS Society Chair

Christopher Dancy II	Bucknell University, USA

SBP Society Chair

Shanchieh (Jay) Yang	Rochester Institute of Technology, USA

BRiMS Steering Committee

Christopher Dancy II	Bucknell University, USA
William G. Kennedy	George Mason University, USA
David Reitter	The Pennsylvania State University, USA
Dan Cassenti	US Army Research Laboratory, USA

SBP Steering Committee

Nitin Agarwal	University of Arkansas – Little Rock, USA
Sun Ki Chai	University of Hawaii, USA
Ariel Greenberg	Johns Hopkins University/Applied Physics Laboratory, USA
Huan Liu	Arizona State University, USA
John Salerno	Exelis
Shanchieh (Jay) Yang	Rochester Institute of Technology, USA

BRiMS Executive Committee

Brad Best	Adaptive Cognitive Systems
Brad Cain	Defense Research and Development, Canada
Daniel N. Cassenti	US Army Research Laboratory, USA
Bruno Emond	National Research Council
Coty Gonzalez	Carnegie Mellon University, USA
Brian Gore	NASA
Kristen Greene	National Institute of Standards and Technology
Jeff Hansberger	US Army Research Laboratory, USA
Tiffany Jastrzembski	Air Force Research Laboratory, USA
Randolph M. Jones	SoarTech
Troy Kelly	US Army Research Laboratory, USA
William G. Kennedy	George Mason University, USA
Christian Lebiere	Carnegie Mellon University, USA
Elizabeth Mezzacappa	Defence Science and Technology Laboratory, UK
Michael Qin	Naval Submarine Medical Research Laboratory, USA
Frank E. Ritter	The Pennsylvania State University, USA
Tracy Sanders	University of Central Florida, USA
Venkat Sastry	University of Cranfield, USA
Barry Silverman	University of Pennsylvania, USA
David Stracuzzi	Sandia National Laboratories, USA
Robert Thomson	Unites States Military Academy, USA
Robert E. Wray	SoarTech

SBP Steering Committee Emeritus

Nathan D. Bos	Johns Hopkins University/Applied Physics Lab, USA
Claudio Cioffi-Revilla	George Mason University, USA
V. S. Subrahmanian	University of Maryland, USA
Dana Nau	University of Maryland, USA

SBP-BRIMS Steering Committee Emeritus

Jeffrey Johnson	University of Florida, USA

Technical Program Committee

Kalin Agrawal
Shah Jamal Alam
Elie Alhajjar
Scott Batson
Jeffrey Bolkhovsky
Lashon Booker
David Broniatowski
Magdalena Bugajska
Jose Cadena
Subhadeep Chakraborty
Rumi Chunara
Andrew Crooks
Peng Dai
Hasan Davulcu
Jana Diesner
Wen Dong
Koji Eguchi
Bruno Emond
William Ferng
Michael Fire
Ariel Greenberg
Kristen Greene
Kyungsik Han
Walter Hill
Shen-Shyang Ho
Tuan-Anh Hoang
Yuheng Hu
Robert Hubal
Terresa Jackson
Aruna Jammalamadaka
Bill Kennedy
Shamanth Kumar
Huan Liu
Yu-Ru Lin
Deryle W. Lonsdale
Stephen Marcus
Venkata Swamy Martha
Elizabeth Mezzacappa
Allen Mclean
Sai Moturu
Keisuke Nakao
Radoslaw Nielek
Kouzou Ohara
Byung Won On
Brandon Oselio
Alexander Outkin
Hemant Purohit
Aryn Pyke
Weicheng Qian
S. S. Ravi
Travis Russell
Amit Saha
Samira Shaikh
Narjes Shojaati
David Stracuzzi
Serpil Tokdemir
Zhijian Wang
Changzhou Wang
Yafei Wang
Xiaofeng Wang
Changzhou Wang
Rik Warren
Elizabeth Whitaker
Paul Whitney
Kevin S. Xu
Xiaoran Yan
Laurence Yang
Yong Yang
Mo Yu
Reza Zafarani
Rifat Zahan
Kang Zhao

Contents

Advances in Sociocultural and Behavioral Process Modeling

Information, Systems, and Network Science

Applications for Health and Well-Being

Military and Intelligence Applications

Cybersecurity

Advances in Sociocultural and Behavioral Process Modeling

Multi-scale Resolution of Cognitive Architectures: *A Paradigm for Simulating Minds and Society*

Mark G. Orr[1](✉), Christian Lebiere[2], Andrea Stocco[3], Peter Pirolli[4],
Bianica Pires[1], and William G. Kennedy[5]

[1] Biocomplexity Institute of Virginia Tech, Blacksburg, USA
morr9@vt.edu
[2] Carnegie Mellon University, Pittsburgh, USA
[3] University of Washington, Seattle, USA
[4] Institute for Human and Machine Cognition, Pensacola, USA
[5] George Mason University, Fairfax, USA

Abstract. We put forth a thesis, the *Resolution Thesis*, that suggests that cognitive science and generative social science are interdependent and should thus be mutually informative. The thesis invokes a paradigm, the reciprocal constraints paradigm, that was designed to leverage the interdependence between the social and cognitive levels of scale for the purpose of building cognitive and social simulations with better resolution. In addition to explaining our thesis, we provide the current research context, a set of issues with the thesis and some parting thoughts to provoke discussion. We see this work as an initial step to motivate both social and cognitive sciences in a new direction, one that represents some unity of purpose and interdependence of theory and methods.

1 Introduction

The degree of overlap between cognitive science and generative social science is small despite a shared interest in human behavior and a reliance on computer simulation. The former focuses, largely, on developing computational and formal accounts of human thought, action, performance and behavior with non-trivial incorporation of neurophysiological principles when warranted. The latter approaches the question of understanding social structure and dynamics using computational and formal accounts that implement simple agents (what we call *sans cognitive*) in social contexts. We submit that the dearth of interdisciplinary

The research is (partially) based upon work supported by the Defense Advanced Research Projects Agency (DARPA), via the Air Force Research Laboratory (AFRL). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the AFRL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 3–15, 2018.
https://doi.org/10.1007/978-3-319-93372-6_1

work between these disciplines does not serve either well. Our central thesis, the *Resolution Thesis*, is this: *the correct resolution of both cognitive and social systems depends on mutual constraints between them in the sense that the dynamics and structure of one system should inform the theoretical nature of the other.* We mean this in the context of theory development and related applications in both cognitive science and generative social science. The method implied by this thesis is what we call the reciprocal constraints paradigm–a bi-directional dependence across levels of scale w.r.t. their respective parameter specifications.[1]

Our thesis implies two claims. First, cognitive models should be able to match and predict the real world dynamics of social systems when embedded in social simulation, and, if not, the cognitive model should be questioned. Second, if an agent-based simulation is not informed by cognitive first principles, it will fail to generalize its account of the dynamics of social system to new situations.

In what follows, we will (1) flesh out the details of the reciprocal constraints paradigm, (2) provide some prior work that is directly relevant to our thesis and puts it in context of recent research, (3) address issues and their potential mitigation, and, (4) close with some brief, but potentially provocative suggestions. We deliberately exercised a narrow focus using the ACT-R cognitive architecture as our vehicle of rhetoric, partly because it reflects our expertise, and partly because this architecture is comparatively well suited for integration of both neural and social constraints. Cognitive architectures, as opposed to any cognitive model, capture what agents, in the scheme of generative social science, are supposed to do–make adaptive decisions that affect the environment.

2 The Reciprocal Constraints Paradigm

Figure 1 captures the core components of the reciprocal constraints paradigm: multiple levels of scale, multiple potentials for model types at each level, and, the constraints among levels. To understand the paradigm, it will be useful to imagine a potential implementation. Consider a modeling problem in which there is a simple social system (e.g., a multi-player repeated economic game). The cognitive model is developed, with some consideration for key neural processes, call this CM [1], and without direct comparison to newly generated individual-level data sources (e.g., running single-subject experiments in pseudo-game like contexts). CM [1] is then implemented in a social network graph that controls information flow (e.g., knowing past decisions of other players) and, given some other parameterizations, a simulation of the multi-player repeated game is conducted; call this SM [1]. Then, SM [1] data is aggregated in some way isomorphic to human data in a similar experimental paradigm and an accuracy/error/confidence metric is computed, call it Constraint [1] to map onto Fig. 1. (Notice, at this point, the only direct comparison to human data was at the social system level.) Constraint [1] would then be used–in an undefined way at this point–to change

[1] Because cognitive systems are sometimes tightly yoked to neurophysiology, we consider three levels as central to our thesis: neurophysiology, cognitive architecture, and social systems.

some aspect of the cognitive model, either directly within the cognitive level or, potentially, through the neurophysiological level. Let's imagine that it makes sense to consider neurophysiological processes as the next step, a step we call Interpret [1] to map onto Fig. 1. Now, a set of targeted neurophysiological measurements are captured by running single-subject experiments in pseudo-game like contexts which yields insight into a potential missing abstraction of neurophysiological process in the cognitive model, which we call Abstraction [1]. The cognitive model is then refactored to incorporate Abstraction [1] and the process is repeated by another simulation using the next generation of the cognitive model CM [2]. Note, this example provides only one of an infinite set of paths; the paths may be consequential to the final model and could include integration of human data at one or more points.

A fundamental part of the paradigm is the acknowledgment that scaling up from the cognitive level to the social level is different, in principle, compared to the scaling up from the neural to cognitive level. The former transition instantiates multiple isomorphic and interdependent cognitive models as a simulated system. The latter, in contrast, abstracts information processing functionalities that are assumed to be interdependent but different in nature (i.e., different functions). This is an important difference in light of what a constraint actually means.

3 Relevant Prior Work

In cognitive science, there are several relevant threads of work that address aspects that are important for the *Resolution Thesis*, e.g., on multi-agent systems [1], computational organizational theory [2], computational social psychology [3]. These efforts, however, were not directly concerned with the *Resolution Thesis.* Instead, these efforts, in the main, attempted to provide both more accurate predictions of social system level behavior and explanations that were grounded in cognitive first principles. In this section, we focus on ACT-R to illustrate efforts to either inform ACT-R from neurophysiology or use implementations of ACT-R as the agent definitions in a social simulation. These efforts, we hope, will illustrate how the state-of-the-art in infusing social simulation with cognition contrasts with the *reciprocal constraint paradigm.* Further, we offer a glimpse into how generative social science has conceptualized the integration of cognitive first principles into the behavior of agents to date.

3.1 The ACT-R Cognitive Architecture

Computational modeling aims to quantitatively capture human cognitive abilities in a principled manner. Cognitive architectures are computational instantiations of unified theories of cognition that specify the structures, representations and mechanisms of the human mind. Cognitive models of any given task can be developed using a cognitive architecture as a principled implementation

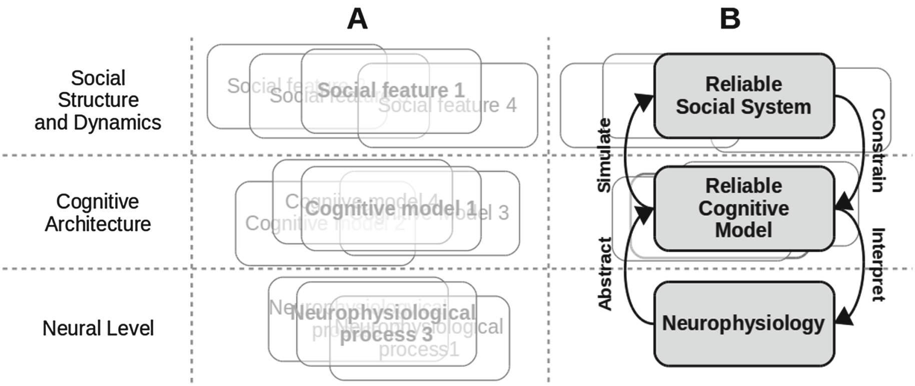


Fig. 1. ***The Architecture and Implementation of the Reciprocal Constraints Paradigm*** Each row represents a level of scale (as labeled in the left-most column). Column A is notational for the degree of variety of potential types of neural processes and cognitive models that could be constructed to capture a phenomenon and the types of features in the social space (e.g., peer-network)–i.e., it captures the feature/model space of a particular implementation. Column B shows the implementation of the reciprocal constraints paradigm; each arrow represents a kind of constraint: *Abstract*–abstraction of neural processes to cognitive processes; *Simulate*–simulating social systems in which humans behavior is defined as a cognitive architecture; *Constrain*–the feedback signal from the accuracy of the social simulation w.r.t. to empirical measurements on human systems; and, *Interpret*–refinement of the selection of neural processes that are implicated in the cognitive model. The former two constraints we call *upward constraints*; the latter are called *downward constraints.* Implementation of the paradigm will require iteration between the feature/model space and the simulation of social and cognitive models. There may be potential for automation of this paradigm once it is well developed.

platform constraining performance to the powers and limitations of human cognition. Cognitive models are not normative but represent Simon's (1991) theory of bounded rationality [4], and can also represent individual differences in knowledge and capacity such as working memory. Cognitive models can be used to generate quantitative predictions in any field of human endeavor.

ACT-R is a highly modular cognitive architecture, composed of a number of modules (e.g., working memory, procedural and declarative memory, perception and action) that operate in parallel asynchronously through capacity-limited buffer interfaces. Each module in turn consists of a number of independent mechanisms, typically including symbolic information processing structures combined with equations that represent specific phenomena and regularities (e.g., power law of practice and forgetting, reinforcement learning). Most notably, the architecture includes a number of learning mechanisms to adapt its processing to the structure of the environment. ACT-R has been applied to model human behavior across a wide range of applications (see ACT-R web site for over a thousand publications), ranging from basic experimental psychology paradigms to language, complex decision making, and rich dynamic task environments. The combination of powerful computational mechanisms and human capacity limitations (e.g., working memory, attention, etc.) provides a principled account of both human information processing capabilities as well as cognitive biases and limitations.

3.2 Neurophysiogical Constraints in ACT-R

The development of ACT-R has been guided and informed, in recent years, by the increased understanding of the computational mechanisms of the brain. For example, independent modules have been associated to specific brain regions and circuits, and this correspondence has been validated multiple times through fMRI experiments. The detailed computations of crucial ACT-R components can also be derived from the neural mechanisms they abstract. For instance, the latency to retrieve declarative information from long-term memory can be derived from the dynamics of the integrate-and-fire neural model [5], and the mechanisms for skill acquisition can be derived from reinforcement learning [5] as well as from the simulation of the large-scale effects of dopamine release in the fronto-striatal circuits [6]. In fact, the modularity of ACT-R permits to easily abstract and integrate lower-level neural principles within the architecture. While this approach does not grant the full flexibility of large-scale neural simulations, it has been repeatedly shown to be very effective in capturing features of human behavior that would otherwise have remained unexplained, while at the same time maintaining the computational parsimony of a cognitive symbolic architecture. For example, implementing the dynamics of memory retrieval permits to capture a variety of decision-making effects and paradoxes, beyond those explained by current mathematical models [7]. The modularity of ACT-R also permits to regulate the degree of fidelity of a module to its biological counterpart, without affecting the entire architecture. As an example, Stocco [8] has shown that the competition between the direct and indirect pathways of the basal ganglia can be captured by splitting production rules into opposing pairs. This procedure captures the cognitive effects of Parkinson's disease, and provides a way to model individual differences in decision-making [8] and cognitive control [9] that are due to individual differences in dopamine receptors in the two pathways. See Fig. 2. This is an example of additional mechanisms that can be added to ACT-R to incorporate further biological details (i.e., the *abstract* constraint in Fig. 1).

3.3 Social Simulation with ACT-R Agents

To study the dynamics of simple systems, work using ACT-R has focused on iterated two-player games, including both adversarial games (e.g., paper-rock-scissors, pitcher-batter in baseball) and social dilemmas allowing both cooperation and competition dynamics such as Prisoner's Dilemma and Chicken Game [10–12]. Even in such simple systems, we have observed the emergence of complex effects such as bifurcations and stochastic resonance [13]. To scale up to

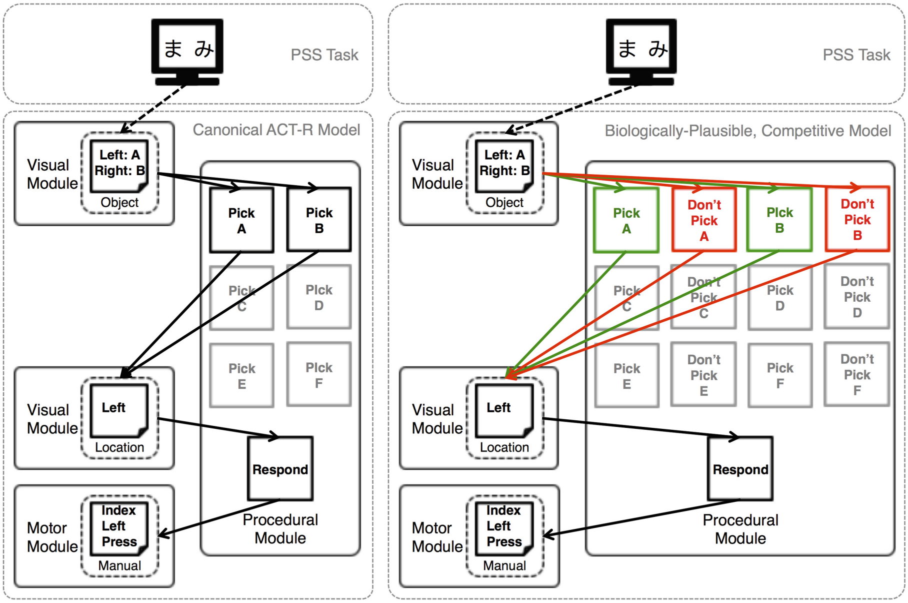


Fig. 2. ***An example (taken from [8] with permission) of how neurobiological constraints can be incorporated in a cognitive architecture.*** The two panels illustrate two alternative ways to implement a forced choice task with six possible options (A through F) in ACT-R. **(Left Panel)** A canonical ACT-R model, in which each option A . . . F is associated with a single, corresponding production rule (Pick A Pick F). In this model, the expected value of the different options is encoded as the expected utility of each production rule. The utility of each rule is learned through reinforcement learning in ACT-R's procedural module, which is associated with the basal ganglia. However, the lack of biological plausibility in ACT-R's procedural module prevents the model from capturing the results of the original study. **(Right Panel)** A biologically-plausible version of the same model, in each of the original production rules is split into two opposite actions (Pick A Pick F and Don't Pick A Don't Pick F), whose utilities are learned separately. This new version abstracts the competition between the direct and indirect pathways of the basal ganglia circuit. When equipped with this biologically-plausible version of production rules, the model can successfully reproduce the results in the neuropsychological literature, as well as capture individual differences in genetics [8] and even correctly predict new findings [9].

more complex yet regular systems, we have modeled the emergence of group consensus and choice differentiation in networks of a few dozen nodes on tasks such as consensus voting and map coloring, respectively, and observed phenomena such as sensitivity to network rewiring parameters [14]. To study complex cognition in complex systems, we have designed and implemented an information foraging task called the Geogame that involves cooperative and competitive problem solving and have observed effects including sensitivity to network topology and tradeoffs between perceptual and memory strategies [15]. Clearly, this work represents well the *simulate* constraint in Fig. 1.

A common pattern in models of social interaction using ACT-R has been to ground agent decisions in previous experiences, whether explicitly in the form of memories or implicitly by reinforcement of existing strategies, as mentioned in the previous section. We will focus here on an example using the former approach, because it has been both more common and more flexible. Models of adversarial interaction usually involve a core capability of detecting patterns in the opponent behavior and exploiting them until they disappear. For instance, playing paper rock scissors involve exploiting the human limitation in generating purely random behavior (the standard game theory solution) by detecting statistical patterns in move sequences. An expectation of an opponent's next move can be generated by matching his most recent moves against previous sequences using statistical memory mechanisms. Once a pattern is being exploited, the opponent is likely to move away from it and in turn exhibit new ones, requiring a cognitive system that constantly unlearns previous patterns and learns emerging ones, rather than traditional machine learning systems that are training on a fixed set of inputs and then frozen. In that sense, social simulation is the ultimate requirement for online learning: unlike physical environments which change relatively slowly and can be mastered in a relatively static way, social interactions (especially competitive and adversarial interactions), as they involve other cognitive entities, are endlessly evolving and require constant learning and adaptivity.

3.4 Comparison to the Generative Social Science Approach

Generative social simulation has historically been concerned with the simulation of interacting agents according to simple behavioral rules. We can often equate the outcome behavior of agents to a simple binary action (e.g., you either riot or don't riot) and the behavioral rules that produce this outcome to simple mathematical and logical formulations (e.g., if/else statements, threshold values). We are in debt to the many classic models that made computational social science the field it is today [16–18]. However, there has been some acknowledgment that to gain further insight into social systems, we need to decompose behavior into its underlying cognitive, emotional, and social (interactions) processes. With this, we are beginning to see a slight shift to developing models with more complex agents [19].

In this vein, an approach that has gained some traction is the use of conceptual frameworks that integrate the varied components of agent decision-making processes [20–23]. Such frameworks include BDI (Beliefs, Desires, and Intentions) and PECS (Physical conditions, Emotional state, Cognitive capabilities, and Social status) [24]. In the BDI framework, beliefs are said to be the individual's knowledge about the environment, desires contain information about the priorities and payoffs associated with the current objective, and intentions represent the chosen course of action [25]. BDI agents use a decision tree process which relies on payoff and utility maximizing functions to select goals and to determine the optimal action sequence for which to achieve those goals. The focus on optimality, however, may pose limits on its ability to model the boundedly rational agent and has been criticized for being too restrictive [25]. PECS views agents as a psychosomatic unit with cognitive capabilities residing in a social environment [26]. The PECS framework is flexible due to its ability to model a full spectrum of behaviors, from simple stimulus-response behaviors to more intricate reflective behaviors, which requires a construction of self that necessitates the agent be fully aware of its internal model. By example, Pires and Crooks [23] used the PECS framework to guide implementation of the underlying processes behind the decision to riot, applying theory from social psychology to create the agent's internal model and to simulate social influence processes that heightened certain emotions and drove the agent's towards certain actions. These frameworks, while helpful for guiding implementation, are not to be considered substitutes for cognitive architectures such as ACT-R. They can, however, provide a meta-framework (sometimes called a macro-architecture) to organize knowledge and skill content in respect to a cognitive architecture (e.g., [27]).

Cognitive architectures and meta-frameworks are fundamentally complementary [28]. Cognitive architectures precisely specify the basic cognitive acts that can be used to compose complex models in a bottom up approach, but provide few constraints to guide those complex structures. Meta-frameworks provide a top down methodology to decompose complex tasks into simpler ones and structure the knowledge required, but do not include a principled grounding for that process. The combination of the two approaches can be achieved in a number of different ways. One approach is to develop integrated environments allowing modelers to flexibly leverage the two methodologies in a way that is best suited to the specific requirements of each application [29]. An alternative is to provide high-level patterns and abstractions that can be formally compiled into cognitive models in a target cognitive architecture [30].

4 Issues and Their Mitigation

4.1 Downward Constraints

Social to Cognitive. This issue was laid out plain by Allen Newell about three decades ago [31] in reference to the *social band* (bands in geometric time of $>$ 10^4 seconds that represent organizational behavior and other social systems). Newell, thinking in terms of the strength of a system's levels, hypothesized that social bands should be characterized as having weak strength, and therefore may not be computing much at all, in a systematic way. If Newell's surmises are correct, then constraining cognitive architectures from the social band makes little sense. Anderson's Relevance Thesis [32], put forth about a decade later, does not address the operation of social systems in terms of constraining cognitive models; his thesis is more focused on the degree to which understanding lower bands, especially the cognitive (10^{-1} to 10^1 time scale), are implicated in qualities of higher bands, e.g., educational outcomes. So, from the cognitive perspective, there might not be much signal from the social band that could serve as a useful constraint on cognitive architectures.

However, there are potential approaches towards mitigation of this problem, Newell's thesis notwithstanding. Online social communities often exhibit emergent empirical regularities. For instance, the World Wide Web exhibits many regularities including the small world organization of link structure and the distribution of the lengths of browsing paths that users exhibit. The latter has been called the "Law of Surfing"?. Many of these regularities have been modeled at the social level using variants of statistical mechanics. The Law of Surfing [33] observes that the frequency distribution of path lengths (number of Web pages visited) is well fit by an Inverse Gaussian Distribution, that has a long positive tail. The key insight at the social level is that a Web surfer can be viewed as moving around in a kind of space analogous to the Brownian motion of a small particle on a liquid surface. In the case of the Web surfer, the movement is in the dimension of expected utility that will be received (or not) when visiting a Web page, where the expected utility from continuing on to the next page is stochastically related to the expected utility of the current page, and the Web surfer continues until a threshold expected utility is reached. This is modeled as a stochastic Wiener process. But, the Law of Surfing can also be predicted from Monte Carlo simulations with ACT-R agents [34]. In contrast to the stochastic social models, these finer-grained ACT-R agents can make predictions for specific Web tasks at specific Web sites, which can be used to predict and engineer improvements [35]. However, the emergence of the Law of Surfing from the ACT-R agent simulations is seen as constraint on the cognitive models.

In short, the social band, at least in some domains, does have structure that could constraint cognitive modeling efforts. A question that remains is to what degree will it be possible to develop general methods across the varieties of social domains for the purpose of constraining cognitive models.

Cognitive to Neurophysiology. The downward "Interpret"? arrow in Fig. 1 could seem paradoxical, given that the underlying neural level is often taken as the ground truth of the entire system. Neurophysiological findings, however, are often only imperfectly understood. For instance, the existence of basal ganglia projections outside of the frontal lobe was considered impossible for a long time until recently [36]. Even when our grasp of neurophysiology is solid, cognitive architectures can be helpful in providing a functional interpretation to existing data by focusing on the computational integration of different circuits, that is, answering the question of "what does this circuit do"?. The most famous example in this sense is the interpretation of the activity of dopamine neurons in terms of reward prediction error signals in reinforcement learning (RL)[37]–an interpretation that borrowed from a decades-old AI theory (temporal difference learning: [38]) to solve decades of seemingly inconsistent empirical findings on the role of dopamine [39,40]. Incidentally, this example perfectly illustrates how the Interpretation is further aided by the use of a comprehensive architecture on an agent's behavior, such as that provided by RL agents. In our case, the adoption of a single architecture (such as ACT-R) to create multiple models provides the unifying framework to interpret neurophysiological data. The fact that the activity of the same neuronal process must be interpreted in the same way across multiple models of different tasks provides additional constraints to maintain the interpretation consistent.

4.2 Upward Constraints

Parsimony and Generative Social Science. By uncovering some new relationship or testing some stylized hypothesis of social phenomena many classic agent-based models (e.g., [16,18]) have demonstrated the value of modeling simple (*sans cognitive*) agents. For instance, Reynolds [41] illustrates how three simple rules of behaviors can result in the emergence of the collective behavior of a flock of birds – what looks like the highly coordinated actions of a "leader" is actually the result of three simple rules.[2] These models and many others in the computational social sciences adhere to parsimony, or keeping the model simple such that the model has just enough of right features and no more, as a main guiding principle [42]. Arguments for this approach stress the intuitive and interpretive appeal of such models [42,43]. The purpose of the model may also dictate that the model be parsimonious (e.g., [41]). In short, parsimony in respect to simple agents has served well as a strategy in generative social science. It is natural, then, to ask if cognitive modeling breaks with this notion of parsimony in modeling social systems.

We think the issue of parsimony in generative social science does not imply anything particular about the use of cognitive architectures in social simulations. Parsimony implies that model simplicity is considered in conjunction with how well a model matches empirical findings. Thus, the issue of whether to include cognitive agents, as defined in the reciprocal constraints paradigm, is largely an empirical issue. We offer that cognitive constraints may provide the right model and thus improve the degree to which a social simulation matches empirical findings. Moreover, because cognitive models inherit mechanistic constraints from cognitive architectures, they might actually end up being more parsimonious than agent-based models without such constraints.

4.3 Mere Parameter Optimization?

To deal with the challenges of scaling up cognitive models beyond the scale of tasks in the cognitive band (seconds to minutes) to tasks in the social band (weeks to months), Reitter and Lebiere [44] formulated a methodology called *accountable modeling.* That approach is not only a technical solution to scaling up the cognitive architecture but also a scientific commitment to an approach

[2] ABMs, however, can range in abstraction, from the stylized models just described to empirically-driven models; although the latter in no way implies incorporation of cognitive constraints.

that explicitly states which aspects of the model are constrained by the architecture and which are free parameters to be estimated from data. This commitment helps determine which aspects of the social-scale simulation reflect the cognitive mechanisms and can be assumed to generalize, and which have been parameterized to reflect aspects of the situation not constrained by first principles, and thus will need to be estimated from data in new situations. Such an approach actually results in simpler, more transparent models that are explicit about their parameters rather than trying to camouflage them under a mechanistic veneer.

5 Closing Thoughts

Crossing levels of scale or analysis inevitably takes one near to deep scientific issues that echo notions pointed out 50+ years ago in Simon's "Architecture of Complexity" paper [45] (see also [46] for similar early example). Our thesis goes counter to Simon's notion of near decomposability in that it puts social structure and dynamics in the realm of convergent evidence for a cognitive theory. In this spirit, we will leave the reader with one final comment.

We see social systems as distributed and symbolic. Thus, insight into them and predictions about them should come through a distributed symbolic system–i.e., a social simulation of interacting cognitive architectures. This argument is not meant to imply that sub-symbolic processes are not part of human information processing, but only to mean that social interactions operate via symbols. For the purposes of simulating social systems, observed social structure and dynamics should be generated from the first principles of interactive artificial symbol systems.

References

1. Sun, R.: Cognition and Multi-agent Interaction: From Cognitive Modeling to Social Simulation. Cambridge University Press, Cambridge (2006)
2. Prietula, M., Carley, K., Gasser, L.: Simulating Organizations: Computational Models of Institutions and Groups, vol. 1. The MIT Press, Cambridge (1998)
3. Vallacher, R.R., Read, S.J., Nowak, A.: Computational Social Psychology. Routledge, Abingdon (2017)
4. Simon, H.A.: Bounded rationality and organizational learning. Organ. Sci. **2**(1), 125–134 (1991)
5. Anderson, J.R.: How can the Human Mind Occur in the Physical Universe?. Oxford University Press, Oxford (2007)
6. Stocco, A., Lebiere, C., Anderson, J.R.: Conditional routing of information to the cortex: a model of the basal ganglia's role in cognitive coordination. Psychol. Rev. **117**(2), 541–574 (2010)
7. Gonzalez, C., Lerch, F.J., Lebiere, C.: Instance-based learning in dynamic decision making. Cognit. Sci. **27**(4), 591–635 (2003)
8. Stocco, A.: A biologically plausible action selection system for cognitive architectures: implications of basal ganglia anatomy for learning and decision-making models. Cognit. Sci. **42**, 457–490 (2018)

9. Stocco, A., Murray, N.L., Yamasaki, B.L., Renno, T.J., Nguyen, J., Prat, C.S.: Individual differences in the simon effect are underpinned by differences in the competitive dynamics in the basal ganglia: An experimental verification and a computational model. Cognition **164**, 31–45 (2017)
10. West, R.L., Lebiere, C.: Simple games as dynamic, coupled systems: randomness and other emergent properties. Cognit. Syst. Res. **1**(4), 221–239 (2001)
11. Lebiere, C., Gray, R., Salvucci, D., West, R.: Choice and learning under uncertainty: a case study in baseball batting. In: Proceedings of the 25th Annual Meeting of the Cognitive Science Society, pp. 704–709. Erlbaum, Mahwah (2003)
12. Lebiere, C., Wallach, D., West, R.: A memory-based account of the prisoner's dilemma and other 2x2 games. In: Proceedings of International Conference on Cognitive Modeling, pp. 185–193. Universal Press, NL (2000)
13. West, R.L., Stewart, T.C., Lebiere, C., Chandrasekharan, S.: Stochastic resonance in human cognition: Act-r vs. game theory, associative neural networks, recursive neural networks, q-learning, and humans. In: Proceedings of the 27th Annual Conference of the Cognitive Science Society, pp. 2353–2358. Lawrence Erlbaum Associates, Mahwah (2005)
14. Romero, O., Lebiere, C.: Simulating network behavioral dynamics by using a multi-agent approach driven by act-r cognitive architecture. In: Proceedings of the Behavior Representation in Modeling and Simulation Conference (2014)
15. Reitter, D., Lebiere, C.: Social cognition: memory decay and adaptive information filtering for robust information maintenance. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI, pp. 242–248 (2012)
16. Schelling, T.C.: Models of segregation. Am. Econ. Rev. **59**(2), 488–493 (1969)
17. Axelrod, R., et al.: A model of the emergence of new political actors. In: Artificial societies The Computer Simulation of Social Life, pp. 19–39 (1995)
18. Epstein, J.M.: Modeling civil violence: an agent-based computational approach. Proc. Natl. Acad. Sci. **99**(suppl 3), 7243–7250 (2002)
19. Epstein, J.M.: Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science. Princeton University Press, Princeton (2014)
20. Caillou, P., Gaudou, B., Grignard, A., Truong, C.Q., Taillandier, P.: A simple-to-use BDI architecture for agent-based modeling and simulation. In: Jager, W., Verbrugge, R., Flache, A., de Roo, G., Hoogduin, L., Hemelrijk, C. (eds.) Advances in Social Simulation 2015. AISC, vol. 528, pp. 15–28. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-47253-9_2
21. Sakellariou, I., Kefalas, P., Stamatopoulou, I.: Enhancing netlogo to simulate BDI communicating agents. In: Darzentas, J., Vouros, G.A., Vosinakis, S., Arnellos, A. (eds.) SETN 2008. LNCS (LNAI), vol. 5138, pp. 263–275. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-87881-0_24
22. Malleson, N., See, L., Evans, A., Heppenstall, A.: Implementing comprehensive offender behaviour in a realistic agent-based model of burglary. Simulation **88**(1), 50–71 (2012)
23. Pires, B., Crooks, A.T.: Modeling the emergence of riots: a geosimulation approach. Comput. Environ. Urban Syst. **61**, 66–80 (2017)
24. Kennedy, W.G.: Modelling human behaviour in agent-based models. In: Heppenstall, A., Crooks, A., See, L., Batty, M. (eds.) Agent-Based Models of Geographical Systems, pp. 167–179. Springer, Heidelberg (2012). https://doi.org/10.1007/978-90-481-8927-4_9
25. Rao, A.S., Georgeff, M.P., et al.: BDI agents: from theory to practice. In: ICMAS, vol. 95, pp. 312–319 (1995)

26. Schmidt, B.: The modelling of human behaviour: The PECS reference models. SCS-Europe BVBA (2000)
27. West, R., Nagy, N., Karimi, F., Dudzik, K.: Detecting macro cognitive influences in micro cognition: using micro strategies to evaluate the SGOMS macro architecture as implemented in ACT-R. In: Proceedings of the 15th International Conference on Cognitive Modeling, pp. 235–236 (2017)
28. Lebiere, C., Best, B.J.: From microcognition to macrocognition: architectural support for adversarial behavior. J. Cognit. Eng. Decis. Mak. **3**(2), 176–193 (2009)
29. Lebiere, C., Archer, R., Best, B., Schunk, D.: Modeling pilot performance with an integrated task network and cognitive architecture approach. Hum. Perform. Model. Aviat. (2008)
30. Ritter, F., Haynes, S.R., Cohen, M., Howes, A., John, B., Best, B., Lebiere, C., Jones, R.M., Crossman, J., Lewis, R.L., St. Amant, R., McBride, S.P., Urbas, L., Leuchter, S., Vera, A.: High-level behavior representation languages revisited. In: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, AAAI, pp. 242–248 (2012)
31. Newell, A.: Unified Theories of Cognition. Harvard University Press, Cambridge (1990)
32. Anderson, J.R.: Spanning seven orders of magnitude: a challenge for cognitive modeling. Cognit. Sci. **26**(1), 85–112 (2002)
33. Huberman, B.A., Pirolli, P., Pitkow, J.E., Lukose, R.M.: Strong regularities in world wide web surfing. Science **280**(5360), 95–97 (1998)
34. Fu, W.T., Pirolli, P.: Snif-act: a model of user navigation on the world wide web. Hum. Comput. Interact. **22**(4), 355–412 (2007)
35. Chi, E.H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., Cousins, S.: The bloodhound project: automating discovery of web usability issues using the infoscent simulator. In: ACM Conference on Human Factors in Computing Systems, CHI Letters, vol. 5, no. 1, pp. 505–512 (2003)
36. Middleton, F.A., Strick, P.L.: The temporal lobe is a target of output from the basal ganglia. Proc. Natl. Acad. Sci. **93**(16), 8683–8687 (1996)
37. Schultz, W., Dayan, P., Montague, P.R.: A neural substrate of prediction and reward. Science **275**(5306), 1593–1599 (1997)
38. Sutton, R.S.: Learning to predict by the methods of temporal differences. Mach. Learn. **3**(1), 9–44 (1988)
39. Bunney, B.S., Chiodo, L.A., Grace, A.A.: Midbrain dopamine system electrophysiological functioning: a review and new hypothesis. Synapse **9**(2), 79–94 (1991)
40. Schultz, W.: Getting formal with dopamine and reward. Neuron **36**(2), 241–263 (2002)
41. Reynolds, C.W.: Flocks, herds and schools: a distributed behavioral model. ACM SIGGRAPH Comput. Graph. **21**(4), 25–34 (1987)
42. Miller, J.H., Page, S.E.: Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton University Press, Princeton (2009)
43. Gigerenzer, G., Todd, P.M., ABC Research Group, et al.: Simple Heuristics That Make Us Smart. Oxford University Press, Oxford (1999)
44. Reitter, D., Lebiere, C.: Accountable modeling in ACT-UP, a scalable, rapid-prototyping ACT-R implementation. In: Proceedings of the 2010 International Conference on Cognitive Modeling (2010)
45. Simon, H.A.: The architecture of complexity. Proc. Am. Philos. Soc. **106**(6), 467–482 (1962)
46. Anderson, P.W.: More is different: broken symmetry and the nature of the hierarchical structure of science. Science **177**(4047), 393–396 (1972)

Detecting Betrayers in Online Environments Using Active Indicators

Paola Rizzo[1(✉)], Chaima Jemmali[1], Alice Leung[2],
Karen Haigh[2], and Magy Seif El-Nasr[1]

[1] Northeastern University, 360 Huntington Avenue, Boston, MA 02115, USA
{p.rizzo,m.seifel-nasr}@northeastern.edu, jemmali.c@husky.neu.edu
[2] Raytheon BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA
{alice.leung,karen.haigh}@raytheon.com

Abstract. Research into betrayal ranges from case studies of real-world betrayers to controlled laboratory experiments. However, the capability of reliably detecting individuals who have previously betrayed through an analysis of their ongoing behavior (after the act of betrayal) has not been studied. To this aim, we propose a novel method composed of a game and several manipulations to stimulate and heighten emotions related to betrayal. We discuss the results of using this game and the manipulations as a mechanism to spot betrayers, with the goal of identifying important manipulations that can be used in future studies to detect betrayers in real-world contexts. In this paper, we discuss the methods and results of modeling the collected game data, which include behavioral logs, to identify betrayers. We used several analysis methods based both on psychology-based hypotheses as well as machine learning techniques. Results show that stimuli that target engagement, persistence, feedback to teammates, and team trust produce behaviors that can contribute to distinguishing betrayers from non-betrayers.

Keywords: Betrayal · Games as experimental methods · Deception
Espionage

1 Introduction

Betrayal of one's group is a phenomenon that occurs in many contexts and organizations [1–4], causing significant economic impacts. According to the Ponemon Institute [5], 874 insider incidents occurred in 2016 in the US, and of those 22% were criminal, costing USD 4.3 million. Therefore, it is important to develop techniques to identify persons who have engaged in such acts. While previous research investigated detecting betrayals within the act of betraying through anomaly detection or other methods (e.g., [3,6,7]), research detecting betrayers after the fact is sparse.

In this paper, we address this topic. In particular, we work with insider threat experts and psychologists who assume that persons who betray their

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 16–27, 2018.
https://doi.org/10.1007/978-3-319-93372-6_2

team or organization have emotional, logical thinking and habitual behaviors, similar to those discussed in [8,9], that are significantly different from those of people who do not betray their teams. We postulate that such behaviors can be evoked through stimuli in the environment producing a distinct fingerprint that is detectable by machine learning or statistical techniques.

In this paper, we focus on the emotional aspects, and expect that the individuals who have betrayed will feel guilty, anxious, trapped, and distant from the group [10]. Consequently, we expect betrayers to have less identification and trust with their teammates, and to exert less focus and diligence on their tasks, compared to other subjects. To investigate this hypothesis, we developed a novel methodological approach composed of multiple techniques. First, we implemented an online social game that allows participants to betray their group by sharing information with a competitor group. Second, in order to make our game a controllable environment with embedded stimuli that can cause subjects to behave in certain ways that can be detectable of malicious intent, we used a technique similar to Sasaki's work [7], according to whom psychological triggers that heighten anxiety in malicious insiders cause them to carry out specific behaviors of deleting evidence and stopping further malicious activities. For instance, a stimulus suggesting that file-searching behaviors may be under surveillance is likely to be ignored by a normal subject engaged in work-related searches, but may cause a malicious subject engaged in espionage to cease certain activities [11]. We developed 13 psychological stimuli called "Active Indicators" (AIs), designed to evoke behaviors that can distinguish betrayers in an online environment. A few of these stimuli, rather than being obtrusive, are integral parts of the background activities, such as stimuli embedded in team text chat, opportunities to react to prompts and cues during the game.

Our work provides the following contributions. First, it presents a novel methodological approach to investigate a set of manipulations to detect betrayers behaviorally after the act of betrayal. The method is based on a game that embeds AIs developed according to previous research. Lessons about the design of the game as well as the utility of previous research can provide a good step for researchers interested in studying this topic within other contexts. Second, we discuss the AIs and the resulting patterns of behaviors and their power in detecting betrayers. We found AIs that target engagement, persistence, feedback to teammates and team trust to be among the most significant and further work is needed to validate these results in real-world environments.

2 Related Work

Computer-based insider threats have been a subject of study for years. Some works focus on anomaly detection, i.e. automated ways using machine learning to distinguish suspicious activities from regular ones [12,13], while other works investigate the utility of eliciting spying behaviors by means of "honeypots", custom-built information system resources (e.g., special files) that can attract and reveal potential insiders [14]. However, none of these works detect insider threats after the fact.

Some works (e.g., [3,6]) use psycho-social and behavioral indicators as antecedents of insider threat activities to assess the chances that an individual will perform specific behaviors. Sasaki [7] assumes that malicious insiders are anxious about their identity being revealed and thus psychological triggers should heighten their anxiety and cause-specific behaviors that will reveal them. We share his hypothesis, however, we focused on emotional stimuli beyond anxiety, including guilt, distance from the group, feeling trapped. We also looked at behavioral patterns other than those concerning insider threats, e.g., willingness to carry out extra work and identification with the team.

Other researchers focus on deceptive communication, looking at nonverbal behaviors such as facial expressions [15], or at linguistic patterns in chats. For instance, Niculae et al. ([16]) studied dyadic communication in an online game where players break alliances through betrayal, but they focused on linguistic cues that foretell betrayal rather than communication patterns of betrayers. Ho and Warkentin [17] developed a game to examine the trustworthiness of spies as measured by teammates engaging in computer-mediated communication. Additionally, Ho et al. [18] used Support Vector Machines to classify deceivers and non-deceivers based on cues in chat data, and found that cues related to time-lag, social attitude and negation in text can discriminate between deceivers and non-deceivers. While such research is relevant, our work uses behavioral measures rather than relying on human-perceived trustworthiness or the contents of chat data.

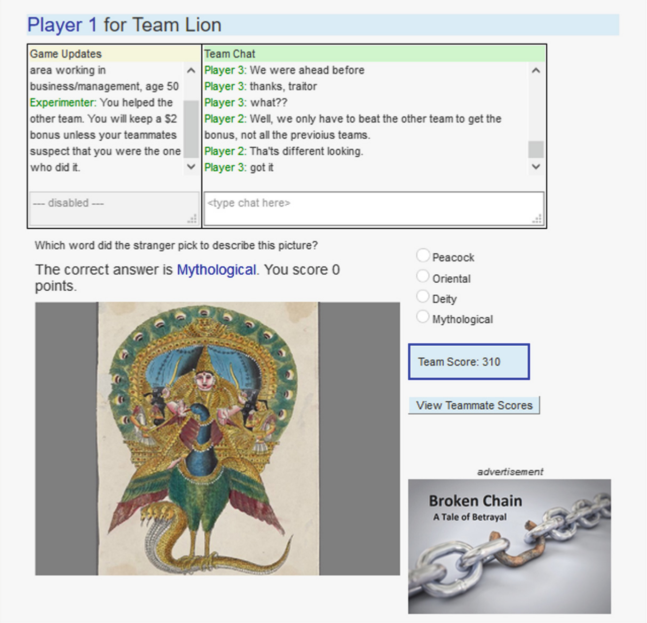
Several works show that deceivers may appear more submissive than truth-tellers when their primary goal is to evade detection (e.g., [19]). However, other research shows that this pattern is reversed when deceivers need to persuade others of their credibility, and thus tend to argue aggressively while simultaneously trying to avoid being detected, a behavior called persuasive deception [20]. In such a case, deceivers may display more dominance, using verbal and non-verbal communication that makes them appear confident [20]. Also, the style of deception can change according to whether the recipient is acquiescent or suspicious [21]. As discussed later in the paper, the textual communicative behavior of betrayers in our game is somewhat similar to that of persuasive deceivers.

3 The Game: ESP

We designed a simple guessing game (see Fig. 1), inspired by Von Ahn and Dabbish [22], lasting about 50 min, where a team plays against another. The goal of the game is to guess the gender, age, location and occupation of a stranger, based on accumulated information about the stranger's reaction to a series of pictures. Each team selects a stranger to be guessed by the opponent team, and then earns points when its members correctly answer questions about its own target stranger. Teammates collaborate to find the right answer by communicating through chat. A game session lasts five rounds, each including 3 pictures of art and 2 questions per picture (first "Which word did the stranger pick to describe this picture?", and then "Did they like the picture?" or "What was their favorite thing about the picture?"). After each round, the scores of each team are revealed. After the last round, 4 high-point value questions about the stranger's demographics, and the final scores of each team are revealed. We let participants play with pre-scripted bot team members, against an imaginary opponent team, to maintain control and comparability across teams (the automated nature of team members and the opponent team was not disclosed to the subjects).

4 Experimental Manipulations and AIs

We had two experimental conditions: the control group played the game with no opportunity to betray their team, while the experimental group was given a message (shown on screen at the end of the first round of the game) asking them if they would secretly pass information to the opponent team about the target stranger and receive a $2 bonus payment in return. It was then up to them to betray their team or not by answering "yes" or "no". After the choice was made, the game announced: "One member of your team was offered money to tell the



Player 1 for Team Lion

Game Updates

area working in business/management, age 50

Experimenter: You helped the other team. You will keep a $2 bonus unless your teammates suspect that you were the one who did it.

--- disabled ---

Team Chat

Player 3: We were ahead before

Player 3: thanks, traitor

Player 3: what??

Player 2: Well, we only have to beat the other team to get the bonus, not all the previoius teams.

Player 2: Tha'ts different looking.

Player 3: got it

<type chat here>

Which word did the stranger pick to describe this picture?

The correct answer is Mythological. You score 0 points.

○ Peacock
○ Oriental
○ Deity
○ Mythological

Team Score: 310

Fig. 1. A screenshot of the ESP game, comprising a chat window (top), a large window where the picture is displayed (bottom left), the question about the picture (below the chat window), the team's score (below the answer options), and a window containing a variable picture (bottom right) used as a small advertisement area for priming and psychological stimuli.

other team the gender of the stranger you picked." In case of betrayal, the text continued with "That player accepted the offer, but they will lose the money if the rest of the team suspects them. Later, the team will vote on who sold the information", otherwise "That player declined the offer and stayed loyal to your team". Hence, we had 3 groups of subjects: controls, betrayers, and decliners.

The game was designed to evoke anxiety and guilt, by showing (a) the negative consequences of discovery, where participants were told they would lose the bonus if their teammates suspected they were the betrayer, and (b) the negative impact of betrayal on the player's teammates, by letting the latter react strongly to the announcement that someone had betrayed the team and produce negative comments about the betrayal during the rest of the game.

At the end of the experiment, we sent participants a debriefing statement that included: "We used this experiment to understand what happens when people put their individual gain before their team's. You were all given the maximum bonus as a reward for your participation". The whole experimental protocol was approved by the IRB.

Based on betrayers' expected behaviors, we developed several AIs (Active Indicators), and corresponding measurements, grouped according to the affective aspect targeted. Our goal is to examine if such AIs when embedded in the environment produce behaviors that can easily discriminate betrayers from the two other groups (controls and decliners). Many AIs were mapped onto several stimuli or "prompts" to provide repeated measures of the same AI while avoiding too much obvious repetition during the experiment. Furthermore, some of the stimuli were embedded in the game rather than being introduced as additional events (for instance, this was the case for AI-Timer and AI-Disengage listed below).

The targeted affective states and corresponding AI stimuli and expected behaviors are as follows:

- **Reluctance to be identified with the group, as a result of emotional distance**
  - *AI-Criticism-of-Group*: Exposure to criticism of the group - *Example stimulus*: presentation of negative information about the team - *Expected behavior*: agreeing with criticism or declining to defend the team from criticism
  - *AI-Assessment-of-Group*: Ask for group assessment - *Example stimulus*: game poll asking the participant to "Please share your impressions about your team" - *Expected behavior*: negative and neutral assessment of the team
- **Less regard and trust for those betrayed, as a result of projection of own lack of trustworthiness and emotional distance**
  - *AI-Peer-Feedback*: Present opportunity to get/give peer feedback - *Example stimulus*: game message asking "Would you like to get feedback from a teammate about your participation? (Y/N)" - *Expected behavior*: "no"
  - *AI-Check-Peer-Performance*: Present opportunity to check peer performance - *Example stimulus*: a button on that can pop up information

  about each team member's contribution to the team score - *Expected behavior*: click on the button
  - *AI-Peer-Trustworthiness*: Ask for an assessment of peer trustworthiness - *Example stimulus*: game message asking "How trustworthy are your teammates?" - *Expected behavior*: low assessment of trustworthiness of teammates
  - *AI-Follow-group*: Provide opportunity to follow group advice - *Stimuli*: presentation of teammate advice throughout the entire task - *Expected behavior*: choose answers that disagree with teammate advice
- **Lack of persistence, as a result of being trapped**
  - *AI-Slack-Off*: Present opportunity to slack off on work - *Stimuli*: opportunities to contribute to team success through chat - *Expected behavior*: low participation in team discussion
- **Impulsive risk taking as a result of anxiety about discovery**
  - *AI-Wager*: Present opportunity to wager - *Stimulus*: an opportunity to wager on an outcome only marginally under the participant's control - *Expected behavior*: decision to wager
- **Does not devote full attention to job, as a result of anxiety about discovery**
  - *AI-Timer*: Present a time-limited window to complete routine work - *Stimuli*: presentation of deadlines for submitting work, as opportunities to submit answers before a timer runs out - *Expected behavior*: not submitting an answer before a timer runs out
  - *AI-Extra-Info*: Present opportunity to gather information to improve probability of success on a task - *Example Stimulus*: presentation of an opportunity to read an extra essay about the stranger - *Expected behavior*: refusal to read the extra essay
  - *AI-Shorter-Game*: Present choice between a smaller task for less reward or a larger task for more reward - *Example Stimulus*: presentation of the poll "Would you prefer to play a longer game (with the opportunity to earn more points) or a shorter game?" - *Expected behavior*: opt for a shorter game
  - *AI-Disengage*: Present opportunities to disengage from task - *Stimuli*: presentation of non-obtrusive task activities throughout the entire game (team text chat, opportunities to react to prompts and cues during the game, and other game content) - *Expected behaviors*: longer time taken to react to game prompts, higher number of times a participant neglects to respond to a game prompt, higher frequency and duration of non-game browser window activity
  - *AI-Cognitive-Challenge*: Present a cognitive challenge - *Stimulus*: a quick test of short memory recall - *Expected behavior*: failure to correctly recall any sequence, caused by not engaging with the test.

4.1 Subjects

We recruited a total of 348 subjects from Amazon Mechanical Turk, with the requirement that participants need to be US residents: 52% males, 48% females, and 88% who attended college; the age mode was 25–29 years, with a frequency of 25%. The compensation was of \$5 + \$2 bonus, set on the basis of several tests aimed at finding a good balance between the percentage of subjects who accepted to betray their team and the percentage of those who did not. We excluded 115 subjects from the analyses because they did not answer the post-game survey, or because they expressed a belief that their teammates were bots or experimenters. They could express such belief either during chat, or as part of their free-text responses to questions about the team during the game, or in the post-game survey. In fact, we assumed that participants would not develop the same social and emotional reactions to betrayal of presumed computer controlled entities or experimenters as they would for presumed human teammates. 76 of our participants were betrayers, 74 were decliners and 83 were controls.

4.2 Data Collected

We collected 2 types of data: behavioral logs of actions in the game and self-report measures. The behavioral data were in the form of time-stamped entries for what the participant saw and did: game content, text chats, button clicks, participant score, etc. As for the self-report measures, immediately after the game each participant was asked to complete a short demographic survey, as well as the validated survey PANAS (Positive Affect Negative Affect) [23], that gauges the respondent's affective state.

We preprocessed the behavioral data to develop measures ready for analysis. There were some variations in the AIs used, in that some AIs had both multiple signals and multiple time segment detection points, and some AIs were continuous measures, others were discrete, and some were human coded rather than automatically labeled. For continuous measures, we aggregated them at three different points: segment 0 (before the opportunity for betrayal), segment 1 (after the betrayal decision point but before the priming cue), and segment 2 (after both the betrayal decision point and the priming cue). For control participants who did not have a betrayal decision point, the corresponding time point in the game task was used. We then normalized the measures across the different segments with the baseline established as the behavior for segment 0 (the AI before inducement). For discrete measures, concatenation was used to make sure to record before and after the stimuli.

4.3 Analysis Methods and Results

Regarding the self-report measures (see Table 1 below), PANAS "Guilt" produced a stronger effect on betrayers, with a significant difference between them and both controls and decliners (one-tail t test $p < 0.0001$). As for anxiety, the "Afraid" and "Scared" measures of the PANAS scale showed some significant differences between betrayers and other subjects (one-tail t-tests $p < 0.05$). We also found a significant difference between betrayers and other subjects regarding the PANAS "Ashamed" measure (one-tail t-tests $p < 0.01$). This is a likely effect of the negative reactions of the team members to the betrayal.

Table 1. Statistics about self-reported measures.

Variable	Group	Statistics	Variable	Group	Statistics
"Guilt"	Betrayers	$\mu = 3$, $\sigma = 1.21$	"Ashamed"	Betrayers	$\mu = 2.82$, $\sigma = 1.14$
"Guilt"	Controls	$\mu = 2.13$, $\sigma = 0.73$	"Ashamed"	Controls	$\mu = 2.25$, $\sigma = 0.60$
"Guilt"	Decliners	$\mu = 2.32$, $\sigma = 0.78$	"Ashamed"	Decliners	$\mu = 2.35$, $\sigma = 0.83$
"Afraid"	Betrayers	$\mu = 2.46$, $\sigma = 0.99$	"Scared"	Betrayers	$\mu = 2.46$, $\sigma = 0.97$
"Afraid"	Controls	$\mu = 2.20$, $\sigma = 0.62$	"Scared"	Controls	$\mu = 2.20$, $\sigma = 0.62$
"Afraid"	Decliners	$\mu = 2.23$, $\sigma = 0.63$	"Scared"	Decliners	$\mu = 2.21$, $\sigma = 0.69$

We used two approaches to analyze the effect of AIs: (1) theory-based detector rules computed on the measures of single AIs and (2) theory-agnostic Machine Learning detector rules computed on the measures of single and multiple AIs. Due to space limitations, we will only discuss the results of AIs that were statistically significant or discriminative.

For (1), we defined and tested simple detector rules based on theory expectations about how betrayers' behavior should differ from controls' on each AI sub-measure. For example, for sub-measures that simply detect whether a participant responded to a prompt, a simple rule was "Betrayers do not respond, other participants do respond." For sub-measures that were scales (e.g. degree of positivity in response) or continuous (e.g. amount of time spent with the game window not activated), a "cut-off" value was selected based on criteria to balance between differentiation (TP/FP ratio) and detection (TP). We used the same process to set the cut-off value for each sub-measure. For this process, an initial cut-off value was selected to divide participants into betrayers vs controls or decliners by calculating the cut-off value such that 20% of betrayers were included. Then, the cut-off value was adjusted in the direction of increasing TP until a local maximum of TP/FP was found. We tested the performance of these rules for both the betrayer/control separation and the betrayer/decliner separation. The best discriminations between betrayers and controls produced by this method are shown in Table 2.

For (2), we used Machine Learning (ML) classification methods with the three experimental conditions (betrayers, decliners, and controls) as labels and AI measures as features. This approach estimates how much discriminative power an AI provides, agnostic to whether the rule follows psychological theory, and can screen both single and composite indicators (the latter made up of two or more individual AIs) to test whether they would provide more discrimination in combination. We also included demographics and post-game surveys as features to see if they could have discriminative power. We ran eight types of algorithms provided by the Weka ML library [24]: Functions (Support Vector Machines using Pearson VII Universal Kernel), Lazy Models (IBk (kNN classifier), KStar (instance-based learner)), Rules (JRip (RIPPER), Ridor (RIpple-DOwn Rule Learner)), Trees (FT (Functional Trees), J48 (C4.5 decision tree)), and Misc (VFI (Voting feature intervals)). The input feature vector is composed of 15 features (13 AIs plus the data from the PANAS and another survey).

This analysis enabled us to estimate whether our behavioral AI measures are more or less discriminative than individual characteristics or self-reported feelings. Each feature is composed of multiple sub-measures. We developed models that used one feature and models that used 381 combinations of features. We examined the results for both the f measure (as a function of precision and recall) and the TP/FP ratio. For each model type, we computed a random 75%/25% test/train split, trained the model and then computed the f measure and TP/FP ratio on the resulting confusion matrix. Table 2 lists the results of the best performing single-feature and multi-feature classifiers for betrayer/control separations.

Table 2. Summative table of best performing *AIs* using Theory-based rules and Machine Learning classifiers

Classifier type	AI	Results
Theory-based rules	*AI-Disengage*	TP/FP ratio = 6.09
	AI-Slack-Off	TP/FP ratio = 5.19
ML single-feature classifiers	*AI-Peer-Feedback*	TP/FP ratio = 6.8, f measure = 0.50
	AI-Peer-Trustworthiness	TP/FP ratio = 2.72, f measure = 0.51
	AI-Slack-Off	TP/FP ratio = 2.66, f measure = 0.37
	AI-Disengage	TP/FP ratio = 3.74, f measure = 0.28
ML multi-feature classifiers	all the best ones included: *AI-Peer-Trustworthiness*	TP/FP ratios > 2.4, f measures > 0.50
	15-feature model	TP/FP ratio = 2.93, f measure = 0.54
	the best 15-feature model included: *AI-Check-Peer-Performance* *AI-Peer-Trustworthiness* *AI-Shorter-Game*	TP/FP ratio = 4.16, f measure = 0.66

The most discriminative sub-measure was *AI-Disengage-B*, the number of text chats typed by the participant: it achieved a TP/FP ratio of 6.24 with TP = 30%. In fact, after the inducement to betray, the number of chats was significantly higher for betrayers than for decliners and controls (segment 1 - after the betrayal decision point but before the priming cue- betrayers: $\mu = 23.49$, $\sigma = 13.99$; decliners: $\mu = 16.58$, $\sigma = 11.33$; controls: $\mu = 12.83$, $\sigma = 9.39$; two-tails t-tests $p < 0.01$).

The best performing classifier varied widely across the different behavior measures. All eight models yielded the best result for some of the input vectors. JRip, FT and J48 tended to do better. See Table 3 for the details.

Even though the betrayers of our experiments were not requested to actively engage in sustained deceptive communication, they may have used communication strategies analogous to those of deceivers in prior studies, in that they chatted more than other groups ([20]), and may have produced effects similar

Table 3. Performance of ML algorithms

ML type	ML technique	Count of testdata best model (TP/FP ratio Yes)	Count of testdata best model (FMeas)
Function	SVM	29	30
Lazy	IBk	16	24
Lazy	KStar	30	27
Misc	VFI	46	25
Rules	JRip	90	122
Rules	Ridor	21	12
Trees	FT	76	67
Trees	J48	73	74
Grand Total		381	381

to those found by Anolli et al. [21] when "lying to a suspicious recipient". In fact, the strong negative reactions of the teammates to the announcement of the betrayal, and the anxiety caused by the risk of being caught, may have caused betrayers to actively attempt to persuade teammates about their innocence by showing an active participation to the game so as to be seen as good team members.

We also informally examined the chats to identify the possible pragmatic strategies used by betrayers vs decliners and controls, but could not find qualitative differences between the three groups of subjects, in that all groups seem to pursue the same communicative goals but with different frequencies. We could not tell the intentions behind the betrayers' increased frequency of such communicative goals without another study that includes interviews or betrayers' reflection on their behaviors.

5 Conclusions and Future Work

In this work, we aimed to study betrayers by placing "Active Indicators" (AIs) in the environment to elicit indicative responses. Our work is the first to use a game to explore emotional indicators as a way to detect betrayers using online behaviors after the act of betrayal. The game provided us with a controllable environment, including replicable teammates, where we could continuously monitor the subjects' behaviors to deduce the effects of stimuli on them. We measured the detector signals of each AI by collecting behavioral data in the form of time-stamped entries for what the participants saw and did. Our results show that some AIs (those that target engagement, persistence, feedback to teammates and team trust) have a promising discriminatory power, both taken singularly and combined with other AIs. One AI related to engagement was the number of chats, which was much higher for betrayers than controls and decliners, confirming other results in the literature [20,21]. In our case, betrayers did "role-playing": they probably produced more chats than decliners and controls to pretend that they were good team members, by participating more in the discussions. For future work, we aim to conduct more qualitative analyses to understand the communicative strategies that betrayers used as opposed to others. We also plan to use other games as experimental environments to analyze the effects of AIs on betrayal behaviors controlled by habits and logical reasoning. These results are promising and follow on studies should take this further to investigate the effect of the successful AIs in different contexts.

Acknowledgments. The research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2016-16031100003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Herbig, K.L., Wiskoff, M.F.: Espionage against the united states by American citizens 1947–2001. Technical report, Defense Personnel Security Research Center Monterey CA (2002)
2. Cummings, A., Lewellen, T., McIntire, D., Moore, A.P., Trzeciak, R.: Insider threat study: illicit cyber activity involving fraud in the us financial services sector. Technical report, Carnegie Mellon University, Software Engineering Institute, Pittsburgh, PA (2012)
3. Kont, M., Pihelgas, M., Wojtkowiak, J., Trinberg, L., Osula, A.M.: Insider threat detection study. NATO CCD COE, Tallinn (2015)
4. Carter, M.: Massively multiplayer dark play: Treacherous play in eve online. The Dark Side of Game Play. Routledge, London (2015)
5. Ponemon Institute: 2016 cost of insider threats. Technical report, Traverse City, MI, USA (2016)
6. Greitzer, F.L., Paulson, P., Kangas, L., Franklin, L.R., Edgar, T.W., Frincke, D.A.: Predictive modelling for insider threat mitigation. Pacific Northwest National Laboratory, Richland, WA, Technical report, PNNL-65204 (2009)
7. Sasaki, T.: A framework for detecting insider threats using psychological triggers. JoWUA **3**(1/2), 99–119 (2012)
8. Lane, J.D., Wegner, D.M.: The cognitive consequences of secrecy. J. Personal. Soc. Psychol. **69**(2), 237 (1995)
9. Kelly, A.E.: The Psychology of Secrets. Springer, Heidelberg (2002). https://doi.org/10.1007/978-1-4615-0683-6
10. Charney, D.: True psychology of the insider spy. Intell.: J. US Intell. Stud. **18**(1), 47–54 (2010)
11. Intelligence Advanced Research Program Agency (IARPA): Iarpa scite program (2015)
12. Rashid, T., Agrafiotis, I., Nurse, J.R.: A new take on detecting insider threats: exploring the use of hidden markov models. In: Proceedings of the 2016 International Workshop on Managing Insider Security Threats, pp. 47–56. ACM (2016)

13. Bindu, P., Thilagam, P.S.: Mining social networks for anomalies: methods and challenges. J. Netw. Comput. Appl. **68**, 213–229 (2016)
14. Spitzner, L.: Honeypots: catching the insider threat. In: Computer Security Applications Conference, 2003. Proceedings. 19th Annual, pp. 170–179. IEEE (2003)
15. Peled, N., Bitan, M., Keshet, J., Kraus, S.: Predicting human strategic decisions using facial expressions. In: IJCAI, pp. 2035–2041 (2013)
16. Niculae, V., Kumar, S., Boyd-Graber, J., Danescu-Niculescu-Mizil, C.: Linguistic harbingers of betrayal: a case study on an online strategy game. arXiv preprint arXiv:1506.04744 (2015)
17. Ho, S.M., Warkentin, M.: Leader's dilemma game: an experimental design for cyber insider threat research. Inf. Syst. Front. **19**(2), 377–396 (2017)
18. Ho, S.M., Liu, X., Booth, C., Hariharan, A.: Saint or sinner? language-action cues for modeling deception using support vector machines. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) Social, Cultural, and Behavioral Modeling. LNCS, vol. 9708, pp. 325–334. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-319-39931-7_31
19. Burgoon, J.K., Dunbar, N.E.: An interactionist perspective on dominance-submission: interpersonal dominance as a dynamic, situationally contingent social skill. Commun. Monogr. **67**(1), 96–121 (2000)
20. Dunbar, N.E., Jensen, M.L., Bessarabova, E., Burgoon, J.K., Bernard, D.R., Harrison, K.J., Kelley, K.M., Adame, B.J., Eckstein, J.M.: Empowered by persuasive deception: the effects of power and deception on dominance, credibility, and decision making. Commun. Res. **41**(6), 852–876 (2014)
21. Anolli, L., Balconi, M., Ciceri, R.: Linguistic styles in deceptive communication: dubitative ambiguity and elliptic eluding in packaged lies. Soc. Behav. Personal.: Int. J. **31**(7), 687–710 (2003)
22. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 319–326. ACM (2004)
23. Watson, D., Clark, L.A., Tellegen, A.: Development and validation of brief measures of positive and negative affect: the panas scales. J. Personal. Soc. Psychol. **54**(6), 1063 (1988)
24. Frank, E., Hall, M.A., Witten, I.H.: Weka 3: Data mining software in java (2017). http://www.cs.waikato.ac.nz/ml/weka/

Forecasting Gang Homicides with Multi-level Multi-task Learning

Nasrin Akhter[1](✉), Liang Zhao[1], Desmond Arias[1], Huzefa Rangwala[1],
and Naren Ramakrishnan[2]

[1] George Mason University, Fairfax, VA, USA
{nakhter3,lzhao9,earias2}@gmu.edu, rangwala@cs.gmu.edu
[2] Virginia Tech, Arlington, VA, USA
naren@cs.vt.edu

Abstract. Gang-related homicides account for a significant proportion of criminal activity across the world, especially in countries of Latin America. They often arise from territorial fights and, distinct from other types of homicides, are characterized by area-specific risk indicators. Current crime modeling and prediction research has largely ignored gang-related homicides owing to: (i) latent dependencies between gangs and spatial areas, (ii) area-specific crime patterns, and (iii) insufficiency of spatially fine-grained predictive signals. To address these challenges, we propose a novel context-aware multi-task multi-level learning framework to jointly learn area-specific crime prediction models and the potential operating territories of gangs. Specifically, to sufficiently learn the finer-grained area-specific tasks, the abundant knowledge from coarse-grained tasks is exploited through multi-task learning. Experimental results using online news articles from Bogotá, Colombia demonstrate the effectiveness of our proposed method.

Keywords: Multi-task learning · Gang homicide · Crime forecasting

1 Introduction

Homicidal violence is concentrated in the Americas [1], especially in Latin American and Caribbean countries [2]. Gang wars, involving narco businesses, are the key contributors to Latin America's homicidal violence problem. Similarly, 90% of gun violence can be attributed to gangs in the United States [3]. Existing homicide prediction research mostly ignores gang involvement and spatial heterogeneity of crime indicators within cities. Often a country or a city has crime pockets dominated by local gangs which are likely to influence the crime scene of neighboring areas. In this study, we aim to identify patterns of activities in multiple locations as indicators for future events. For instance, the arrest of a gang leader could incite aggression by rivals gangs in the neighborhoods, leading to homicides.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 28–37, 2018.
https://doi.org/10.1007/978-3-319-93372-6_3

Forecasting gang-related homicides from online news demands several challenges to be solved: (1) **Scarcity of fine-grained location information**. City-level news articles report local crimes as well as crimes with nationwide impact. Although there may be a reasonable amount of city-level data available from city-level newspapers, dividing them into even finer levels, i.e., suburbs within a city, often suffers from data scarcity. (2) **Heterogeneity of geographical locations**. Although nearby locations may be influenced by the same regional phenomena, each region has its own exclusive set of characteristics and principle actors affecting that particular region. Accounting for this location heterogeneity is crucial in predicting future area-specific homicidal violence. (3) **Multi-resolution feature structure**. The set of keywords in the homicide-reporting news articles often exhibit a subtle hierarchical structure. On top is the common homicide and violence related keywords, area-specific entity names and keywords lie at the bottom level. A model, trained on a global set of keywords, is unlikely to learn this two-level feature structure. In order to address these challenges, we propose a novel multi-task learning framework that learns area-specific patterns for predicting area-specific violence intensity attributed by gang-homicides. The study was carried out on Bogotá, a Colombian city with a high level of violent crime [1].

2 Related Work

Hotspot mapping is one of the most popular approaches for mapping crime-prone locations [4,5]. Crime has also been predicted using time series model ARIMA [6]. Twitter has been effective in identifying risk indicators [7,8]. The crime prediction literature has featured a variety of methods such as regression models [9], Bayesian approaches [10] and neural networks [11].

Of all the crime prediction models, only a handful of them focus solely on homicide. The use of hotspot maps to predict homicide and gun-crime can be found in [12]. Berk et al. [13] studied murder rates among probationers and parolees. Nineteen years of data were utilized to forecast homicide, robbery, burglary, and motor vehicle theft in [14].

Multi-task learning (MTL) is concerned with learning multiple related tasks simultaneously to achieve a better generalization performance [15]. Different assumptions on task relatedness result in different MTL strategies. For instance, assumptions can be made that the task parameters share a common subspace [16], or that they use a tree-structured model to share a common underlying structure [17]. MTL has been successfully employed in various applications including text classification, natural language processing, and computer vision.

3 Problem Formulation

We learn two different sets of keywords for two different levels of features: area-agnostic common keywords such as ‘cocaine’ (cocaine), ‘levantón’ (kidnapping), 'sicarios' (hitmen), and area-specific keywords focusing on gangs based on specific locations. Given a geographical region, the set of newspaper articles published in the past h days covering the news focused on the *i-th* area in that region is denoted by $X_i = \{x_{i,1}, x_{i,2}, \ldots, x_{i,h}\}$, where $x_{i,j}$ denotes a collection of news articles published on day j based on the *i-th* area. Violence intensity over a window of w days is denoted by $Y^{(1)}$. The number of homicides per day is denoted by $Y^{(2)}$. The prediction problems can be formulated as two different mappings. Firstly, we seek to learn $f : X_i \rightarrow {Y^{(1)}}_{i,h+k}$, where the right hand side denotes the intensity of violence in the *i-th* area, predicted k days before the target time window by reading past h days of news articles (classification problem). Another function $f : X_i \rightarrow {Y^{(2)}}_{i,h+k}$ maps the sets of news articles from past h days to the number of homicides that may be committed in future, again k days in advance (regression problem). From the ground truth data, we compute homicide statistics such as the average and median of actual number of homicides over our study period in each area. If the average number of homicides committed over a given time window exceeds the median, we identify a 'large scale' violence for that time period. Otherwise, it is 'small scale'. *History days*, denoted by h, refers to the number of days' articles in the past that the model would use to make a prediction. *Lead time*, denoted by k, refers to how many days in advance the model would make a forecast.

4 Models

We model S different tasks for S different locations. Our proposed strategy simultaneously learns models for all S locations in a multi-task feature learning framework. We divide our feature matrix W row-wise into G and R such that R rows follow the top G rows as shown in Fig. 1. G denotes the general features and R denotes the area-specific features. Our model minimizes the following:

$$\min_{W} \quad f(W) + \lambda_1 g_1(G) + \lambda_2 g_2(R), \tag{1}$$

where $f(W)$ is the empirical loss. We use the least squares loss which is smooth and convex. g_i represents the regularization function. The tunable parameter λ_i controls the model sparsity and balances the emphasis between the loss and the penalty. We propose: (I) multi-level multi-task (MLMT) model, and (II) constrained multi-level multi-task (cMLMT) model that simultaneously learn features for all areas in a multi-task, multi-level feature learning framework.

4.1 MLMT Model

We apply regularization at two levels to capture the multi-level feature representation. The models need to be able to take advantage of shared common features across locations and learn location-specific features. We apply $\ell_{2,1}$-norm to jointly learn a set of across-task features. Area-specific features for each task are selected in the next level when we directly apply ℓ_1-norm regularization on gang-related feature set R. The objective function for our proposed model is:

$$\min_{W=[G;R]} \sum\nolimits_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda_1 \|G\|_{2,1} + \lambda_2 \|R\|_1. \tag{2}$$

S denotes the total number of tasks. $\mathcal{L}$ denotes the loss function. $\lambda_1\|G\|_{2,1}$ denotes the $\ell_{2,1}$-norm on G. $\lambda_2\|R\|_1$ is the ℓ_1-norm regularization which enforces individual sparsity on each task. λ_1 controls the group sparsity, and λ_2 controls sparsity in area-specific features. The $\ell_{2,1}$-norm on G makes the model select a common set of features for all the tasks while ℓ_1-norm on R learns features exclusively associated with each area.

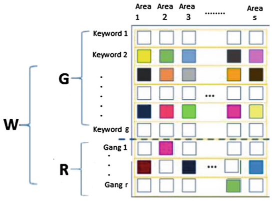


Fig. 1. Illustration of MLMT. Each column represents a model for an area. The rows represent the feature vectors. The general features are specified by first G rows. The area-specific features are represented by the rest (R rows).

4.2 cMLMT Model

This model offers a way to constrain the feature learning process. It is often desirable to identify the level of correlation between gangs and locations. In this formulation, we prohibit the area-specific features to take on negative scores. As a result, they become either zero or take on positive weights. This affects the overall scores distribution across tasks. As the models try to minimize the empirical loss, the gang-area correlation, modeled by area-specific features, are rearranged in such a way that the empirical loss do not increase significantly. The resulting area-specific weights offer insight into each gang's contribution in the ongoing violence in each location. Below is the constrained form of MLMT:

$$\begin{aligned} \min\nolimits_{W=[G;R]} \quad & \sum\nolimits_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda_1 \|G\|_{2,1} + \lambda_2 \|R\|_1. \\ & \text{s.t.} \quad R \geq 0 \end{aligned} \tag{3}$$

5 Algorithm

Both of our optimization problems have two non-smooth terms. Equation (3) is a constrained form of Eq. (2) such that $R \geq 0$. To solve these problems, we develop an algorithm based on proximal gradient descent. The basic idea is to first use the gradient at the current search point and apply proximal operator on $W^i - \frac{1}{L}\nabla F(W^i)$ to find an approximate solution point. In other words, we find an approximate solution point by applying $prox_{\lambda g}(W^i - \frac{1}{L}\nabla F(W^i))$. This is a gradient step towards the optimal solution point. Line 5 of our algorithm can also be viewed as proximal operator of first order approximation. The approximate solution point found in the current iteration would be used as the *current* search point in the next iteration. The step size is $\frac{1}{L}$, and L is determined by a line search method. The details are given in Algorithm 1 where,

$$\nabla F(W) = X(X^T W - Y). \tag{4}$$

Algorithm 1. The Proposed Algorithm

```
Require: : X, Y, $\rho, \eta > 1$
Ensure: : solution W
 1: Initialize $W^0, \eta = 0.5$
 2: for $i \leftarrow 1, 2, 3, \ldots$ do
 3:     Initialize L = 1
 4:     repeat
 5:         $\hat{W}^i \leftarrow W^i - \frac{1}{L}\nabla F(W^i)$        ▷ $\hat{W}^i = [\hat{G}^i; \hat{R}^i]$
 6:         $G^i \leftarrow prox_{2,1}(\hat{G}^i)$
 7:         $R^i \leftarrow prox_1(\hat{R}^i)$
 8:         $L \leftarrow \eta L$
 9:     until line search criterion is satisfied
10:     if the objective stop criterion is satisfied then
11:         Return $W^i$
12:     end if
13: end for
```

Note that we divide W into G and R such that $W = [G; R]$. We have two subproblems to solve: proximal $\ell_{2,1}$ regularized problem in line 6 and proximal ℓ_1 regularized problem in line 7, both of which have closed form solutions. We solve the proximal operator with $\ell_{2,1}$-norm on G by,

$$Prox_{2,1}(G) = (max(\|G\|_2 - \lambda, 0)/\|G\|_2)G. \tag{5}$$

Juxtaposition of two quantities implies matrix multiplication. Recall that G is the set of general features that occupies the top G rows in our feature matrix. For proximal of ℓ_1 on R, the closed form solution is given by,

$$Prox_1(R) = sign(R).max(abs(R) - \lambda, 0). \tag{6}$$

The dot denotes element-wise multiplication. For the constrained optimization problem given in (3), the solution to proximal operator with ℓ_1-norm is given by,

$$Prox_1(R) = max(abs(R) - \lambda, 0). \tag{7}$$

In every iteration, the algorithm finds an approximate solution point that gets closer to the optimal solution point. The algorithm iterates until the optimal point is found, or the maximum number of iteration is reached.

6 Experiments

6.1 Dataset

The experiments were carried out on 10,672 newspaper articles collected from several news agencies such as El Colombiano, El Universal, RCN Radio, El Tiempo, El Confidencial, NTN24, and El Nuevo Herald between April 2015 and May, 2016. The articles were in Spanish. We used police records on homicides in Bogotá for evaluating our model's performance.

6.2 Data Preprocessing

We worked on three regions in Bogotá: far Northwest, center and center south, and far south. Often the articles refer to multiple locations. We use the geometric median of the GPS coordinates of the localities appearing in an article to determine the finer-grained location information. Each news article is assigned a location based on the geometric median, m, of the GPS locations, L, of the areas mentioned in that news article.

$$m = \underset{x \in L}{\operatorname{argmin}} \sum_{y \in L} \mathit{distance}(x, y), \tag{8}$$

where $\mathit{distance}(x, y)$ is the orthonormic distance calculated using Vincentry's formula [18]. There is a possibility that some news articles are not assigned to any of our three target areas even though it may belong to one. To compensate this situation, we accommodate a fourth task that contains all the news articles that are based on Bogotá, but are not assigned to any of our three pre-defined regions.

6.3 Experimental Setup

We denote 'large scale' violence by 1, and 'small scale' violence by 0. Our model outputs either 0 or 1 in the violence intensity setting. Examples of general keywords can be found in Sect. 3. For area-specific features, we use names of gangs, armed groups, and members of those groups such as 'Los Rastrojos', 'Clan Úsuga', 'Pastor Alape', which are either drug-trafficking paramilitary groups or members of those groups.

The input for each task is an $n \times m$ matrix where n denotes the number of input samples, and m denotes number of features. Each input sample (i.e., row) is constructed by counting the frequencies of the features occurring in the news articles published in past h days starting from a particular date. Each cell in that row, therefore, represents the frequency of a specific feature (i.e., general keyword or area-specific keywords) in the news articles over the same time period. Imagine a sliding window that starts from the starting date of the training period with a window width of h. We slide this window over time until the right end of that window touches the end date of the training period. While the window slides, the input matrix gets constructed.

We have three tunable parameters in our model: lead time k, history days h, and time window w are the tunable parameters. As an example, if k is 3, h is 5, and w is 4, then the model will read past 5 days of news articles to predict the violence intensity over 4 days; the prediction will be made 3 days in advance. Changing the values of these variables would yield different models, each having their own specification of lead time, time window, and history days. In this article, we show the results when $k = 1$, $h = 5$, and $w = 2$. The regularization parameters λ_1 and λ_2 were selected via a 5-fold cross-validation.

6.4 Comparison Methods

For the classification task, we compare our proposed models with Support Vector Machine (SVM), Logistic Regression, regularized LASSO and the baseline approach **monotonic multi-task** (MMT) given by:

$$\min_{W} \sum\nolimits_{s=1}^{S} \mathcal{L}(f(X_s, W_s), Y_s) + \lambda \|W\|_{2,1}. \tag{9}$$

The baseline method does not distinguish between general and area-specific features. The regularization parameter λ was selected via a 5-fold cross-validation. These models are area-ignorant in the sense that they do not capture the multi-level feature structure. We use an L_2-penalized logistic regression and the *liblinear* solver. For the SVM, we use the radial basis function (*rbf*) kernel with co-efficient gamma set to 0.7. The regularization parameter λ for LASSO were determined via 5-fold cross-validation. For the regression task, we compare our model with seasonal ARIMA and Support Vector Regression (SVR). We use the police record data to build the seasonal ARIMA model. For SVR, we use the *rbf* kernel with the penalty parameter set to 0.8. Note that the parameter values for the comparison methods were selected by a trial and error method. We select the values that give the best performance for each comparison model.

6.5 Results and Discussion

For the homicidal violence prediction task, we consider four performance metrics: precision, recall, F1-score, and ROC AUC (Area Under the Receiver Operating Characteristic Curve). Table 1 shows the performance comparisons between our proposed model and the comparison methods for the classification task. The results show that our proposed model MLMT performs better, on average, than other methods. MLMT outperforms the baseline method by 5% to 10% in precision, recall, and F1-score in Area 1. In Area 3, the baseline is outperformed by MLMT by 10.3% to 20.8% in precision, recall, F1-score, and AUC. This implies

Table 1. Violence intensity prediction performance comparison (precision, recall, F1-score, AUC).

Methods	Area 1 p, r, f1, auc	Area 2 p, r, f1, auc	Area 3 p, r, f1, auc	Area 4 p, r, f1, auc
Logistic regression	0.44, 0.5, 0.47, 0.23	0.40, 0.41, 0.40, 0.46	0.46, 0.46, 0.46, 0.49	0.49, 0.49, 0.49, 0.57
SVM	0.44, 0.5, 0.47, 0.35	0.22, 0.5, 0.31, 0.53	0.4, 0.5, 0.44, 0.59	**0.79**, 0.51, 0.62, 0.57
LASSO	0.47, 0.44, 0.46, 0.45	0.50, 0.50, 0.50, 0.47	0.59, 0.58, 0.59, 0.59	0.51, 0.51, 0.51, 0.51
MMT	0.64, 0.83, 0.72, **0.97**	**0.69, 0.72, 0.71**, 0.77	0.69, 0.81, 0.75, 0.79	0.66, **0.68, 0.67**, 0.69
MLMT	**0.74, 0.88, 0.81, 0.97**	**0.69, 0.72**, 0.70, **0.78**	**0.71, 0.83, 0.76, 0.80**	0.65, 0.67, 0.66, **0.7**
cMLMT	0.64, 0.83, 0.72, 0.94	0.67, 0.71, 0.69, 0.72	0.68, 0.79, 0.73, 0.79	0.64, 0.64, 0.64, 0.69

Table 2. Homicide count prediction performance comparison (RMSE, MAE).

Methods	Area 1 rmse, mae	Area 2 rmse, mae	Area 3 rmse, mae	Area 4 rmse, mae
SVR	0.87, 0.75	1.58, 1.25	1.65, 1.34	1.65, 1.34
SARIMA	**0.11, 0.01**	0.88, 0.62	0.81, 0.60	0.79, 0.56
LASSO	0.37, 0.14	0.92, 0.64	0.52, 0.27	1.37, 0.86
MMT	0.28, 0.08	**0.53, 0.29**	0.37, 0.14	**0.60, 0.36**
MLMT	**0.26, 0.07**	0.54, **0.29**	**0.36, 0.13**	**0.60**, 0.37
cMLMT	0.28, 0.08	0.55, 0.30	0.38, 0.14	0.61, 0.38

that each location does have its own specific factors that affect the intensity of homicide-induced violence in that area.

Table 2 shows a performance comparison for the regression task. We use RMSE and MAE as the performance metrics. Note that the seasonal ARIMA model was not constrained with *history days* and *lead time*. It enjoyed as much data as we had for the training period with no restriction on history days, which may have attributed to better RMSE and MAE scores for Area 1. However, if we compare MLMT with only the baseline MMT, it outperforms the baseline. We present a comparison of the performances by varying lead time in Fig. 2. **History days** was fixed to 5. Figure 2 (left panel) shows that the MLMT model achieves better F1 score than the others, especially with increased lead time. This is explained by the fact that a precursor incident such as an arrest or a murder committed by a rival group will not necessarily generate an immediate reaction. Often, the rival group's attempt to take control of a local business controlled by another group, or a retaliatory murder may take some time to happen. This gap between an event and the reactive violence may be a reason for why the models generally perform better with an increasing lead time.

Figure 2 (right panel) also shows a performance comparison in AUC when the number of history days varies. *Lead time* was fixed to 1 day with varying number of history days. We compare only the baseline method and MLMT since other models perform worse. Figure 2 shows that MLMT mostly performs better, or no less than the baseline method. This consistency in better F1 score and AUC when the lead time and history-days vary shows the necessity of capturing the multi-level feature structure for predicting gang-related homicides.

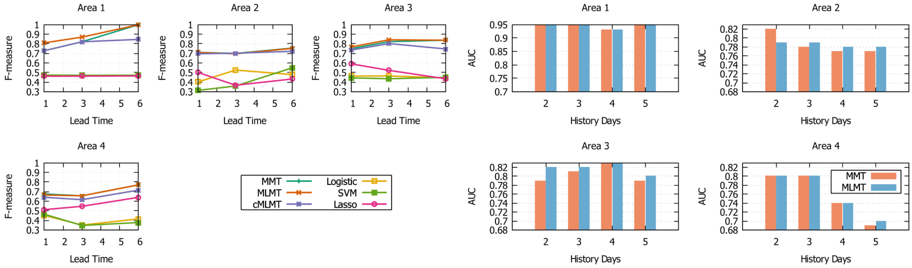


Fig. 2. Visualization of performance comparison. F-measure comparison when the lead time varies (left) and AUC comparison when the number of history days varies (right).

Table 3. Model-selected top 4 area-specific features

Area 1	Area 2	Area 3	Area 4
Mao	Roman Ruiz	El Médico	ELN
Clan Úsuga	AUC	EPL	Otoniel
Comba	FARC	El Coronel	FARC
Omar	clan Úsuga	Roman Ruiz	El Coronel

Table 3 shows the model-selected gang-related features. While the general keywords present an area-agnostic global view of the feature space, the gang-related features demonstrate a subtle dependency on the spatial areas. For instance, violence in Area 2 connects highly with three armed groups: FARC, AUC (Autodefensas Unidas de Colombia), and Clan Úsuga. AUC was a rival of FARC, and Clan Úsuga emerged when AUC was being demobilized. While Clan Úsuga is also a top contributor to violence in Area 1 with its leaders Omar and Mao, another group Rastrojos also contributes via its leader Comba in the same area. Rastrojos is a rival of Clan Úsuga. Rivalry leads to more violence in general. We find a different group EPL (Ejército Popular de Liberación) affecting Area 3. FARC is also present in Area 4 together with its another former rival ELN (National Liberation Army). The rest in Table 3 are members of the aforementioned groups. The area-agnostic features together with these gang-related area-specific features indicate a multi-level feature structure. Note that the sets of top contributors for each task are mostly different from each other.

7 Conclusion

We present a novel approach that learns features at two different resolutions to predict gang-homicide and violence intensity. Existing homicide prediction works do not distinguish between shared common information across locations and location-specific information. Our proposed method addresses these issues by simultaneously learning models for multiple tasks while capturing the multi-level structure of the features. Empirical results show that our proposed model can effectively predict gang-homicides and homicidal-violence intensity.

Acknowledgments. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

1. Igarapé Institute: Homicide monitor (2017)
2. Muggah, R.: Interactive Map Tracks Murder Rate Worldwide (2015)
3. Saul, J.: Why 2016 Has Been Chicago's Bloodiest Year in Almost Two Decades (2016)
4. Gorr, W.L., Lee, Y.: Early warning system for temporary crime hot spots. J. Quant. Criminol. **31**(1), 25–47 (2015)
5. Weisburd, D., Braga, A.A., Groff, E.R., Wooditch, A.: Can hot spots policing reduce crime in urban areas? An agent-based simulation. Criminology **55**(1), 137–173 (2017)
6. Chen, P., Yuan, H., Shu, X.: Forecasting crime using the ARIMA model. In: Fifth International Conference on FSKD 2008, vol. 5, pp. 627–630. IEEE (2008)
7. Gerber, M.S.: Predicting crime using Twitter and kernel density estimation. Decis. Support Syst. **61**, 115–125 (2014)
8. Wang, X., Brown, D.E., Gerber, M.S.: Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In: 2012 IEEE International Conference on ISI, pp. 36–41. IEEE (2012)
9. Shingleton, J.S.: Crime trend prediction using regression models for Salinas, California. Ph.D. thesis. Naval Postgraduate School, Monterey, California (2012)
10. Liao, R., Wang, X., Li, L., Qin, Z.: A novel serial crime prediction model based on Bayesian learning theory. In: 2010 International Conference on ICMLC, vol. 4, pp. 1757–1762. IEEE (2010)
11. Kang, H.W., Kang, H.B.: Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE **12**(4), e0176244 (2017)
12. Mohler, G.: Marked point process hotspot maps for homicide and gun crime prediction in chicago. Int. J. Forecast. **30**(3), 491–497 (2014)
13. Berk, R., Sherman, L., Barnes, G., Kurtz, E., Ahlman, L.: Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. J. R. Stat. Soc.: Ser. A **172**(1), 191–211 (2009)
14. Pepper, J.V.: Forecasting crime: a city-level analysis. In: Understanding Crime Trends: Workshop Report, National Research Council, pp. 177–210 (2008)
15. Zhao, L., Sun, Q., Ye, J., Chen, F., Lu, C.T., Ramakrishnan, N.: Multi-task learning for spatio-temporal event forecasting. In: Proceedings of the 21th ACM SIGKDD, pp. 1503–1512. ACM (2015)
16. Acharya, A., Mooney, R.J., Ghosh, J.: Active multitask learning using supervised and shared latent topics. In: Pattern Recognition and Big Data, p. 75 (2016)
17. Kim, S., Xing, E.P.: Tree-guided group lasso for multi-task regression with structured sparsity (2010)
18. Vincenty, T.: Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Surv. Rev. **23**(176), 88–93 (1975)

Feature Selection of Post-graduation Income of College Students in the United States

Ewan Wright[1](✉), Qiang Hao[2], Khaled Rasheed[3], and Yan Liu[4]

[1] University of Hong Kong, Pokfulam, Hong Kong SAR
etwright@hku.hk
[2] Western Washington University, Bellingham, WA, USA
[3] University of Georgia, Athens, GA, USA
[4] University of British Columbia, Vancouver, Canada

Abstract. This study investigated the most important attributes of the 6-year post-graduation income of college graduates who used financial aid during their time at college in the United States. The latest data released by the United States Department of Education was used. Specifically, 1,429 cohorts of graduates from three years (2001, 2003, and 2005) were included in the data analysis. Three attribute selection methods, including filter methods, forward selection, and Genetic Algorithm, were applied to the attribute selection from 30 relevant attributes. We discuss how higher numbers of students in a cohort who grew up in Zip code areas where over 25% of the population hold a Professional Degree was predictive of more college graduates being classified as High income.

Keywords: Attribute selection · Feature selection
Post-graduation income classification · Post-graduation income prediction
Social stratification

1 Introduction

Higher education is an excellent "investment" that should be encouraged by families, schools, communities, and policy makers. The returns of a college degree vis-à-vis a high school diploma has expanded considerably in recent decades. Autor [1] found that this "graduate premium" doubled in real terms between 1979 and 2012. The gap in earnings between the median college educated worker and the median high-school educated worker increased from $17,411 to $34,969 for men, while also increasing from $12,887 to $23,280 for women. Research by Chetty et al. [2] underscores the role of higher education as a key pathway to intergenerational social mobility in the U.S. Further, Hout [3] contends that higher education "makes life better" through a host of social benefits in community relations, health, family stability, and social connections.

Yet as higher education participation has expanded [4], college graduates have become an increasing heterogeneous population with increasingly disparate labor market outcomes [5, 6]. While some graduates are highly successful, others face challenges to gainful employment. Data from the Federal Reserve Bank of New York [7] shows that 43.4% of college graduates aged between 22 and 27 graduates are under-employed or employed in a job that "typically does not require a college degree",

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 38–45, 2018.
https://doi.org/10.1007/978-3-319-93372-6_4

while 12.7 are employed in “low-wage jobs” that tend to pay below $25,000 per annum. Research has established that field of study [8] and institutional selectivity [9] are important features in post-graduation incomes. Building on the literature, this study explored the most important attributes of 6-year post-graduation income of college graduates who used student aid from the U.S. Department of Education, and to what extent of accuracy the select attributes can be used to classify post-graduation income. The research questions were: (1) What are the most important attributes of post-graduation income of college students who graduate with debt repayment obligations? and (2) To what extent can the selected attributes classify post-graduation income of college students who graduate with debt repayment obligations?

2 Research Design

2.1 Data Collection

The data for this study was the latest dataset – released in October 2015 – by College Scorecard under the U.S. Department of Education [10]. This dataset only covered students who used financial aid during their college study period. Each row in the data stands for a student cohort admitted to a certain university. The data ranged from 1996 to 2013, but the 6-year post-graduation income data are only available for the years 1997, 1999, 2001, 2003 and 2005. The response variable in the present study is the mean value of the 6-year post-graduation income of a student cohort. Attributes were filtered based on domain knowledge. Factors deemed less relevant were excluded, such as latitude of the institution and percent of students who passed away within 6 years after graduation.

30 potential attributes (see Appendix A) under five groups were included in this study. The groups are: (1) School, (2) Admission, (3) Cost, (4) Student Cohort, and (5) Socioeconomic Status of Students. Some attributes in certain groups are not available before 2000, such as admission rate in the Admission Group. Thus, only three years of data, including 2001, 2003, and 2005 were used. 1,429 cohorts were included for the data analysis. The response variable, mean income value of each cohort, was discretized into four classes based on the American Individual Income Distribution; including Very low (0 to 25,000), Low (25,000 to 37,500), Middle (37,500 to 50,000), and High (Above 50,000) [11].

2.2 Data Analysis

Two steps of preprocessing were applied to the collected data before the analysis: (1) *Standardization*: Standardization, transforming raw scores to z-scores, was applied to all the numerical attributes. There were 28 numerical attributes in total; (2) *One-hot encoding*: One-hot encoding techniques were applied to all the nominal attributes. There were 2 nominal attributes.

Three attribute selection methods were applied and compared, including filter methods, stepwise wrapper methods, and naturally inspired algorithms. The filter methods applied in this study included five algorithms: (1) OneR algorithm,

(2) Relief-based selection, (3) Chi-square selection, (4) Gain-ratio-based selection, and (5) Information-gain-based selection.

Both stepwise wrapper methods and naturally inspired algorithms need to have an evaluation function to work. Logistic regression was chosen as the evaluation function of both for stability and efficiency. The stepwise wrapper methods included forward and backward selection. Forward selection starts with no attributes in the model, and tests the addition of each attribute using certain comparison criteria. Backward selection starts with all candidate attributes, and tests deletion of each attribute using certain criteria. Only forward selection was used in this study.

The naturally inspired algorithm implemented was the Genetic Algorithm. Genetic Algorithm is a computational algorithm with origins in the field of biology. The tools that Genetic Algorithm uses have marks of genetic systems, including generation selection, crossover, and mutation [12]. We implemented the simple form of Genetic Algorithm described by Goldberg [13].

Weighted average F1-score was chosen as the primary evaluation criterion, because there exists an imbalance in the four income classes. A classifier that primarily guesses based on the majority class would achieve a small advantage in accuracy, but would perform worse in terms of the F1-score. Also, classification accuracy rate was used as the secondary evaluation criterion. Ten-fold cross validation was used for the estimation of both F1-score and accuracy rate.

3 Results

Five filter methods, including (1) OneR algorithm, (2) Relief-based selection, (3) Chi-square selection, (4) Gain-ratio-based selection, and (5) Information-gain-based selection, were applied to the attribute selection. The 10-fold cross validation scheme was implemented in Weka [14]. As opposed to the cross-validation in prediction or classification, no training or testing is involved in the cross-validation scheme of attribute selection. Under such a scheme, the dataset was randomly sectioned into 10 folds, and only 9 folds were used for subset attribute selection in each round. There were 10 rounds in total. The 10 selection results were summarized afterwards. The attributes selected by at least three out of the five methods (60%) were selected, yielding 14 selected attributes in total. The arithmetic mean of each attribute's ordinal ranking across all selection methods was also calculated, to enable measuring of attribute usefulness. For each single-attribute evaluator, the output of Weka showed the average merit and average rank of each attribute over the 10 folds (see Table 1).

Same as the implementation of filter methods, 10-fold cross validation scheme in Weka was used for more stable estimates. Attributes selected by at least six out of ten folds (60%) were selected, yielding 9 selected attributes in total. The selected attributes are presented in Table 2.

In alignment with the prior two attribute selection approaches, 10-fold cross validation scheme in Weka was used. Attributes selected by at least six out of ten folds (60%) were selected, yielding 22 selected attributes in total (see Table 3).

Table 1. Selected attributes subset using filter methods

Attributes	Votes*/ Average rank*	Attributes	Votes*/ Average rank*
% of Population from Students' Zip Codes over 25 with a Professional Degree	5/2.88	Admission Rate	5/12.42
Average Faculty Salary	5/3.50	Instructional Expenditure per Student	4/7.25
Average SAT Score	5/5.22	% of Students Whose Parents Have Post-High School Degree	4/9.23
Degree Completion Rate	5/6.10	Out-of-State Tuition Fee	4/10.18
% of Asian Students	5/7.22	% of Students Whose Parents were 1st Generation College Student	4/10.33
% of Students Whose Parents Have a High School Degree	5/8.58	% of 1st Gen. College Students	4/10.63
In-State Tuition Fee	5/10.88	% of Students whose Family Income classified Very High	4/11.30

**Votes Column: The number of filter methods that selected the corresponding attributes; Average Rank Column: The averaged rank values among the filter methods that selected the corresponding attributes.*

Table 2. Selected attribute subset using forward selection

Attributes	Votes*	Attributes	Votes*
Predominant Degree Type	90%	% of Students whose Parents were 1st Generation College Student	60%
Ratio between Part-time and Full-time Students	100%	% of the Population from Students' Zip Codes over 25 with a Professional Degree	100%
Degree Completion Rate	100%	% of Female Students	100%
Admission Rate	100%	Average Age of Entering College	100%
% of Asian Students	100%		

**Votes Column: The percentage of folds that selected the corresponding attributes.*

The Genetic Algorithm (GA) was the third option for attribute selection. The settings of the GA were as follows:

- Population size: 500
- Fitness function: Classification accuracy derived from Logistic Regression
- Selection Method: Tournament selection
- Crossover Type: Two-point crossover
- Crossover Rate: 0.6
- Mutation Rate: 0.03
- Stopping Criteria: 60 generations.

Table 3. Selected attributes subset using genetic algorithm

Attributes	Votes*	Attributes	Votes*
School Type	60%	% of Asian Students	100%
Predominant Degree Type	70%	% of Hispanic Students	100%
Student Size	100%	% of Students whose Family Income classified Higher Middle	80%
Instructional Expenditure per Student	90%	% of Students whose Family Income Classified Very High	100%
Ratio between Part-time and Full-time Students	100%	% of Students whose Parents have a Middle School Degree	70%
Degree Completion Rate	100%	% of Students whose Parents have a Post-High-School Degree	60%
Admission Rate	100%	% of Population from Students' Zip Codes over 25 with a Professional Degree	100%
Average SAT Score	90%	% of Female Students	100%
Out-of-State Tuition	100%	% of 1st Generation Students	60%
% of White Students	90%	Average Age of Entering College	100%
% of Black Students	60%	Average Debt	70%

**Votes Column: The percentage of folds that selected the corresponding attributes.*

Logistic Regression and Support Vector Machine with Pearson VII function kernel were used to compare the performance of the three selected attribute subsets. Ten-fold cross validation was used to estimate the classification accuracy for each classification method (see Tables 4 and 5 for individual classification results). As the most selective feature selection method (*9 attributes selected*), Forward Selection achieved acceptable F-measure. Although less selective (*22 attributes selected*), Genetic Algorithm outperformed the other two methods by both F-measure and accuracy.

Table 4. Comparisons among three selected attribute subsets using logistic regression

Attribute numbers	Accuracy	Weighted average		
		Precision	Recall	F-measure
Filter methods (N = 13)	0.691	0.688	0.691	0.686
Forward selection (N = 9)	0.736	0.733	0.736	0.731
Genetic algorithm (N = 22)	0.746	0.746	0.746	0.745

Table 5. Comparisons among three selected attribute subsets using support vector machine with Pearson VII function kernel.

Attribute numbers	Accuracy	Weighted average		
		Precision	Recall	F-measure
Filter methods (N = 13)	0.708	0.697	0.708	0.701
Forward selection (N = 9)	0.733	0.723	0.733	0.726
Genetic algorithm (N = 22)	0.755	0.745	0.755	0.747

4 Conclusion

Using College Scorecard data [10], we selected the most important factors predicting the 6-year post-graduation income of college students who used financial aid during their time at college. We compared three attribute selection methods: filter methods, forward selection, and Genetic Algorithm, in terms of classification accuracy on students' post-graduation income. We found that the attribute subset selected by the Genetic Algorithm outperformed the other two subsets when using logistic regression and support vector machine as the classification algorithm.

We wish to draw attention to how higher numbers of students in a cohort who grew up in Zip code areas where over 25% of the population hold a Professional Degree was predictive of more college graduates likely being classified as High income. This finding is aligned with evidence about how geography or "where you grow up" impacts life outcomes. Chetty et al. [15] identified that areas with lower racial segregation and income inequality, but higher social capital[1] and family stability are associated with greater opportunities for intergenerational social mobility. In the current research, the role of geography for post-graduation incomes in the case of neighborhood Professional Degree attainment signifies social stratification in graduate labor markets. The finding may stem from unequal access to support for education and careers. This would reinforce the Effectively Maintained Inequality model that predicts that as access to education widens, higher socio-economic status students will seek "horizontal differentiation" by accessing *qualitatively* distinctive or superior types of education that maintain their advantage in society [17, 18].

We are *not* arguing that young people from disadvantaged neighborhoods should not attend higher education. Attaining a Bachelor's degree remains an excellent "investment" to enhance career prospects. Yet our findings showing a disparity of post-graduation income according to "where you grow up" suggests a need for greater support for students *both* in college access and in transitions to the labor market, especially given rising tuition fees and associated concerns about student debt [19].

Appendix A

The dataset analyzed in this study can be accessed at https://collegescorecard.ed.gov/data/.
30 potential attributes include:
Group One: School information

1. School Type (e.g. private school)
2. Predominant Awarded Degrees (e.g., bachelor degree)
3. Student Size
4. Instructional Expenditure per Student

[1] Social capital represents trust, solidarity, and reciprocity in collective social interactions and engagement in community-based activities [16].

5. Ratio between Part-time and Full-time Students
6. Degree Completion Rate
7. Average Faculty Salary

Group Two: Admission information

8. Admission Rate
9. Average SAT Score

Group Three: Cost information

10. In-State Tuition
11. Out-of-State Tuition

Group Four: Student information

12. Percentage of White Students
13. Percentage of Black Students
14. Percentage of Asian Students
15. Percentage of American Indian Students
16. Percentage of Hispanic Students
17. Percentage of Female Students
18. Percentage of First-Generation Students
19. Average Age of Entering College
20. Average Debt

Group Five: Family and community information

21. Percentage of Students whose Family Income was classified as Low
22. Percentage of Students whose Family Income was classified as Lower Middle
23. Percentage of Students whose Family Income was classified as Higher Middle
24. Percentage of Students whose Family Income was classified as High
25. Percentage of Students whose Family Income was classified as Very High
26. Percentage of Students whose Parents were 1st Generation College Student
27. Percentage of Students whose Parents Have a Middle School Degree
28. Percentage of Students whose Parents Have a High School Degree
29. Percentage of Students whose Parents Have a Post-High-School Degree
30. Population from Students' Zip Codes over 25% with a Professional Degree.

References

1. Autor, D.H.: Skills, education, and the rise of earnings inequality among the 'other 99 percent'. Science **344**(6186), 843–851 (2014)
2. Chetty, R., Friedman, J., Saez, E., Turner, N., Yagan, D.: Mobility report cards: the role of colleges in intergenerational mobility. Technical report, Stanford University (2017)
3. Hout, M.: Social and economic returns to college education in the United States. Ann. Rev. Sociol. **38**, 379–400 (2012)

4. National Center for Educational Statistics [NCES]. Percentage of 18- to 24-year-olds enrolled in degree-granting postsecondary institutions, by level of institution and sex and race/ethnicity of student: 1970 through 2015. http://nces.ed.gov/programs/digest/d15/tables/dt15_302.60.asp?current=yes. Accessed 1 Mar 2018
5. Beaudry, P., Green, D.A., Sand, B.M.: The declining fortunes of the young since 2000. Am. Econ. Rev. **104**(5), 381–386 (2014)
6. Valletta, R.G.: Recent flattening in the higher education wage premium: polarization, skill downgrading, or both? In: Education, Skills, and Technical Change: Implications for Future US GDP Growth. University of Chicago Press (2017)
7. Federal Reserve Bank of New York. The Labor Market for Recent College Graduates. https://www.newyorkfed.org/research/college-labor-market/index.html. Accessed 1 Mar 2018
8. Altonji, J.G., Arcidiacono, P., Maurel, A.: The analysis of field choice in college and graduate school: determinants and wage effects (no. w21655). National Bureau of Economic Research (2015)
9. Witteveen, D., Attewell, P.: The earnings payoff from attending a selective college. Soc. Sci. Res. **66**, 154–169 (2017)
10. U.S. Department of Education. https://www.newyorkfed.org/research/college-labor-market/index.html. Accessed 1 Mar 2018
11. U.S. Census Bureau: Distribution of Personal Income 2010 (2010). https://www.census.gov/2010census/data/. Accessed 1 Mar 2018
12. Beasley, J.E., Chu, P.C.: A genetic algorithm for the set covering problem. Eur. J. Oper. Res. **94**(2), 392–404 (1996)
13. Goldberg, D.: Genetic Algorithms in Optimization, Search and Machine Learning. Addison-Wesley, Reading (1989)
14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
15. Chetty, R., Hendren, N., Kline, P., Saez, E.: Where is the land of opportunity? The geography of intergenerational mobility in the United States. Q. J. Econ. **129**(4), 1553–1623 (2014)
16. Putnam, R.D.: Our Kids: The American Dream in Crisis. Simon and Schuster, New York (2016)
17. Lucas, S.R.: Effectively maintained inequality: education transitions, track mobility, and social background effects. Am. J. Sociol. **106**, 1642–1690 (2001)
18. Lucas, S.R., Byrne, D.: Effectively maintained inequality in education: an introduction. Am. Behav. Sci. **61**(1), 3–7 (2017)
19. Avery, C., Turner, S.: Student loans: do college students borrow too much—or not enough? J. Econ. Perspect. **26**(1), 165–192 (2012)

From Language to Location Using Multiple Instance Neural Networks

Sneha Nagpaul(✉) and Huzefa Rangwala

George Mason University, Fairfax, VA, USA
snagpaul@gmu.edu, rangwala@cs.gmu.edu

Abstract. Language patterns pertaining to a geographic region has various uses including cultural exploration, disaster response and targeted advertising. In this paper, we propose a method for geographically locating short text data within a multiple instance learning framework augmented by neural networks. Our representation learning approach tackles minimally pre-processed social media discourse and discovers high level language features that are used for classification. The proposed method scales and adapts to datasets relating to 15 cities in the United States. Empirical evaluation demonstrates that our approach outperforms state of the art in multiple instance learning while providing a framework that alleviates the need for subjective feature engineering based approaches.

Keywords: NLP · MIL · Text geolocation · Neural networks

1 Introduction

Due to privacy concerns, users of social media often chose not to share the geographic location while they generate content. Besides commercial and malicious uses such as targeted advertising and recommender systems [3], this information could also be used to facilitate better disaster response and help law enforcement [1] with crime prevention [8]. Thus, a system which geo-tags user generated text is valuable for its social applications.

Prior work on text geolocation frames the problem as classification of user discourse into regions based on words that appear in the text [9]. Since the phrase level structure is distorted by this Bag Of Words approach, these models often lose context because word order is lost. Additionally, the data requirements tend to move away from short text to body of text produced by a user. Hence, they end up predicting a user's location rather than locating a stand alone piece of content.

Hence, this is a problem where location is available for users rather than an individual tweet. This allows us to express the problem for distilling information from group level labels to individual parts within the group. This is referred to as multiple instance learning (MIL) and has found extensive use in semi-supervised learning and sentiment analysis. Since MIL research makes strong assumptions

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 46–53, 2018.
https://doi.org/10.1007/978-3-319-93372-6_5

about the membership of instances inside a bag and/or use feature engineering based approaches, there is scope to augment this work. This paper makes a contribution to the tractability and abstraction mechanisms employed within an MIL exercise.

The approach proposed in this work provides a flexible and scalable framework for transferring geographic location labels from user to tweet level without the use of explicit features or kernels. The flexibility is present in its modular structure that separates instance level predictions from bag level aggregations while still enabling a backward flow of information of labels from the aggregate to the individual instance level. Since theI underlying method is a neural network that can be trained using optimization techniques and learns internal representations in the datasets, it scales well to the size and types of datasets under consideration.

2 Related Work

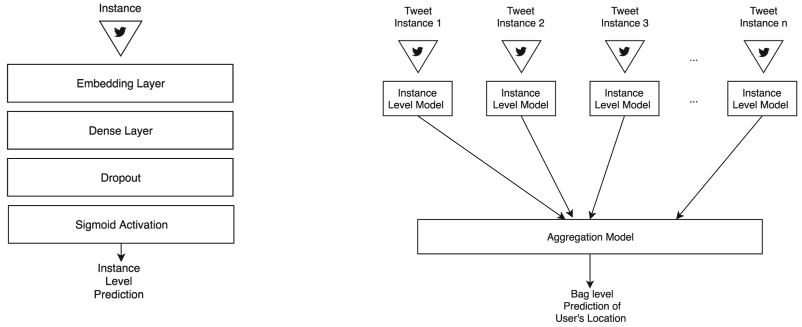
Multiple Instance Learning. Within the standard formulation, a group of instances referred as *bags* are labeled but individual instances are not [11] The bag level label is associated with its contents by a membership assumption and an aggregation function.

Single Instance Learning (SIL) is a naive and noisy way to accomplish this task [10] wherein every instance is assigned the label of its bag. Recently, neural networks have been used with adapted cost functions to accomplish the task of relaxing aggregation assumptions while using custom similarity measures [5]. Most of these prior methods are kernel based, as they require substantial feature engineering and are thus hard to scale. Additionally, in prior applications instances share heavy context, whereas tweets within a user-bag need not share context or even temporal origins.

Geographic Information Retrieval. Geographic Information Retrieval refers to methods that deal with mapping language to location [7]. Classically, a supplementary dataset or gazetteer, that maps words to locations along with heuristics to disambiguate place names was used. However as scale of datasets grew language modeling became prevalent in GIR which solves the problem for the user level with Bag of Words models using traditional bayesian techniques and using neural networks [7].

3 Methods

To overcome the gaps in prior work, we leverage basic feed forward neural network architectures like the multi layer perceptron (MLP) [6]. We make simple changes to this basic architecture to enable it to perform MIL with higher level of modular abstraction for instance level classification and bag level aggregation, as shown in Fig. 1b. We consider a user-bag, labeled with a binary location label which contains tweet-instances that are devoid of labels at the training stage.



(a) Instance Level Model

(b) Model Architecture

Fig. 1. (a): Instance level model: the model for each tweet and (b) model architecture: to achieve MIL, the instance level models feed their predictions to a bag level aggregation layer to be able to share the weights from retro-propagated losses.

3.1 Problem Statement

Given a user $\mathrm{U_i}$ with a binary location label $\mathrm{y_i} \in \{0, 1\}$ where 1 denotes that the user is from a particular city and 0 denotes otherwise. Each $\mathrm{U_i}$ is a collection of tweets $\mathrm{t_{ij}}, \mathrm{j} = 1, 2, \ldots \mathrm{N}$ and the task is then to devise a function $\mathrm{f(t)} \rightarrow \mathrm{y}$ which essentially labels individual tweets as belonging to the city under consideration.

For a treatment of the problem as formulated here, an end-to-end trainable neural network architecture is proposed in this work and is called milNN. The model's architecture is illustrated in Fig. 1a and b.

Instance Level Classifier. The tweet level classifier consists of an embedding layer that feeds into the fully connected hidden layer component and is designed as though labels were available (Fig. 1a). The embedding layer learns representations that can be viewed as an intuitive language model as opposed to a symbolic language model that stems from rigid grammatical rules or engineered features [6]. This also makes the model well suited to social media content which often deviates from traditional language use.

Bag Level Classifier. The instance level classifier is then applied to individual tweets and the results are averaged to get the bag level labels as shown in Fig. 1b. This is the component of the architecture that addresses the relationship between bag and instance level labels.

Loss Function and Training. At this stage a label is available and losses (binary cross entropy) can be back-propagated (Adam Optimizer [4]) throughout the network. Thus, the instance level classifier gets trained as a result of gradients of the bag level losses.

3.2 Method Characteristics

Due to being fully embedded in the neural network and representation learning paradigm, milNN relies on learning distributed representations and is devoid of subjective feature engineering requirements. This also equips it to handle any changes that might occur organically in the data. Moreover, this framework does not have high computational and memory requirements and learned using stochastic gradient descent which is easy to parallelize.

Additionally, the assumptions of membership proportions are relaxed and aggregation assumptions are not particularly stringent as the sigmoid layer does not provide exact labels for the instances, but rather a probabilistic average across all tweet classifications outputs for a user level label. The architecture is also flexible and the model described here can be seen as one example of the most basic possibilities.

The datasets were created by reverse geocoding information from Twitter North America dataset [12]. The latitude and longitude readings were recorded when the user registered on Twitter and provided the location. Subsequent tweets were recorded for this user. For use in this work, the top cities in the data were split into fifteen datasets of equal number of positive and negative samples. The negative samples for each city were randomly selected from the rest of the dataset after stratified sampling from other cities.

3.3 Experimental Setup

At the instance level, the tweets are preprocessed by changing URLs, @mentions, and hashtags to a generic tags for each. Subsequently they are tokenized and vocabulary size is chosen to be 5000. The tweet is then padded to a 20 word maximum and then fed through an embedding layer with 32 dimensions which is randomly initialized. Following this, there is a single hidden layer with 100 nodes that process the various language level relationships and feed the relu activations to the sigmoid layer for classification after adding a dropout of 25% for regularization. 10 tweets from each user are averaged from the instance model at a higher layer for the bag level output. At this level binary cross entropy loss is calculated using the bag level labels and back-propagated using the Adam optimizer. A batch size of 256 bags at a time is chosen and trained for 200 epochs with a learning rate of 0.0001. An early stopping condition is included which breaks out of training when the loss of the epoch converges and waits for 5 iterations to confirm the convergence. The hyper-parameters chosen for the model are described in this section and were chosen using half the training data for validation. The performance of all considered choices was comparable except for a running time increase for models with more parameters.

3.4 Results

As seen in Table 1, in terms of the accuracy metric milNN outperforms state of the art on 14 of the 15 datasets considered here. When considering the F-score, it outperforms the other methods on 10 of the 15 datasets. It is important to notice that when it loses to older methods, it is for smaller datasets and not by a lot of margin. However, when it outperforms it is significantly better (eg. Boston performance is better by 20%). Also, it is consistently good on datasets of varied sizes and needs while the other methods don't seem to be able to adapt to the feature requirements and scale of the data.

Table 1. Accuracy and F-scores for milNN and prior methods. milNN scores a 14/15 and 10/15 on Accuracy and F-score respectively on the 15 datasets of varied sizes (train-test split was 80:20)

City	Acc:SIL	Acc:GICF	Acc:milNN	F1:SIL	F1:GICF	F1:milNN	Total
Atlanta	0.5780	0.6025	**0.6602**	0.6900	**0.6982**	0.6568	4414
Austin	0.6070	0.6501	**0.7015**	0.6650	0.6291	**0.6848**	2915
Baltimore	0.5180	0.6248	**0.6858**	0.6560	**0.6957**	0.6816	2700
Boston	0.5460	0.5774	**0.6276**	0.5820	0.5960	**0.6130**	2389
Chicago	0.5760	0.6429	**0.6502**	0.5180	0.5163	**0.6420**	8286
New Orleans	0.5230	0.6365	**0.6962**	0.6690	0.6976	**0.7041**	2592
New York City	0.5740	0.6476	**0.7024**	0.6230	0.6349	**0.6988**	19000
Paradise	0.6060	**0.6629**	0.6565	**0.6510**	0.6365	0.6468	3095
Philadelphia	0.5190	0.6195	**0.6644**	0.6620	**0.7006**	0.6830	5792
San Diego	0.6300	0.6504	**0.6850**	0.6080	0.6325	**0.6548**	2452
San Francisco	0.6300	0.7322	**0.7542**	0.6980	0.7398	**0.7603**	7710
Seattle	0.6210	0.6970	**0.7269**	0.6840	0.7112	**0.7215**	3350
Toronto	0.6390	0.6895	**0.7520**	0.6810	0.7056	**0.7485**	5037
Washington, D.C.	0.5740	0.6298	**0.6437**	**0.6190**	0.4910	0.6108	5732
Weehawken	0.6160	0.6727	**0.7000**	0.6590	0.6588	**0.6827**	2196

3.5 Case Study

Since San Francisco was the biggest dataset and milNN outperformed the state of the art on all counts due to its scalability. The structure of the following analysis is to go over examples of users that belong to SF for exploring language patterns. Additionally the anti-patterns are explored by analyzing tweets from users that do not belong in order to discover how SF users don't tweet.

The word cloud visual is created using all the test instance tweets that were classified to be over 0.95 by the instance level classifier. Location entities were subsequently extracted from these tweets using StandfordNER [2] and then weighted into a word cloud (Fig. 2).

While the San Francisco model caught traditional place names such as 'Alcatraz' and 'San Franciso', more exotic language patterns were also discovered in this dataset. For the second user that was from SF, technology related tweets were identified as indicative of the region. Additionally, a proclivity to tweet with correct grammar is also discovered when a single word change causes probability to increase as grammatical usage becomes less awkward. This point is further reinforced when a user that is not from SF is seen to have no high probability tweets due to use of slang (Fig. 3).

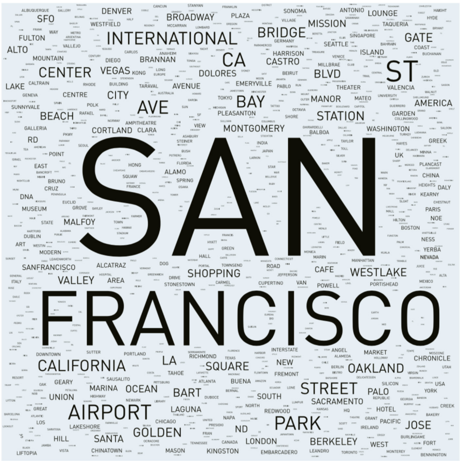
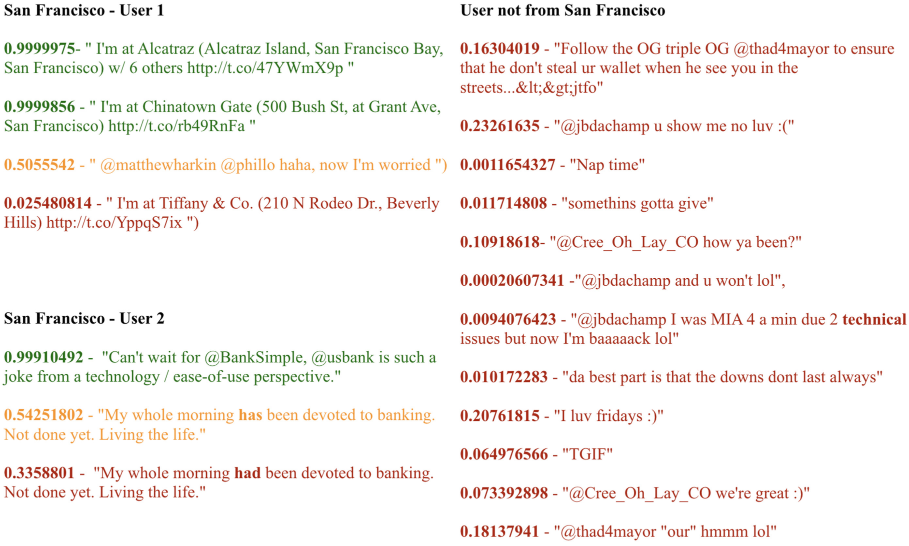


Fig. 2. Word cloud - SF


San Francisco - User 1

0.9999975- " I'm at Alcatraz (Alcatraz Island, San Francisco Bay, San Francisco) w/ 6 others http://t.co/47YWmX9p "

0.9999856 - " I'm at Chinatown Gate (500 Bush St, at Grant Ave, San Francisco) http://t.co/rb49RnFa "

0.5055542 - " @matthewharkin @phillo haha, now I'm worried ")

0.025480814 - " I'm at Tiffany & Co. (210 N Rodeo Dr., Beverly Hills) http://t.co/YppqS7ix ")

San Francisco - User 2

0.99910492 - "Can't wait for @BankSimple, @usbank is such a joke from a technology / ease-of-use perspective."

0.54251802 - "My whole morning **has** been devoted to banking. Not done yet. Living the life."

0.3358801 - "My whole morning **had** been devoted to banking. Not done yet. Living the life."

User not from San Francisco

0.16304019 - "Follow the OG triple OG @thad4mayor to ensure that he don't steal ur wallet when he see you in the streets...&lt;&gt;jtfo"

0.23261635 - "@jbdachamp u show me no luv :("

0.0011654327 - "Nap time"

0.011714808 - "somethins gotta give"

0.10918618- "@Cree_Oh_Lay_CO how ya been?"

0.00020607341 -"@jbdachamp and u won't lol",

0.0094076423 - "@jbdachamp I was MIA 4 a min due 2 **technical** issues but now I'm baaaaack lol"

0.010172283 - "da best part is that the downs dont last always"

0.20761815 - "I luv fridays :)"

0.064976566 - "TGIF"

0.073392898 - "@Cree_Oh_Lay_CO we're great :)"

0.18137941 - "@thad4mayor "our" hmmm lol"

Fig. 3. San Francisco examples

4 Conclusion

This work demonstrated a flexible multiple instance learning framework applied to identifying geographic location of short text data. The experiments showed that milNN was scalable and capable of discovering high level language features such as grammar in addition to place names in data. Thus this work contributed to multiple instance learning and geographic information retrieval literature by designing a novel model architecture that was end-to-end trainable.

While recurrent neural networks and attention mechanisms might have lent themselves to the problem considered here, we chose to focus this exercise on MIL in order to augment prior research that has suffered from the intractability of kernel based methods. Given the flexibility of the neural network architecture, future work could focus on developing recurrent neural network based models that can take a variable length input both in terms of tweet length and number of tweets. Additionally, extensions of standard aggregation functions could be developed for instance level data that have bag-internal structure that needs to be exploited, like in reviews. Transfer learning approaches could also be explored given the embedding based input of the neural network as and when reliable twitter word vectors become available.

Acknowledgment. This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

References

1. Ashktorab, Z., Brown, C.D., Nandi, M., Culotta, A.: Tweedr: mining Twitter to inform disaster response. In: ISCRAM (2014)
2. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, pp. 363–370. Association for Computational Linguistics, Stroudsburg (2005). https://doi.org/10.3115/1219840.1219885
3. Ho, S.S., Lieberman, M., Wang, P., Samet, H.: Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2012, pp. 25–32. ACM, New York (2012). http://doi.acm.org/10.1145/2442810.2442816
4. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014). http://arxiv.org/abs/1412.6980
5. Kotzias, D., Denil, M., de Freitas, N., Smyth, P.: From group to individual labels using deep features. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2015, pp. 597–606. ACM, New York (2015). http://doi.acm.org/10.1145/2783258.2783380
6. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–44 (2015)
7. Melo, F., Martins, B.: Automated geocoding of textual documents: a survey of current approaches. Trans. GIS **21**(1), 3–38 (2016). https://doi.org/10.1111/tgis.12212
8. Ning, Y., Muthiah, S., Rangwala, H., Ramakrishnan, N.: Modeling precursors for event forecasting via nested multi-instance learning. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2016, pp. 1095–1104. ACM, New York (2016). http://doi.acm.org/10.1145/2939672.2939802

9. Rahimi, A., Cohn, T., Baldwin, T.: A neural model for user geolocation and lexical dialectology. CoRR abs/1704.04008 (2017). http://arxiv.org/abs/1704.04008
10. Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of the 22nd International Conference on Machine Learning, ICML 2005, pp. 697–704. ACM, New York (2005). http://doi.acm.org/10.1145/1102351.1102439
11. Foulds, J., Frank, E.: A review of multi-instance learning assumptions. Knowl. Eng. Rev. **25**(1), 1–25 (2010)
12. Roller, S., Speriosu, M., Rallapalli, S., Wing, B., Baldridge, J.: Supervised text-based geolocation using language models on an adaptive grid. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2012, pp. 1500–1510. Association for Computational Linguistics, Stroudsburg (2012). http://dl.acm.org/citation.cfm?id=2390948.2391120

Detecting Agreement and Disagreement in Political Debates

Mahboubeh Ahmadalinezhad(✉) and Masoud Makrehchi

University of Ontario Institute of Technology, Oshawa, ON L1H 7K4, Canada
{mahboubeh.ahmadalinezhad, masoud.makrehchi}@uoit.ca

Abstract. In this paper, the task of agreement/disagreement detection in political debates is studied. The main goal of this study is to detect agreement/disagreement between two individuals on a topic based on their conversations. This is a challenging task due to the lack of annotated corpora in this field. A self-labeling method is introduced for data collection and generating the training data. A new approach based on text classification is proposed for this task. The experimental results on Canadian Parliamentary debates and the United State 1960 Presidential Campaign datasets have proven the efficiency of the developed methodology and outperforms the baseline methodologies. In addition, the validity of the proposed self-labeling method is evaluated, and its efficiency is confirmed.

Keywords: Sentiment analysis · Agreement/disagreement detection · Text classification · Political analytics

1 Introduction

Recently, analyzing social interactions and mining public opinions have attracted a great deal of attention due to its practical applications in providing better services to the users. A number of studies have focused on mining for dispute in online interactions [3,9]. One of the main argumentative dataset to analyze both agreement/disagreement is the political debates, which contain official and unofficial documents. Detecting agreement/disagreement in political debates in the US congress has been studied [2,13]. In these methods, determining the single stance of a debate participant with respect to a specific topic was investigated.

In this paper, the goal is to introduce a new approach to generate training data for political analytics and introduce a methodology to detect agreement/disagreement in political debates. The cost of gathering such a dataset is not very high, but handling this kind of information is difficult and requires further development.

In Sect. 2, related works are presented. The problem statement and the notion of agreement/disagreement are defined in Sect. 3. A new methodology is presented in Sect. 4. In Sect. 5, further details about collecting the data and labeling are discussed. Implementation results are demonstrated in Sect. 6. Finally, conclusion and future works are mentioned in Sect. 7.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 54–60, 2018.
https://doi.org/10.1007/978-3-319-93372-6_6

2 Related Works

Two major categories of methods are used to classify political statements as supporting/opposing for a debated topic.

The first category are those approaches that utilize the common information of the text structure, in which the focus is on sentiment analysis and contextual information. Somasundaran and Wiebe [12] proposed an approach, which classifies a stance as approve or disapprove about a debated topic.

In [1], determining disagreement in online political forums between a pair of quoted text and a given response is studied. Anand et al. [2] improved the results of unigram and classification for various topics using contextual information and opinion dependencies. To classify controversial discussion topics on the political domain, an LM-based method is proposed [4]. In [6], U.S Congressional floor debate transcripts are used as a dataset and sentiment classification is applied to determine agreement/disagreement.

The second category is based on corpus-specific features. In [13], a new variant of Latent Semantic Analysis (LSA) is proposed to detect the support and opposition to legislation in congressional debates using information such as speech transcriptions, records on voting, and the relation between the speakers. Moreover, some approaches are proposed to identify agreement/disagreement in consecutive speech transcription segments. Different speakers talk either positively or negatively against the discussed topic by using lexical, structural, and prosodic features [7].

As opposed to the previous studies, in which the opinion of one speaker about a specific topic is investigated and his agreement/disagreement is detected, the objective of this study is to detect the type of the interaction between two speakers regardless of the topic.

3 Problem Statement

In Canadian Parliamentary, parties are categorized into two groups: governing party and opposition party. In this study, there are two main assumptions: first, a representatives of the governing party and a representative from the opposition party disagree on a topic, and second two representatives of the same party agree on the topic. In addition, it is assumed that each conversation between two individuals is a document. In this scenario, first, the document collections are processed and the text are extracted to compared and classified. Second, an approach is proposed to classify each pair of conversations based on supervised learning, which considers the features capturing the relevant dimensions.

4 The Proposed Method

The major problem in an agreement/disagreement task is to represent the conversation between two individuals. The core idea in the proposed approach is conversation modeling using Bag-of-Words representation of interpolation discussions. Three different operators are used to build a document representation and apply a text classifier such as Support Vector Machine to train the prediction model.

The objective is to learn the features for the automatic detection of agreement/disagreement that would provide useful information about the conversations between people without knowledge of their topic. We focus on oral speech, which has less information in comparison to written text such as punctuation, non-lexical features, and time between posts, etc. Since we work on Hansard - the printed version of what members of Parliament expressed in the House of Commons- we also lose utterance information of the oral conversations. In addition to the main proposed method, three other methods based on similarity and sentiment analysis are also implemented and compared to the text classification methods.

1. **Text classification:** Each document is considered separately and two vectors for each document are computed. Each document is converted to a fixed-size representation to be used as an input to the classifier. Three different operators are applied to interpolate a document for conversation modeling:
   Concatenation operator: The conversation between two members is represented by concatenation of two vectors. The length of the document is *2n.*
   OR operator: This operator is used to represent the conversation by using this operator, the length of the document is n.
   AND operator: The conversation is represented by applying AND operator. The length of the document after using AND operator is n.
2. **Lexicon based analysis (Sentiment):** In order to implement the sentiment analysis of a conversation, the Linguistics Inquiry Word Count tool (LIWC-2001) [11] is utilized.
3. **Cosine similarity:** This measure is used as a metric to compute similarity between two documents [14].
4. **Cosine similarity and Lexicon based analysis:** Both of them are combined and considered as features for a document.

5 Dataset

The proposed method is evaluated on three different datasets.

Parliament of Canada: The first one is the debates of Parliament of Canada are collected from January to May 2016. The data includes 55 debates and more than 5000 documents. Conservative Party, Liberal Party, and New Democratic Party are the three major political parties in Canada.

To analyze effectiveness of these assumptions, two other datasets are considered which have been annotated by independent annotators using the CrowdFlower crowd sourcing.

1960 Presidential Campaign Dataset: The transcription of discourses and official declarations issued by Nixon and Kennedy during 1960 presidential campaign are collected. This data includes 881 documents.

Table 1. Results of determining agreement/disagreement by running various methods on Parliament of Canada dataset

Methods	Accuracy	F-score
Sentiment	0.48	0.37
Cosine similarity	0.52	0.41
Sentiment and similarity	0.53	0.44
Text classification with concatenation	**0.81**	**0.80**
Text classification with OR	0.68	0.67
Text classification with AND	0.63	0.63

Extended 1960 Elections: This data is extended version of the second dataset includes 1,400 pairs.

6 Results

A Support Vector Machine (SVM) [10] is used to train a model. The results are based on 10-fold cross-validation and the average prediction accuracy and F-scores are reported for all experiments.

6.1 Classification and Self-labeling for Parliament of Canada Data

The classification accuracy is shown in Table 1, where the results are demonstrated for all methods. The accuracy of the proposed method, text classification, is significantly improved in comparison to the others. An interesting point about the text classification method is the reduction in the difference between accuracy and F-score. For other methods, the results of accuracy is 10% higher than the results of F-score. This observation shows a strong capability of text classification in detecting both agreement and disagreement in a conversation.

6.2 Evaluation the Sensitivity of the Classification to the Amount of the Training Data

In order to evaluate the sensitivity of our models, the proposed method is evaluated on varied percentage of the training data. This is done to investigate the effect of a change in the percentage of training and its impact on the results. Since the best result was obtained with text classification using concatenation operator, its sensitivity to the training percentage against the test sample is investigated. The results are reported in Fig. 1 which is proving the fact that the proposed approach can detect the agreement/disagreement without much dependency on the training percentage.

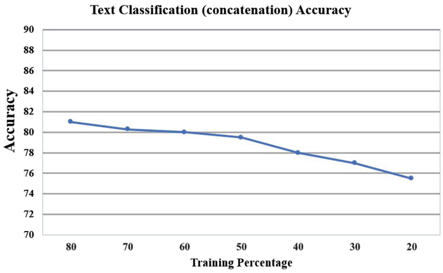


Fig. 1. Results of text classification (concatenation) at different training percentage on Parliament of Canada dataset

Table 2. Results of text classification (concatenation) on annotated dataset

Methods	1960 Elections	Extended 1960 Elections
Menini and Tonelli [8]	0.83	0.80
Self-labeling (Sect. 6.3)	0.79	0.74
Text classification (Sect. 6.4)	0.87	0.93

6.3 Evaluation the Proposed Self-labeling Method

Furthermore, the proposed self-labeling is evaluated. Therefore the US 1960 Presidential Campaign dataset is used [8] which is an annotated dataset and the transcription of discourses during the campaign. In this scenario, the proposed method (Text classification with concatenation) is run on this dataset. The model is constructed based on the two main assumptions and self-labeling, however, test phase is evaluated based on the goal labels which achieved 79% accuracy. The goal of this experiment is to test the proposed self-labeling method and the strength of the training model. The results are compared to [8] which uses negation/overlap, entailment, sentiment, cosine, word embeddings as features. According to Table 2 which is confirmed that the proposed self-labeling method works properly.

By applying transfer learning, we achieved 60% accuracy. The difference in the accuracy from the literature may be the result of the two political domains and the fact that language used has changed between 1960 and 2016.

6.4 Evaluation the Efficiency of Text Classification Method

In addition, the efficiency of the text classification approach by using interpolation compares to the proposed method of [8]. According to the Table 2 the results of text classification approach approve the efficiency of text classification method by using concatenation operator. We can conclude that our approach is a reliable solution to the task of detection both agreement and disagreement.

6.5 Domain Adaptation and Transfer Learning

In this experiment, transfer learning methods are applied and the training model is updated with the data which achieved high probability of predictions. By using transfer learning [5], the effort for annotating reviews for each document can be reduced, and the model which is based on training documents is used to learn classification models of other datasets. In this case, transfer learning can save a significant amount of labeling effort. The Parliament of Canada debates are considered as training data and the US 1960 Elections as test data. In each iteration, samples with high probability are added to improve the model. By applying transfer learning, we achieved 60% accuracy. The difference in the accuracy from the literature may be the result of the two political domains and the fact that language used has changed between 1960 and 2016.

7 Conclusion

In this paper, detection of agreement/disagreement in Canada's Parliament debates and the US 1960 election datasets were investigated. The detection was done by inputing the conversations and determining the agreement/disagreement of two individuals without respect to the topic. The input data were the oral debates between the parties of the Parliament without written information and utterance, which makes it a challenging task to detect the agreement/disagreement. As the data is not annotated, the data labeling was done based on two main assumptions: a representative of the governing party and a representative of the opposition party disagree on a topic, and two representatives of the same party agree on the topic. A new method was introduced for data collection and a novel algorithm based on classical text classification was proposed to detect agreement/disagreement. Different classification methods and different types of interpolation were examined and text classification with concatenation operator was found to be the best one by 81% accuracy.

Moreover, validity of two main hypotheses of this study was investigated. The US 1960 Presidential Campaign dataset was used to evaluate the assumptions, is labeled manually. We achieved 79% accuracy which proves efficiency of the proposed self-labeling and two assumptions.

In addition, text classification method is applied to the US 1960 Presidential Campaign dataset and observed significant improvement in comparison to proposed features and classifier of [8]. Overall, 87% accuracy is attained for the US 1960 elections and 93% for extended the US 1960 elections. Furthermore, semi-supervised learning and applying domain adaptation achieve acceptable results in comparison to the previous work. In the future, we also want to use other methods such as skip-gram to represent a document and apply advanced text classification algorithms.

References

1. Abbott, R., Walker, M., Anand, P., Fox Tree, J.E., Bowmani, R., King, J.: How can you say such things?!?: recognizing disagreement in informal political argument. In: Proceedings of the Workshop on Languages in Social Media, pp. 2–11. Association for Computational Linguistics (2011)
2. Anand, P., Walker, M., Abbott, R., Tree, J.E.F., Bowmani, R., Minor, M.: Cats rule and dogs drool!: classifying stance in online debate. In: Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, pp. 1–9. Association for Computational Linguistics (2011)
3. Andreas, J., Rosenthal, S., McKeown, K.: Annotating agreement and disagreement in threaded discussion. In: LREC, pp. 818–822. Citeseer (2012)
4. Awadallah, R., Ramanath, M., Weikum, G.: Language-model-based pro/con classification of political text. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 747–748. ACM (2010)
5. Blitzer, J., Dredze, M., Pereira, F., et al.: Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: ACL, vol. 7, pp. 440–447 (2007)
6. Burfoot, C.: Using multiple sources of agreement information for sentiment classification of political transcripts. In: Australasian Language Technology Association Workshop, vol. 6, pp. 11–18 (2008)
7. Galley, M., McKeown, K., Hirschberg, J., Shriberg, E.: Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, p. 669. Association for Computational Linguistics (2004)
8. Menini, S., Tonelli, S.: Agreement and disagreement: Comparison of points of view in the political domain. In: Proceedings of the 26th International Conference on Computational Linguistics, pp. 2461–2470 (2016)
9. Misra, A., Walker, M.A.: Topic independent identification of agreement and disagreement in social media dialogue. In: Conference of the Special Interest Group on Discourse and Dialogue, p. 920 (2013)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**(Oct), 2825–2830 (2011)
11. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001, vol. 71. Lawrence Erlbaum Associates, Mahway (2001)
12. Somasundaran, S., Wiebe, J.: Recognizing stances in ideological on-line debates. In: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 116–124. Association for Computational Linguistics (2010)
13. Thomas, M., Pang, B., Lee, L.: Get out the vote: determining support or opposition from congressional floor-debate transcripts. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 327–335. Association for Computational Linguistics (2006)
14. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Intell. Res. **37**, 141–188 (2010)

Tipping Points for Norm Change in Human Cultures

Soham De[1,2(✉)], Dana S. Nau[1,2], Xinyue Pan[3], and Michele J. Gelfand[3]

[1] Department of Computer Science, University of Maryland, College Park, USA
{sohamde,nau}@cs.umd.edu
[2] Institute for Systems Research, University of Maryland, College Park, USA
[3] Department of Psychology, University of Maryland, College Park, USA
{xypan,mgelfand}@umd.edu

Abstract. Humans interact with each other on a daily basis by developing and maintaining various social norms and it is critical to form a deeper understanding of how such norms develop, how they change, and how fast they change. In this work, we develop an evolutionary game-theoretic model based on research in cultural psychology that shows that humans in various cultures differ in their tendencies to conform with those around them. Using this model, we analyze the evolutionary relationships between the tendency to conform and how quickly a population reacts when conditions make a change in norm desirable. Our analysis identifies conditions when a tipping point is reached in a population, causing norms to change rapidly.

Keywords: Cultural evolution · Human behavior · Social norms
Norm change · Evolutionary Game Theory

1 Introduction

Social norms are critical in enabling human populations across the world to coordinate and accomplish different tasks. The *strength* of these social norms, however, differs widely around the globe, as has been established by past neuroscience, field and experimental research [9,12,13,16,17,32,38]. Some cultures are said to be *tight*, with strong social norms, typically characterized by high degrees of norm adherence and strict punishment directed towards norm-violators. Other cultures are said to be *loose*, with weaker norms characterized by a higher acceptance of deviant behavior [13,17,38].

Tightness-looseness is a dynamic construct, yet to date, there has been little research on the evolutionary processes that lead to *changes* in societal norms, the *rate* at which such changes occurs, and how these processes *vary* across different cultures. In this paper, we aim to study how cultural differences in the way humans interact and influence each other heavily influence how societal norms are established and the rate at which they change across the world. We use evolutionary game theory (EGT) to examine the causal relationship between an

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 61–69, 2018.
https://doi.org/10.1007/978-3-319-93372-6_7

individual's tendency to conform with those around them and the rate at which norms are changed in different cultures (see Sect. 2 for a brief discussion on EGT). More specifically, our primary contributions in this paper are as follows:

- Drawing on recent research in cultural psychology, we propose a game-theoretic model of a culture based on the tendency of an individual to conform with others, vs. being more individualistic in their behavior.
- Using this model, we provide conditions under which a population is open to changing the current norm in a society, depending on the pressure of conformity and the abruptness of the boundary between following and violating the norm.
- Finally, we analyze the *rate* at which such norm changes occur. We show that tighter cultures are more likely to be initially resistant to norm changes compared to looser cultures. Further, we analyze conditions under which tighter cultures sometimes reach a *tipping point*, where a large proportion of the population suddenly switch to a new norm.

The rest of the paper is organized as follows. Section 2 provides background and related work. We introduce our proposed model in Sect. 3 and study the rate of norm change in Sect. 3.2. In Sect. 4 we discuss the significance of our results.

2 Background and Related Work

Evolutionary Game Theory (EGT) was initially proposed to model biological evolution [23,39,40], but has been increasingly used to study human cultural evolution. The idea is to represent an interaction among individuals as a normal-form game, where individuals can use different strategies. The game's payoffs represent an individual's evolutionary fitness. In EGT models of cultural evolution, biological reproduction represents *social learning*: individuals are more likely to adopt strategies from others that produce high fitness and thus strategies that lead to higher fitness become more prevalent over time. While such models use highly simplified abstractions of complex human interactions, they aim to capture the essential nature of the interactions of interest, and thus have been increasingly used to study a wide variety of social and cultural phenomena, such as, cooperation and altruism [2,14,33,34,36], punishment [3–5,35,37], trust and reputation [11,19,27], ethnocentrism [8,15,18], etc.

There has been widespread interest in studying the *emergence* of social norms in a population both from an evolutionary perspective [1,20,21,31,41], as well as in empirical research [6,25,26]. There has been, however, much less work done on understanding the processes that lead to *change* in an already established norm in a population. A related concept, the propagation of information in social networks, has been well-studied (see [7,10,24] for an overview), but these works typically do not account for the differences in how individuals interact and influence each other in different cultures. Data science approaches have also explored this question, however, it is very challenging to separate out the various confounding factors (such as institutional influence) to establish clear causal relationships [28–30,42]. Finally, note that [22] and [9] have previously studied the processes of norm change in societies, and our work extends these studies in important directions. Using a more general model of conformist transmission that depends on the degree to which there is an abrupt boundary between following and violating the norm, we study the *speed* of norm change in different cultures, and show how *tipping points* during such norm change depends critically on the model of conformist transmission.

3 Proposed Evolutionary Game-Theoretic Model

Research in cross-cultural psychology has established that the strength of social norms varies considerably across cultures. Further, using historical and ecological data, it has been shown that this is related to the degree of threat that populations face [13,17]. Stronger norms and sanctions are needed in high-threat situations to coordinate and survive, leading to tighter cultures. By contrast, populations that lack exposure to serious ecological threats can afford to have weaker norms and tolerance for deviance given that they have less need for coordinated social action (looser cultures). We now describe our model.

Consider an infinite, well-mixed population (i.e., each individual can interact with any other individual in the population) that evolves according to the well-known replicator dynamic [23]. For simplicity of presentation, suppose each agent may choose one of two possible actions: A and B (see Sect. 4 for a discussion on the assumptions used in our model). The two actions A and B correspond to possible norms that the society could settle on. Let x_A and x_B denote the proportions of the population using actions A and B respectively, with $0 \leq x_A, x_B \leq 1$ and $x_A + x_B = 1$, and let $x = (x_A, x_B)$. According to the replicator dynamic, the rate of change in the proportions of agents using each action is given by the following differential equation:

$$\dot{x}_i = x_i[f_i(x) - \phi(x)], \tag{1}$$

where $i \in \{A, B\}$, $\dot{x}_i = dx_i/dt$ (i.e., rate of change of x_i), $f_i(x)$ is the fitness of action i, and $\phi(x)$ denotes the average fitness of the population, i.e.: $\phi(x) = x_A f_A(x) + x_B f_B(x)$. The replicator dynamic is based on the idea that the proportion of agents with a particular strategy increases when it achieves expected fitness higher than the average fitness, and vice versa.

Let u_A and u_B denote the payoffs associated with actions A and B, where $0 < u_A, u_B < 1$ and $u_A + u_B = 1$. To define the fitness function f_i, we use the key insight that in loose cultures, individuals tend to choose the action that is most beneficial to them; but in tight cultures, individuals tend to conform to the same action that others use, even if a different action might be more beneficial to each individual. To model this mathematically, we let f_i be a weighted combination of the payoff u_i and an additional *conformism fitness* measure θ_i that depends on whether the individual is conforming to others in the population. Let m denote the parameter controlling the weighting between these two fitness measures, i.e., the amount of conformist transmission in a population. Thus, we define f_i as:

$$f_i(x, m) = (1 - m)u_i + m\theta_i(x, k), \tag{2}$$

where $0 \le m \le 1$, and we define the conformism fitness measure θ_i as:

$$\theta_i(x, k) = \Big[1 + \exp\big(-k(x - 0.5)\big)\Big]^{-1}, \tag{3}$$

where $k > 0$. Note that we can vary the behavior of the conformism fitness measure θ_i using the parameter k (see Fig. 1). For example, when k is large, θ_i is close to a step function where there is an abrupt boundary between following and violating the norm and agents have a non-zero conformism fitness only if they conform with the majority action:

$$\theta_i^{\infty}(x) = \lim_{k \to \infty} \theta_i(x, k) = \begin{cases} 0, & \text{if } x_i < 0.5; \\ 0.5, & \text{if } x_i = 0.5; \\ 1, & \text{if } x_i > 0.5. \end{cases}$$

Note that with no conformism whatsoever ($m = 0$), each action's fitness depends solely on its payoff, i.e., typical of a very loose culture. On the other hand, with 100% conformist transmission ($m = 1$), i's fitness depends solely on the conformism fitness measure (for the case of θ_i^{∞}, this means that i's fitness depends solely on whether i is in the majority or the minority of the population). This is more indicative of a very tight culture. For simplicity, for the rest of the paper, we denote $\theta_i(x, k)$ as θ_i and $\theta_i^{\infty}(x)$ as θ_i^{∞}.

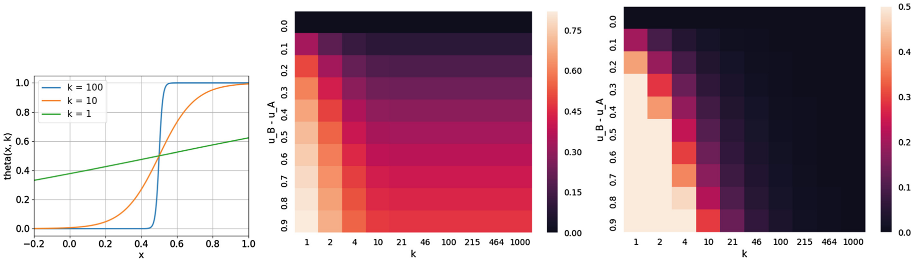


Fig. 1. *Left:* Plot of (3) for different values of k. *Middle:* Heatmap of the right-hand side in (6) when $x_B = 0.1$, for various $u_B - u_A$ and k values. *Right:* Heatmap of the right-hand side in (8), for various $u_B - u_A$ and k values. Best viewed in color. (Color figure online)

3.1 When Does Norm Change Occur?

Suppose norm B has a higher utility compared to A, i.e., $u_B > u_A$. We are interested in analyzing the conditions for which a population shifts from norm A to B (norm change). We can re-write the average fitness to be:

$$\phi(x) = (1 - m)(x_A u_A + x_B u_B) + m(x_A \theta_A + x_B \theta_B). \tag{4}$$

We are interested in anlayzing the rate of change in the proportion of B individuals. From (1), (2), and (4), we get:

$$\dot{x}_B = x_B(1 - x_B)\big[(1 - m)(u_B - u_A) + m(\theta_B - \theta_A)\big]. \tag{5}$$

Note that $\theta_B \geq \theta_A$ when $x_B \geq 0.5$. Since $u_B > u_A$, we see that $\dot{x}_B > 0$, i.e., x_B will converge to 1 ($\lim_{t\to\infty} x_B = 1$) when $x_B \geq 0.5$. If $x_A > x_B$, i.e., if the current norm in the population is A, norm change takes place only if:

$$m < \frac{u_B - u_A}{(u_B - u_A) + (\theta_A - \theta_B)}. \tag{6}$$

Thus, norm change takes place only if the population is loose enough, while tighter cultures are more resistant to change. Further, note that $\theta_A^\infty - \theta_B^\infty = 1$ when $x_A > x_B$. Thus, when the conformist fitness measure is a step-function θ_i^∞, (6) becomes: $m < (u_B - u_A)/(u_B - u_A + 1) < 0.5$. Thus, for θ_i^∞, norm change occurs only if individuals in a population weigh their individual payoff more than whether they conform with others. Figure 1 (middle) shows a heatmap of how condition (6) varies with $u_B - u_A$ and k when $x_B = 0.1$. We see that the bound on m increases as $u_B - u_A$ increases, i.e., a population becomes more likely to switch the norm. On increasing k, we see that the bound on m decreases. This makes intuitive sense, since a higher k makes the difference in conformist fitness between A and B clearer. Thus, in tight cultures, where people tend to agree more on what behaviors are appropriate vs. inappropriate in different situations [13], a higher k would lead to more resistance to norm change.

3.2 Rate of Norm Change in Tight vs. Loose Cultures

We are now interested in studying the speed with which norms change in different populations. Consider two possible values of m, namely m_1 and m_2, with $m_2 > m_1$ (i.e., m_2 is a more conformist culture than m_1). Let the corresponding values of $\dot{x}_B$ be denoted by $\dot{x}_B^1$ and $\dot{x}_B^2$, respectively. Assume further that both m_1 and m_2 satisfy (6), i.e., norm change takes place in both cultures. Analyzing the difference in the rates of change, from (5) we get:

$$\dot{x}_B^2 - \dot{x}_B^1 = x_B(1 - x_B)(m_2 - m_1)\big[(\theta_B - \theta_A) - (u_B - u_A)\big]. \tag{7}$$

Note that when $x_B \leq 0.5$, $\theta_B - \theta_A \leq 0$, which would mean: $\dot{x}_B^2 - \dot{x}_B^1 \leq 0$, i.e., the more conformist culture would be slow to change initially. To analyze the case when $x_B > 0.5$, let's assume $x_B = 0.5 + \epsilon$, for $\epsilon > 0$. Thus, $x_A = 0.5 - \epsilon$. From (7), we see that for $\dot{x}_B^2 - \dot{x}_B^1 > 0$, the following condition needs to hold:

$$\epsilon > \big[\ln(1 + u_B - u_A) - \ln(1 - (u_B - u_A))\big]/k. \tag{8}$$

Figure 1 (right) plots a heatmap of how this bound varies with $u_B - u_A$ and k. We see that as k increases, the point at which the more conformist culture starts changing faster moves closer to the point $x_B = 0.5$. Note that when $k \to \infty$,

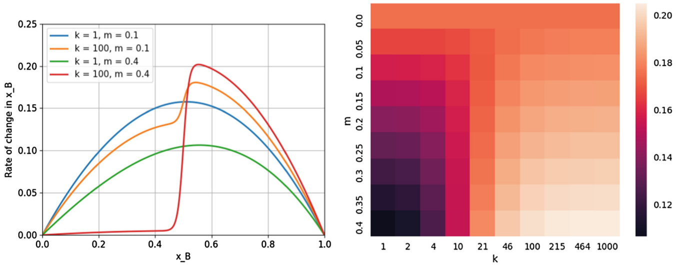


Fig. 2. *Left:* Plot of (5) at $u_B - u_A = 0.7$. *Right:* Heatmap of $\max_{x_B} \dot{x}_B$ for various k and m values, with $u_B - u_A = 0.7$. Best viewed in color. (Color figure online)

(8) reduces to $\epsilon > 0$, i.e., as soon as x_B becomes a majority, greater conformism would produce a larger rate of change.

We are also interested in studying how the rate of change $\dot{x}_B$ varies with x_B as a population switches to norm B, and how this relates to different levels of conformism. We first look at the *maximum* rate of change $\dot{x}_B^{\max} = \max_{x_B} \dot{x}_B$ (we numerically calculate this given values for k, m and $u_B - u_A$). Figure 2 (right) plots $\dot{x}_B^{\max}$ for different m and k values, where we set $u_B - u_A = 0.7$. These values were chosen such that the norm changes from A to B for all the considered combinations (using the bounds from Sect. 3.1). We see that when k is low, lower conformism leads to higher $\dot{x}_B^{\max}$. However, as k increases (i.e., as θ_B approaches θ_B^{∞}), there is a clear transition, where more conformist cultures end up having a higher maximum rate of change.

This effect is clearer in the left plot of Fig. 2, where we show how $\dot{x}_B$ varies with x_B. We see that when k is low, i.e., when there is no clear difference between A and B in its conformist fitness measure θ, $\dot{x}_B$ changes slowly for both tight and loose cultures, with the loose culture having a higher rate of change. With high k, however, we see that the tighter culture faces a *tipping point*, resulting in a sudden increase in $\dot{x}_B$ with the tighter culture adopting a higher rate of change $\dot{x}_B$ than the loose culture. Thus, using a more general model of conformist transmission over [22], we find that the parameter k, which makes the difference in conformist fitness between following and violating the norm clearer, has a big influence on the pressure of conformity, and thus on the rate at which the norm changes in a society. In summary, in a more conformist culture, initially peer pressure impedes the switch to the more beneficial norm B. But once enough of the population has switched, a tipping point is reached where peer pressure causes the rest of them to switch very rapidly.

4 Discussion

This paper presents an EGT model that aims to investigate how the cultural dynamics of norm maintenance and norm-change differ across various cultures. We show that tight cultures sometimes experience a tipping point for norm change while loose cultures typically face a more gradual change. The results presented assume an infinite well-mixed population with two possible actions. We believe it would be relatively straightforward to extend our results for multiple actions. Assuming an infinite well-mixed population made our model mathematically tractable and to provide exact conditions for norm change. As future work, it would be interesting to extend this model to the finite population case, where interactions between individuals are dictated by a social network.

References

1. Bicchieri, C.: The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge University Press, Cambridge (2005)
2. Bowles, S., Gintis, H.: The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor. Popul. Biol. **65**(1), 17–28 (2004)
3. Boyd, R., Gintis, H., Bowles, S., Richerson, P.J.: The evolution of altruistic punishment. Proc. Nat. Acad. Sci. **100**(6), 3531–3535 (2003)
4. Brandt, H., Hauert, C., Sigmund, K.: Punishment and reputation in spatial public goods games. Proc. R. Soc. Lond. B: Biol. Sci. **270**(1519), 1099–1104 (2003)
5. Brandt, H., Hauert, C., Sigmund, K.: Punishing and abstaining for public goods. Proc. Nat. Acad. Sci. **103**(2), 495–497 (2006)
6. Centola, D., Baronchelli, A.: The spontaneous emergence of conventions: an experimental study of cultural evolution. Proc. Nat. Acad. Sci. **112**(7), 1989–1994 (2015)
7. Chen, W., Lakshmanan, L.V., Castillo, C.: Information and influence propagation in social networks. Synth. Lect. Data Manag. **5**(4), 1–177 (2013)
8. De, S., Gelfand, M.J., Nau, D., Roos, P.: The inevitability of ethnocentrism revisited: ethnocentrism diminishes as mobility increases. Sci. Rep. **5**, 17963 (2015)
9. De, S., Nau, D.S., Gelfand, M.J.: Understanding norm change: an evolutionary game-theoretic approach. In: Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp. 1433–1441 (2017)
10. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, Cambridge (2010)
11. Fang, C., Kimbrough, S.O., Pace, S., Valluri, A., Zheng, Z.: On adaptive emergence of trust behavior in the game of stag hunt. Group Decis. Negot. **11**(6), 449–467 (2002)
12. Gelfand, M.J., Harrington, J.R., Jackson, J.C.: The strength of social norms across human groups. Perspect. Psychol. Sci. **12**(5), 800–809 (2017)
13. Gelfand, M.J., Raver, J.L., Nishii, L., Leslie, L.M., Lun, J., Lim, B.C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al.: Differences between tight and loose cultures: a 33-nation study. Science **332**(6033), 1100–1104 (2011)
14. Hamilton, W.D., Axelrod, R.: The evolution of cooperation. Science **211**(27), 1390–1396 (1981)
15. Hammond, R.A., Axelrod, R.: The evolution of ethnocentrism. J. Confl. Resolut. **50**(6), 926–936 (2006)
16. Harrington, J.R., Boski, P., Gelfand, M.J.: Culture and national well-being: should societies emphasize freedom or constraint? PloS One **10**(6), e0127173 (2015)
17. Harrington, J.R., Gelfand, M.J.: Tightness-looseness across the 50 United States. Proc. Nat. Acad. Sci. **111**(22), 7990–7995 (2014)

18. Hartshorn, M., Kaznatcheev, A., Shultz, T.: The evolutionary dominance of ethnocentric cooperation. J. Artif. Soc. Soc. Simul. **16**(3), 7 (2013)
19. Hauert, C.: Replicator dynamics of reward & reputation in public goods games. J. Theor. Biol. **267**(1), 22–28 (2010)
20. Hechter, M., Opp, K.D.: Social Norms. Russell Sage Foundation, New York City (2001)
21. Helbing, D., Yu, W., Opp, K.D., Rauhut, H.: Conditions for the emergence of shared norms in populations with incompatible preferences. PloS One **9**(8), e104207 (2014)
22. Henrich, J.: Cultural transmission and the diffusion of innovations: adoption dynamics indicate that biased cultural transmission is the predominate force in behavioral change. Am. Anthropol. **103**(4), 992–1013 (2001)
23. Hofbauer, J., Sigmund, K.: Evolutionary game dynamics. Bull. Am. Math. Soc. **40**(4), 479–519 (2003)
24. Jackson, M.O.: Social and Economic Networks. Princeton University Press, Princeton (2010)
25. Judd, S., Kearns, M., Vorobeychik, Y.: Behavioral dynamics and influence in networked coloring and consensus. Proc. Nat. Acad. Sci. **107**(34), 14978–14982 (2010)
26. Kearns, M., Judd, S., Tan, J., Wortman, J.: Behavioral experiments on biased voting in networks. Proc. Nat. Acad. Sci. **106**(5), 1347–1352 (2009)
27. Kimbrough, S.O.: Foraging for trust: exploring rationality and the stag hunt game. In: Herrmann, P., Issarny, V., Shiu, S. (eds.) iTrust 2005. LNCS, vol. 3477, pp. 1–16. Springer, Heidelberg (2005). https://doi.org/10.1007/11429760_1
28. Kooti, F., Yang, H., Cha, M., Gummadi, P.K., Mason, W.A.: The emergence of conventions in online social networks. In: ICWSM (2012)
29. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in Twitter. In: Proceedings of the 21st international conference on World Wide Web, pp. 251–260. ACM (2012)
30. Lin, Y.R., Margolin, D., Keegan, B., Baronchelli, A., Lazer, D.: # bigbirds never die: understanding social dynamics of emergent hashtag. arXiv preprint arXiv:1303.7144 (2013)
31. Merton, R.K.: Science and the social order. Philos. Sci. **5**(3), 321–337 (1938)
32. Mu, Y., Kitayama, S., Han, S., Gelfand, M.J.: How culture gets embrained: cultural differences in event-related potentials of social norm violations. Proc. Nat. Acad. Sci. **112**(50), 15348–15353 (2015)
33. Nowak, M.A.: Five rules for the evolution of cooperation. Science **314**(5805), 1560–1563 (2006)
34. Nowak, M.A., Sigmund, K.: Tit for tat in heterogeneous populations. Nature **355**(6357), 250–253 (1992)
35. Rand, D.G., Nowak, M.A.: The evolution of antisocial punishment in optional public goods games. Nature Commun. **2**, 434 (2011)
36. Riolo, R.L., Cohen, M.D., Axelrod, R.: Evolution of cooperation without reciprocity. Nature **414**(6862), 441–443 (2001)
37. Roos, P., Gelfand, M., Nau, D., Carr, R.: High strength-of-ties and low mobility enable the evolution of third-party punishment. Proc. R. Soc. Lond. B: Biol. Sci. **281**(1776), 20132661 (2014)
38. Roos, P., Gelfand, M., Nau, D., Lun, J.: Societal threat and cultural variation in the strength of social norms: an evolutionary basis. Organ. Behav. Hum. Decis. Process. **129**, 14–23 (2015)
39. Smith, J.M.: Evolution and the Theory of Games. Cambridge University Press, Cambridge (1982)

40. Weibull, J.W.: Evolutionary Game Theory. MIT press, Cambridge (1997)
41. Young, H.P.: Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press, Princeton (2001)
42. Zhang, L., Zhao, J., Xu, K.: Who creates trends in online social media: the crowd or opinion leaders? J. Comput.-Mediat. Commun. **21**(1), 1–16 (2015)

Model Co-creation from a Modeler's Perspective: Lessons Learned from the Collaboration Between Ethnographers and Modelers

Jose J. Padilla[1](✉), Erika Frydenlund[1], Hege Wallewik[2], and Hanne Haaland[2]

[1] Old Dominion University, Suffolk, VA 23435, USA
jpadilla@odu.edu
[2] University of Agder, Universitetsveien 25, 4630 Kristiansand S, Norway

Abstract. This paper reports on the authors' ongoing collaboration on model co-creation, a process that involves not only the reconciliation of methodologies (qualitative vs. quantitative), but also of epistemologies (empirical vs empirical/rationalist) and ontologies (observable referent vs. abstracted referent). The co-creation process has taken place over several months, from early 2017, both in person, teleconferencing and via email. The result was an ethnographic model of the refugee situation in Lesbos, Greece. The qualifier "ethnographic" means that the simulation's purpose was to capture the problem situation described by ethnographers in a manner that resembles their observations, not to answer a research question. Ethnographers used the modeling process – mostly elicitation and variable identification - to think about questions they had not considered in the field. Further, the used the prototype model to further narrow their desired modeling scope and ask new questions. Lastly, notes captured by the ethnographers in the field highlight the challenges of the modeling situation.

Keywords: Ethnography · Qualitative data · Simulation

1 Introduction

Starting off our work together the modelers asked: "So what is the question?" and after initial fumbling, the ethnographers replied – "here is the story."[1]

Modelers increasingly encounter work with social scientists as modeling and simulation makes its way beyond the realm of engineering and into questions about social and behavioral phenomena. Quantitative and mixed-methods social science researchers may find the transition from statistical models to simulation-based thinking relatively natural and uncontroversial. The relationship between qualitative social sciences and M&S, however, is not necessarily straightforward. Further, there are few published cases where this combination of approaches have taken place. Some examples include agent-based modeling and grounded theory (Dilaver 2015), system dynamics and case study methods (Rwashana et al. 2009), and agent-based modeling (ABM) and ethnography

[1] Italicized paragraphs at the beginning of each section are from ethnographers' field notes.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 70–75, 2018.
https://doi.org/10.1007/978-3-319-93372-6_8

(Ghorbani et al. 2015). However, little has been explored regarding how researchers across disciplines collaborate to implement difficult-to-reconcile approaches.

At the very outset of a research endeavor, M&S and qualitative research approaches have two fundamental features in common. First, both are arguably a creative process. Neither has an explicit formula for producing the desired conceptualization of the real-world phenomena of interest, yet they rely on procedural frameworks. Second, M&S and qualitative research are both iterative. Where simulation models are refined and iteratively become more specific or complex, qualitative analysis requires deep immersion in the data and iterative processing of theoretical constructs and possible interpretations.

Foundationally, there is a possible point of contention between M&S and qualitative research agendas. For social scientists, "One of the functions of qualitative analysis is to find patterns and produce explanations" (Gibbs 2008). Where social scientists seek explanation, M&S addresses a number of different objectives including prediction. This is not inherently a conflict between M&S and qualitative research. There are many reasons to model other than prediction including focusing research questions, data collection, and dialogue about policy decisions (Epstein 2008). Qualitative researchers who are particularly interested in contextualized, small sample approaches to understanding the lived reality of individuals find the emphasis of prediction in M&S descriptions to be particularly off-putting. This is an important point for reconciling the worldview of modelers and qualitative social scientists. Particularly of ethnographers and those who use participant observation, interviews, focus groups, or even participatory community-based approaches to gather data, their primary objective is generally explanatory. This type of research is grounded in the particular: context matters to the extent that generalizability is often outside the scope of the research agenda. Additionally, qualitative field studies often acknowledge the bias of the researcher, where bias does not necessarily connote a negative impact to the study as it might in traditionally quantitative approaches based on the scientific method.

2 Parallels Between Simulation and Qualitative Field Study Approaches

> *"Something happened yesterday," the seminar leader, and leader of the overall project, started the morning session with these simple words that something happened yesterday. Agreeing to that and explaining to the modelers the experience of a major setback during a plenary the day before, doubts were voiced from the ethnographers' side. Enthusiasm was replaced with skepticism as a discussion unfolded. Through dialoging across, with the modelers acknowledging the skepticism of ethnographers, especially on the epistemological doubts, trust was regained. The link between grounded theory and modeling, and why the two go so well together.*

For simplicity purposes, we compare the research framework of Tolk et al. (2013) and John and Lyn (1995) as they provide step-wise approaches for each group. Simplicity is emphasized as approaches vary due to unique factors brought by each area of study. We will not discuss all the steps for brevity.

Tolk et al. (2013) proposed a framework focused on simulation creation. The main premise of the framework lies on the nature of the problem for which a simulation is being created. A “problem situation” is one that is difficult to formulate due to lack of consensus about the nature and existence of the problem (Vennix 2000). A simulation is then the computer implementation of a potential explanation of the problem situation. The framework attempts to guide the modeler through three main processes: reference modeling (a comprehensive model that captures theories, explanations, models, data, and assumptions), conceptual modeling (a sub-set of the reference model that focuses on answering a modeling question), and simulation modeling (a sub-set of the conceptual model that can be implemented in a computer). Other activities like verification and validation also take place, but they are usually comparisons between models or between the model and phenomenon.

John and Lyn (1995) organize their description of social science research into three phases: gathering data, focusing data, and analyzing data. These three phases lead to an explanation of the phenomenon of interest. “Focusing data” represents a phase that overlaps data collection and data analysis.[2] During the data collection, they propose a “form of consciousness” that begins to organize and make sense of the data and eventually gives way to data analysis. Social scientists often ask eight broad questions about the phenomena of interest. What are the (1) types, (2) frequencies, (3) magnitudes, (4) structures, (5) processes, (6) causes, and (7) consequences observed; and (8) “how [do] people strategize their actions in and toward situations and settings” (i.e. human agency) John and Lyn (1995).

We can argue that problem situations are what ethnographers explore in their research. The challenge becomes structuring ethnographic data to reconcile modelers and social scientists’ perspectives.

If we adopt as a starting point the terminology “Reference Modeling” in M&S and “Focusing Data” in social science, it is relatively straightforward to illustrate that these two processes follow similar paths in both disciplines (Fig. 1). Both approaches attempt to understand a problem situation, collect data collection, and synthesize that data.

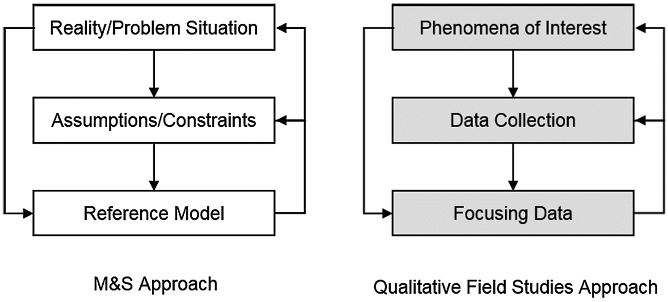


Fig. 1. Parallel between an M&S and qualitative field study approaches (initial stages)

Similarly, both groups of researchers develop conceptual models. As mentioned above, while qualitative researchers do not necessarily use the same terminology, they

[2] We follow this terminology here for simplicity of explanation, acknowledging that social scientists and modelers may have a number of different ways to describe the process.

do in fact develop mental models of how the data is organized and the relational aspects of units of analysis within their study. The difference is how the data is structured and represented.

3 The Story

> *As the story about a crisis unfolded it was carefully noted down and put into boxes on a piece of paper. The modeler had started the work, asking questions along the way that forced the ethnographers to think about the data, answering questions yet not thought of and reorganizing the material – in other words making explicit the implicit model that had already been built through engaging with the data. Along with the narration the dialogue also gave birth to new questions to be explored. And surprisingly also guiding us as ethnographers towards the lack of data and where to go next.*

Ethnographic work, captured and conveyed in narrative form, is not easy to distill into a model. The most obvious question was *what do we focus on*? After a few minutes, the decision was just to listen and document as much as possible of the story, focusing on factors, actors, and relations. In particular, we were discussing the evolution of the refugee situation on the island of Lesbos, Greece from the start until present day, scoping it from 2014 to 2016. The driving interest for the ethnographers was to study the phenomenon of citizen initiatives in Lesbos. However, they needed to understand it in the context of the refugee situation.

The narrative, from the ethnographers' perspectives, would partly read:

> *During three field visits to Lesvos we conducted interviews with people from different CIGS*[3] *working in Lesvos attending to the crisis, during and after the peak in 2015. In these interviews we were focusing on their doings (work knowledge in the institutional ethnography language). Exploring and understanding how they work from within CIGS concerning the role they play, both during a crisis and in the aftermath, necessitates talking to other actors as well. …Moreover, observation was done in two of the refugee camps on the island, and we have also had numerous informal talks with people while staying in different places in Lesvos during fieldwork.*

As modelers, we attempted to capture, among the most relevant items, the chronological evolution of events, stakeholders, potential variables, and the "context" surrounding each year. For instance, the context from 2014 was a climate of poor economic conditions due to austerity measures imposed by the European Union. These austerity measures had a negative impact on the island. Despite these measures, residents were accommodating to the plea of refugees, generating what ethnographers were interested in: citizen initiatives or people self-organizing to provide care for victims of forced migration. Figure 2 shows a snapshot of the notes from one of the discussions.

[3] Citizen Initiatives for Global Solidarity (CIGS) is the term coined to reflect emergent actors in the humanitarian sector, the main focus of the ethnographers' research.

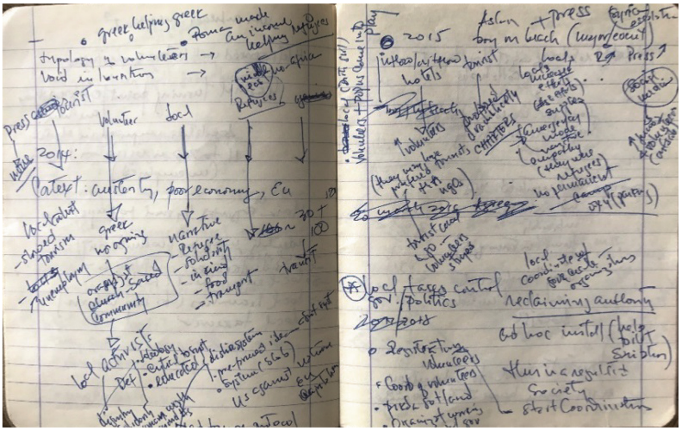


Fig. 2. Snapshot of modeler's notes

4 The Model

As modelers, we don't think in terms of capturing a story. We think in terms of purposeful models that "go beyond" a description of events. This endeavor forced us to think about how a model can tell a story. Not a story reflected in variables and numbers but one that should convey the challenges communities encounter and the concerns of stakeholders. While the world "self-organizing" seemed like a pointer towards an agent-based implementation, the narrative was pointing to a system dynamics (SD) implementation. We followed the narrative and generated an initial SD model.

The initial prototype was not intended to be correct or complete. Its goal was to capture the narrative put forth and provide a platform for discussion. We, as the modelers, wanted to assess if we had captured the story from the ethnographers and identify whether data would be available to explore model.

Briefly, the model attempts to capture the impact refugees have on Lesbos in areas ranging from infrastructure, island appeal, tourism, and quality of life. The role of other actors, like NGOS and citizen initiatives (called AHGO in the model) are also captured in the prototype. The response to the prototype by the ethnographers was:

> *[It] has so many interesting things to tell us…look at the links between solidarity and quality of life. There is something there we need to pay attention to!*

5 Conclusions and Future Work

One of the lessons learned is to establish a division of work and what that entitles. Modelers are constraint by what theories to use and deciding what the relevant data from fieldwork is. This side is supported by ethnographers as domain experts. Social scientists, on the other hand, need to engage in a process of simplification to abide by the constrains of what can be structured and/or programmed in a computer. As such, like in any other collaboration process, there is a negotiation stage where trust must be established. The understanding of limitations and identification of common modeling objectives through extensive dialogue were essential for the collaboration to work. Another lesson was that for the ethnographers, being new to modeling and simulation, knowledge co-creation is appealing as models can be used for interpreting qualitative data, generating new questions and for making implicit models explicit. Similarly, for the modelers, thinking about a meaningful way of capturing these stories in a model so they can provide insight into a phenomenon of interest pushes the role of what models can and should do to support insight generation. In this case, the generated model's role as a description of a phenomenon is less familiar with modelers than a model used for experimentation.

Lastly, since our last meeting, we have identified data sources for the model and data that we need but do not have. In addition, we have established preliminary equations and implemented them in the model. Further, we have isolated the citizen initiatives portion of the model and started implementation both in SD and ABM frameworks.

References

Rwashana, A.S., Williams, D.W., Neema, S.: System dynamics approach to immunization healthcare issues in developing countries: a case study of Uganda. Health Inform. J. **15**, 95–107 (2009)

Dilaver, O.: From participants to agents: grounded simulation as a mixed-method research design. J. Artif. Soc. Soc. Simul. **18**, 15 (2015)

Epstein, J.M.: Why model? J. Artif. Soc. Soc. Simul. **11**, 12 (2008)

Ghorbani, A., Dijkema, G., Schrauwen, N.: Structuring qualitative data for agent-based modelling. J. Artif. Soc. Soc. Simul. **18**, 2 (2015)

Gibbs, G.R.: Analysing Qualitative Data. Thousand Oaks, SAGE (2008)

John, L., Lyn, L.H.: Analyzing Social Settings: A Guide to Qualitative Observation and Analysis. Wadsworth Publishing Company, Belmont (1995)

Tolk, A., Diallo, S.Y., Padilla, J.J., Herencia-Zapana, H.: Reference modelling in support of M&S: foundations and application. J. Simul. **7**, 69–82 (2013)

Vennix, J.A.M.: Group model-building: tackling messy problems. Syst. Dyn. Rev. **15**, 379–401 (2000)

Multi-Agent Accumulator-Based Decision-Making Model of Incivility (MADI)

Jordan Richard Schoenherr(✉) and Kim Nguyen

Department of Psychology, Carleton University, Ottawa, Canada
Jordan.Schoenherr@Carleton.ca,
KimNguyen3@cmail.carleton.ca

Abstract. While behaviour can either be perceived as respectful or disrespectful, incivility reflects relatively minor violations of social norms within a group. In the present study, we used an accumulator-based model of decision-making, assuming that social agents attempt to classify behaviour as respectful or disrespectful based on available social cues and reciprocate toward other group members once a criterion amount of evidence is accumulated. Perceived incivility is derived from the model by taking the balance of evidence of the respectful and disrespectful social cues, reflecting uncertainty in decision-making. In multi-agent interactions, the model averages perceived incivility (i.e., uncertainty) over multiple trials. We demonstrate that this model can differentiate between attitudes and behavior in a single social agent as well as how incivility can arise within a group as a result of small differences in response threshold to disrespectful behaviour and biases in social cue identification accuracy.

Keywords: Incivility · Certainty · Disrespectful behaviour
Accumulator model

1 Introduction

1.1 Incivility: Uncertainty and Reciprocity Within a Group

Workplace deviance is a growing concern. While considerable research has focused on bullying, harassment, and discrimination, comparatively less work has considered *incivility*: workplace behaviour that reflects minor violations of expected group norms (e.g., Cortina et al. 2001). Incivility can include making demeaning remarks, withholding information, undermining the credibility of an employee, making unwarranted accusations, and giving someone the "silent" treatment. A critical feature of incivility is uncertainty (Andersson and Pearson 1999): while potentially disrespectful actions might have occurred, a social agent's intentions might not be clear. Moreover, Andersson and Pearson (1999); Pearson et al. (2000) have suggested that group members use a tit-for-tat response strategy (i.e., a reciprocal norm) such that group members respond with incivility when they observe incivility. This in turn results in a "downward spiral" over time. We consider a multi-agent accumulator-based decision-making model of incivility (MADI) that uses trial-to-trial uncertainty as a means to differentiate respectful behaviour from perceived incivility.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 76–81, 2018.
https://doi.org/10.1007/978-3-319-93372-6_9

2 Decision-Making and Uncertainty in Classification

While incivility has yet to be modelled, conceptual (Brown and Levinson 1987) and computational models (e.g., Miller et al. 2006) have considered politeness. In this case, a social agent weighs a number of factors including the severity of the transgression and potential outcomes of redressing this behaviour. Rather than considering incivility as a distinct set of behaviours and beliefs, incivility might be understood in terms of general decision-making processes wherein a social agent attempts to interpret social cues (e.g., Brunswik 1952). While a number of decision-making models have been proposed (e.g., dynamical systems, random-walk diffusion, and signal detection theory), we consider an accumulator-based model of decision-making as these models have been used to examine decision-making and response certainty (for a review, see Baranski and Petrusic 1998). These models are defined by accumulators that retain information for the response alternatives and produce a response when a criterion threshold of evidence has been accumulated. The proportion of evidence favouring the dominant response (e.g., respectful social cues) relative to all accumulated evidence (i.e., respectful and disrespectful social cues) can be used to scale the uncertainty of the social agents decision-making process (i.e., the ambiguous of the intentions of another social agent). We first consider a possible accumulator-based model that uses response uncertainty as a measure of incivility and then simulate multiple agents interacting in dyads over a short time interval.

2.1 Single-Agent Accumulator-Based Model

In order to simulate social cues, MADI is presented with a randomly selecting values from a normal distribution with range 0 to 1. The accumulation of evidence is modelled by means of a fixed criterion, such that if a randomly generated value is below criterion (e.g., $c = 0.5$) the model will perceive a respectful social cue, otherwise it will perceive a disrespectful social cue. Social cue accuracy is modelled by multiplying this randomly value by a fixed criterion, c. For instance, difficult decisions can be modelled by setting a low criterion (e.g., $c = 0.5$) such that social cues for either respectful or disrespectful behaviour are equally likely whereas easy decisions can be modeled by setting a high criterion (e.g., $c = 0.8$) such that a greater proportion of the distribution is associated with respectful behaviour. Another basic assumption of MADI is that a response is produced when one of the two accumulators reaches a threshold (e.g., 5 social cues). The model also uses separate response threshold for the respectful cues accumulator and the disrespectful cues accumulator, allowing MADI to be more or less sensitive to certain behaviours. Consequently, a social agent might require very few social cues to perceive respectful or disrespectful behaviour.

A defining feature of incivility is that it reflects behaviour that is assigned ambiguous intentions (Andersson and Pearson 1999). Thus, a social agent experiences uncertainty as to whether the social cues that were presented suggest respectful or disrespectful behaviour. The products of accumulator-based decision models can also be used to scale response confidence (Baranski and Petrusic 1998). For instance, Vickers (1979) suggested confidence could be understood in terms of the proportion of evidence accumulated for the dominant response relative to the total accumulated evidence for dominant and nondominant responses, i.e., Certainty = $A_D/(A_D + A_N)$. We additionally assume Incivility reflects *either* respectful (R) or disrespectful (D) decisions associated with *low response certainty*. Thus, we additional assume that:

$$\text{Incivility} = (A_R \text{ and } A_D \mid \text{Certainty} \leq .66).$$

Thus, our model suggests that either respectful or disrespectful responses associated with low certainty reflect perceived uncivil behaviour. We used .66 as the criterion for incivility given that it would create an equal frequency of responses, e.g., high-certainty in respectful responses, high-certainty in disrespectful responses, and low-certainty in either respectful or disrespectful responses.

2.2 Multi-agent Accumulator-Based Model

In our multi-agent accumulator model, we assume that social agents interact in dyads and adhere to a reciprocity norm (Andersson and Pearson 1999). When a social agent obtains a criterion amount of evidence (i.e., social cues) that they are being treated in a respectful manner, they will act in a respectful manner to another group member: the output of one model is used as the input of the next model. We additionally assume that each model retains a memory from previous interactions, thus incivility on one trial is modelled as the average of the uncertainty on the current trial and the certainty from the previous trial. At the end of a multi-agent simulation, we can examine the average respectful behaviour produced within a group as well as the average perceived incivility.

3 Results and Discussion

3.1 Simulation 1: Single-Agent Accumulator-Based Model

In our first simulation, we examined the effects of varying response threshold and accuracy in social cue identification for a single model. Each model was run 10 times to reduce variability, approximating 60 different interactions. Social competency was modelled by using three levels: no competence (0.50), intermediate (0.65) and high competence (0.70). Three response thresholds were also examined: the same threshold for respectful and disrespectful behaviour (No Bias), a bias toward responding with respect (Respect Bias), or a bias toward responding with disrespect (Disrespect Bias). In the No Bias condition, the threshold for both accumulators was set to 5 social cues. In the bias conditions, one accumulator threshold was set to 4 social cues whereas the other accumulator threshold was set to 6 social cues.

Incivility was scaled by dividing certainty into three categories: respect ($A_R \mid \text{Certainty}_R > .66$), incivility ($A_R$ and $A_D \mid \text{Certainty}_R \leq .66$), and disrespect ($A_D \mid \text{Certainty}_R > .66$) and were then recoded as 1, 2, and 3, respectively. Table 1 contains the mean output for behaviour and perception of respect.

Table 1. Effects of response threshold bias and social cue competence of respectful behaviour and perceived incivility

	Respect bias			No bias			Disrespect bias		
	.50	.65	.70	.50	.65	.70	.50	.65	.70
Behavioural response	.78	.93	.975	.88	.975	.917	.58	.80	.85
Perceived respect	2.08	2.37	2.47	2.1	2.32	2.5	1.88	2.12	2.17

As Table 1 indicates, both respectful bias and social cue competency affected model behaviour and perception. A mixed repeated-measures ANOVA included Respectful Bias (No Bias, Respect Bias, and Disrespect Bias) as a between-subjects variable and Social Competency (low, intermediate, and high) as a within-subject variable. An ANOVA revealed that both Social Competency ($F(2, 54) = 47.8$, $MSE = .006$, $p < .001$), Respectful Bias ($F(2, 27) = 92.01$, $MSE = .004$, $p < .001$), and their interaction ($F(4, 54) = 2.77$, $MSE = .006$, $p = .036$) significantly affected the model's respectful responses.

Importantly, our ANOVA of perceived incivility ratings produced results with an important difference. Main effects were obtained for both Social Competency ($F(2, 54) = 26.24$, $MSE = .038$, $p < .001$), Respectful Bias ($F(2, 27) = 17.03$, $MSE = .037$, $p < .001$), while their interaction was not significant ($p = .78$). The discrepancy between behavioural response measures and perceived incivility can interpreted as comparable to psychological exit: while a social agent might *act* in a respectful manner, they might in fact *perceive* considerable disrespect in their social environment. Moreover, while respectful responses and perceived incivility were affected by different factors, we did find a positive correlation, $r(90) = .729$, $p < .001$. Following from the finding that attitudes and behaviour are only somewhat related, our measure of incivility and respectful behaviour appear to suggest that a social agent might not always act in a manner that reflects the perceived incivility in their social environment.

3.2 Simulation 2: Multi-agent Accumulator-Based Model

Simulation 2 examined how respectful behaviour and perceptions of incivility change over time within a closed group. In this case, we assume that 12 interactions occur in dyads (i.e., one-to-one). In each iteration, a decision that is made by one social agent is used as input by the next social agent. Each social agent also retains an average of the perceived incivility from the current trial and the previous trial reflect one's attitudes toward the group as a whole.

To examine the effect of differences in group composition, we varied the number of individuals who had sensitivities in identifying disrespectful behaviour. We selected four situations where no individuals had a disrespect bias (0%), or where 3 (25%), 6 (50%), or 12 (100%) individuals had a disrespect bias. We considered 12 sets of random interactions of the respectful and disrespectful models. Figure 1 presented the results of the simulations, averaged over 3 simulation periods producing 4 aggregated periods (e.g., simulations periods 1 through 3 were averaged together to reduce trial-to-trial variability).

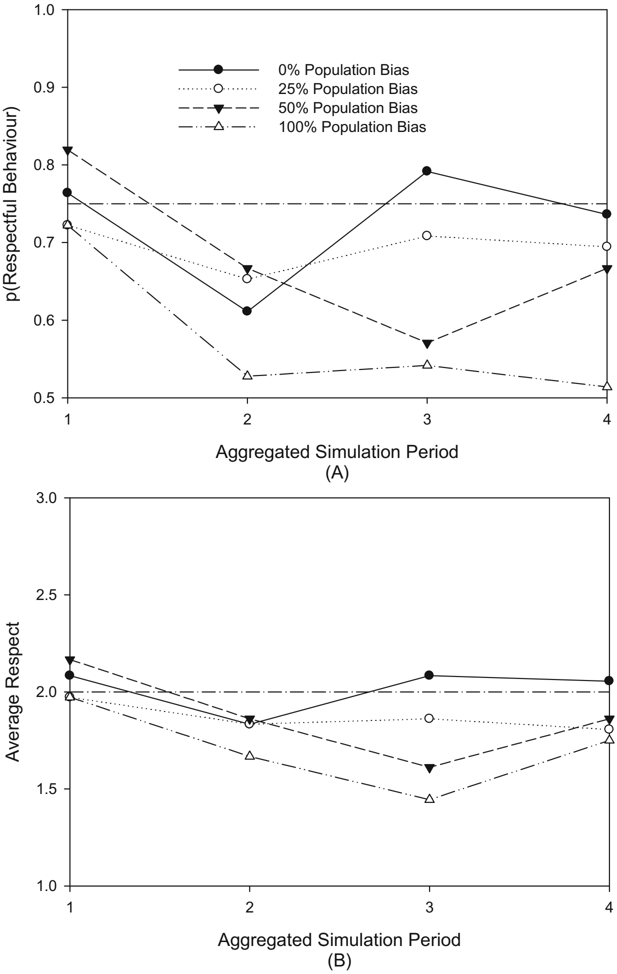


Fig. 1. Effects of group composition on respectful behaviour (A) and civility (B). Groups varied in terms of whether 0%, 25%, 50%, or 100% of members had a bias toward perceiving disrespectful behaviour.

As Fig. 1 indicates, respectful behavior generally decreased over the course of the simulated interaction in all groups (Fig. 1A) as evidenced by the fact that the majority of data points for all functions fell near or below the threshold for disrespectful behavior (i.e., p(respect) = 0.75). Notably, populations that had fewer members with disrespectful bias were displayed more respectful behavior. An ANOVA including trial as a within-subjects variable and disrespectful bias as a between-subject variable provided support for finding: both population composition ($F(3, 44) = 3.63$, $MSE = .055$, $p = .020$), aggregated simulation period ($F(3, 132) = 7.63$, $MSE = .024$, $p < .001$), and their interaction ($F(9, 132) = 2.12$, $MSE = .024$, $p < .001$) significantly affected the models respectful responses.

Similarly, even in groups with a smaller proportion of members that were sensitive to disrespectful behavior, many behaviours were classified as incivility (Fig. 1B) as evidence by the location of functions around the mid-point representing perceived incivility (i.e., Average Respect = 2). However, an ANOVA revealed that the only significant difference was the simulation period ($F(3,\ 132) = 2.12$, $MSE = .024$, $p < .001$). Whereas the interaction of simulation period and population composition was not significant ($p = .335$), population composition was marginally significant ($F(1, 44) = 2.75$, $MSE = .227$, $p = .057$). While the lack of a significant effect of population composition is likely a result of the small number of group-level simulations, it is notable that by the end of the simulation, most of the social agents perceived incivility rather than respectful or disrespectful behaviour.

4 Conclusions

The results of the current study suggest that MADI can be used to model differences in respectful behavior and perceived incivility that emerge within a group over time. By varying the sensitivity of a social agent to disrespectful social cues as well as the accuracy of social cue detection, we found that individual differences in identification and production of disrespectful behavior can give rise to group-level behaviour. Overall, small differences in the ability to correctly identify social cues introduce more disrespectful behaviour as the result of the reciprocity norms. Moreover, we found that respectful behaviour and incivility were not perfectly correlated, replicating the weak relationship between attitudes and behaviour more generally.

References

Andersson, L., Pearson, C.: Tit for tat? The spiraling effect of incivility in the workplace. Acad. Manag. Rev. **24**, 452–471 (1999)

Baranski, J.V., Petrusic, W.M.: Probing the locus of confidence judgments: experiments on the time to determine confidence. J. Exp. Psychol. Hum. Percept. Perform. **24**, 929–945 (1998)

Brown, P., Levinson, S.: Politeness: Some Universals in Language Usage. Cambridge University Press, Cambridge (1987)

Brunswik, E.: The Conceptual Framework of Psychology. University of Chicago Press, Chicago (1952)

Cortina, L.M., Magley, V.J., Williams, J.H., Langhout, R.D.: Incivility in the workplace: incidence and impact. J. Occup. Health Psychol. **6**, 64–80 (2001)

Miller, C., Wu, P., Funk, H., Wilson, P., Johnson, L.: A computational approach to etiquette and politeness: initial test cases. In: Proceedings of the 15th Conference on Behavior Representation in Modeling and Simulation (BRIMS), 15–18 May 2006, Baltimore, MD (2006)

Pearson, C.M., Andersson, L.M., Porath, C.L.: Assessing and attacking workplace incivility. Org. Dyn. **29**(2), 123–137 (2000)

Vickers, D.: Decision Processes in Visual Perception. Academic Press, New York (1979)

Legislative Voting Dynamics in Ukraine

Thomas Magelinski(✉) and Kathleen M. Carley

Center for Computational Analysis of Complex and Organized Systems,
Institute for Software Research, Carnegie Mellon University, Pittsburgh, USA
tmagelin@andrew.cmu.edu
http://www.casos.cs.cmu.edu/

Abstract. Current work in roll call modeling focuses on the legislative decision process and does not take advantage of the dynamic nature of legislation. Some political systems, such as Ukraine's Verkovna Rada, are inherently dynamic, and should be modeled as such. In the model proposed, the entire legislative body is modeled together and bills are viewed as a dynamic process. This model requires no contextual information about individual legislators and predicts the amount of favorable votes a bill will receive within 6.2%, on average. Additionally, we find differences in behavior of bills proposed by the President and those proposed by parliament members or the Cabinet. This work only uses a simple differential model, opening the door to the use of more complex models capable of leveraging contextual information in the future.

Keywords: Differential modeling · Roll call prediction

1 Introduction

The legislative process in the Ukrainian Parliament, the Verkovna Rada, is inherently dynamic. Each bill is voted on several times before being passed or rejected. Current work in roll call modeling focuses on the legislative decision process and does not take advantage of the dynamic nature of legislation. This work seeks to answer the question: can a simple model that takes advantage of bill dynamics be used to predict future legislative outcomes? To answer this question, we propose a new model for dynamic legislation, in which the entire legislative body is modeled together and bills are viewed as a dynamic process. First, background information on the Verkovna Rada is provided. Then, a review of previous voting models is discussed. Following, the model, the procedure for its use, and the results are shown. Finally we discuss the implications of the models vis-a-vis avenues for future work.

This work was supported in part by the Office of Naval Research (ONR) through a MURI Grant N00014-17-1-2675, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing official policies, either expressed or implied, of the Office of Naval Research or the U.S. government. Thomas Magelinski was also supported by an ARCS Foundation scholarship.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 82–88, 2018.
https://doi.org/10.1007/978-3-319-93372-6_10

2 Voting in the Verkovna Rada

The Verkovna Rada, Ukraine's parliament, consists of 440 members. Bills may be sponsored by one of three subjects: a parliament member, the Government of Ukraine (the Cabinet), or the President. Additionally, bills from any subject may be tied to a committee. It is believed that bills sponsored by the President are passed quicker than those sponsored by others. The legislative procedure in the Verkovna Rada requires multiple votes on a bill before it is put into law. Technically a bill passes after 3 separate votes with at least 225 legislators voting "for" it. This leads to bills going through many iterations, sometimes as many as 9, before the bill is passed or given up on. This work uses bills from the most recent convocation, which began on November 27, 2014 and will continue until November 27, 2019.

3 Voting Models

Many strategies for modeling roll call voting have been employed in the past. Early work focused on accurately representing individual legislator's decision process using as much contextual information as possible [4,6]. Part of this contextual analysis relies on political dimensions of the legislation. Early work by MacRae showed the presence of political dimensions in legislation from the United States Congress [5]. Clausen, among others, used these dimensions as the basis for legislative decision making [1]. Most recently, the dynamics of political systems have been leveraged through game theory [2,3]. In contrast to previous work which focused on modeling legislative decision making, we focus on the group dynamic process behind roll call voting.

4 The Data

In total 78 bills were analyzed. Note that the model needs 2 vote iterations before prediction can begin, so the "predictable points" are the following vote iterations. As such, bills with less than 3 iterations would have no predictable points, so they were not used. Each bill has a sponsoring subject, and committee. The votes on each bill iteration are collapsed into a single number, percentage of votes in favor, in order to model the legislative body as a whole, rather than predicting member's individual votes. The summary statistics for the data are shown in Table 1.

Since the percentage of votes for a bill is changing dynamically in time, it can be viewed in phase-space. The phase space diagram shown in Fig. 1. This diagram shows votes spiraling in towards an equilibrium point on the x-axis, indicating that it can be modeled using a differential equation.

Table 1. Bill summary statistics by subject

Subject	Bills	Max iterations	Predictable points	Committees	Mean	Std. Dev.	Min	Max	Median
Parliament	71	9	359	21	45.454	10.681	10.090	71.760	48.813
Government	4	8	21	4	44.640	10.476	22.421	56.502	48.380
President	3	5	12	1	47.160	12.897	26.233	74.664	48.318

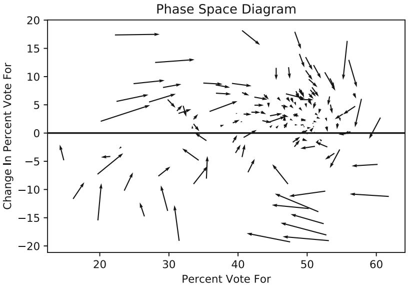


Fig. 1. Percentage of votes for, viewed in phase space.

5 Methodology

5.1 The Model

The goal of the work is to find a model equation, $v(t)$ that approximates the percentage of favorable votes that it receives at time t. Since bills are voted on at discrete intervals, i, we say that votes occur during regular intervals proportionate to the bill iteration:

$$t_i = \alpha * i\ . \tag{1}$$

Based on the structure seen in the phase diagram, Fig. 1, it seems appropriate to use a differential model. This work's model equation follows the form of a second order linear homogeneous differential equation:

$$A * \frac{d^2v}{dt^2} + B * \frac{dv}{dt} + C * v = 0, \tag{2}$$

where A, B, C, and D are constants that can be fit to the data. This model was selected for its simplicity; higher order and/or non-homogeneous models are left for future work. The discrete nature of the bill data also makes estimating derivatives challenging. At least 3 data points are needed to calculate the second derivative. The best estimate of the first derivative at the center point is the average of the change before and the change after. Thus, only bills with 3 or more votes could be used to fit the model. The advantage of having the simple model in Eq. 2 is its analytic solutions. With an analytic solution, the first two iterations of a bill can be fit exactly, instead of relying on the poor estimate of the derivative at point 2, as would be needed for numerical integration.

Following this, $\frac{d^2v}{dt^2}$ and $\frac{dv}{dt}$ will be referred to as v'' and v', respectively. When calculating derivatives from the raw data, the unknown factor α must also be accounted for, thus, the model equation becomes:

$$\frac{A}{\alpha^2}v'' + \frac{B}{\alpha}v' + C * v = 0. \tag{3}$$

Finally, the model will be solved analytically, and will be written in terms of bill iteration (instead of t). Based on the phase portrait, real eigenvalues are expected. In this case, the solution will be of the form:

$$v(i) = C_1 e^{\lambda_1 i} + C_2 e^{\lambda_2 i}. \tag{4}$$

5.2 Fitting Parameters

For each possible vote iteration, the derivatives were calculated, resulting in N data points, each having a value for v, v', and v''. Note that consecutive data points may or may not belong to the same bill. For example, a bill with 4 iterations will become two data points: $[v_2,v_2',v_2'']$ and $[v_3,v_3',v_3'']$. The data does not follow the model exactly, so an objective function was defined:

$$F = \sum_{b=1}^{N} (v_b'' + C_{v'}\alpha v_b' + C_v \alpha^2 v_b)^2. \tag{5}$$

This function is simply Eq. 3 set equal to zero with the v'' coefficient normalized to one, squared, and summed over the whole dataset. Squaring the values is used in place of an absolute value sign, making differentiation smooth. The objective function was minimized using SciPy's implementation of Nelder and Mead's simplex method for minimization [7].

Once the parameters for Eq. 4 are found, bills can be predicted. First, the coefficients C_1 and C_2 are calculated:

$$\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ e^{\lambda_1} & e^{\lambda_2} \end{bmatrix}^{-1} \begin{bmatrix} V_0 \\ V_1 \end{bmatrix}. \tag{6}$$

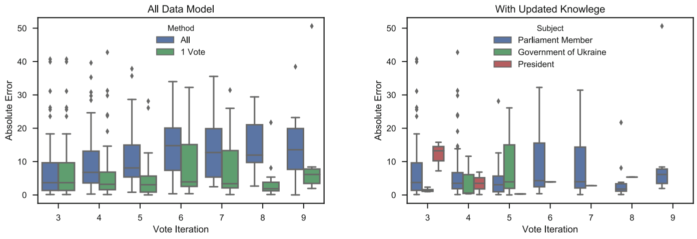
From here, the model was used in two ways. First, Eq. 4 was evaluated at all desired times. Second, the model equation was updated after every iteration by refitting Eq. 6 to the incoming data. The updated knowledge method is more accurate but less powerful as only one iteration is predicted at a time.

6 Results

6.1 All Data

First, the bills sponsoring subject and committee were ignored, and a model was fit to the entire dataset. The resulting model from Eq. 4 has the values

$\lambda_1 = -0.424$ and $\lambda_2 = -0.008$. The negative exponents indicate that the model is indeed following the sink behavior observed in the phase diagram. The average absolute difference between prediction and actual percentage vote for the entire data set is 6.2% using this model. Figure 2 shows the models absolute error by vote iteration. The non-updated prediction performs worse on later iterations. While average error is low, the upper bound on error is very high. Figure 2 also shows that the overall model performs differently based on the initial subject.



(a) (b)

Fig. 2. Predictions using updated knowledge are in a. Updated knowledge model error by subject are in b.

6.2 Bill Subject

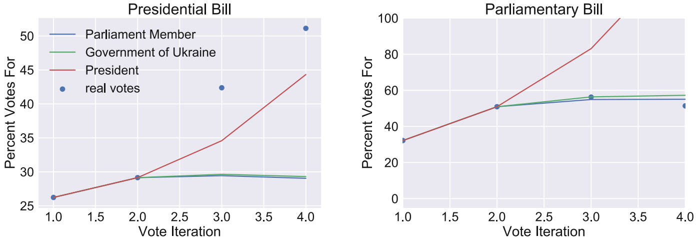
After seeing the difference in model performance by subject, each subject was modeled separately. The parameters λ_1 and λ_2 in Eq. 4 are $(-0.284, -0.005)$, $(-0.494, -0.012)$, and $(0.478, -0.026)$ for bills from parliament, government, and the President, respectively. It is noteworthy that while the parliamentary and government models have two negative coefficients, the presidential model has one positive and one negative. The first two models are sinks, while the last is a saddle. Thus, presidential bill behavior is quantitatively different. See Sect. 7.

On average, the absolute difference was 6.25%, 3.78%, and 2.97% for Parliament, Government, and President respectfully. The subject modeling improved the average accuracy for every subject. The presidential and government models had no instances of greater than 12%. About 25% of parliamentary votes are poorly predicted, >10% absolute error.

Since all the bills with greater than 10% error are from Parliamentary Members, the Bill Committee analysis was performed on only the parliamentary bills. Fitting models based on committee does improve overall error, but there is still many instances of >20% error.

7 Discussion

While fitting a single model for all bills achieves an average error of nearly 6%, modeling initial bill subjects separately increases accuracy and shows that presidential bill have different behavior than those of other subjects. This gives evidence to the hypothesis that presidential bills tend to have different trajectories than bills from other sources. Parliamentary and government bills are modeled to stabilize quickly, while presidential bills continue to increase or decrease based on what happened in the initial two iterations.



(a) (b)

Fig. 3. Predictions and votes for a presidential bill (a) and a parliamentary bill (b).

The difference in model behavior is visualized in Fig. 3. There are two reasons that the presidential model increases rapidly for the parliamentary bill. First, the presidential data has less bills with many iterations, and only a max iteration of 5, so the model is fit mostly to predict iteration 3. Second, the presidential bills all changed slowly first, with a maximum increase of only 5% from the first iteration to the second.

While modeling parliamentary bills separately based on their committee increased the overall accuracy, there were still several instances of >20%, and these errors were spread across many committees. Thus, fitting for committee and fitting for subject cannot completely explain why some iterations of bills are poorly predicted. It is also noteworthy that the parliamentary model has a mean error of 3.5% for the "Committee on Legal Policy and Justice," and a maximum error of 12%. This is the only committee sponsoring presidential bills, indicating that the difference in models cannot be just explained by the committee.

8 Conclusion

This work takes advantage of the dynamic nature of legislative voting in the Verkhovna Rada to build predictive models for future legislative votes. Without considering the initial subject of a bill, the model predicted the percentage of votes in favor of bills within 6.2% on average. Validation on 600 unlabeled bills from convocation 8 resulted in similar accuracy. Modeling by initial subject improved prediction for bills initiated by the President and by the Cabinet to 3.0% and 3.8% respectively. The models imply that presidential bills change significantly after the first two votes, while other bills stabilize quickly.

About 20–25% of parliamentary votes are predicted with over 10% error. Accounting for committee did not resolve this error, so it seems that some votes simply do not follow the model equation. Still, given how little information the model uses, its success shows the potential of dynamic modeling. Future work may use more complex models to better predict the remaining bills.

References

1. Clausen, A.R.: How Congressmen Decide: Policy Focus. St. Martin S Press, New York City (1973)
2. Duggan, J., Kalandrakis, T.: Dynamic legislative policy making. J. Econ. Theory **147**(5), 1653–1688 (2012)
3. Kalandrakis, A.: A three-player dynamic majoritarian bargaining game. J. Econ. Theory **116**(2), 294–322 (2004)
4. Kingdon, J.W.: Models of legislative voting. J. Polit. **39**(3), 563–595 (1977)
5. MacRae, D., Goldner, F.H.: Dimensions of Congressional Voting. University of California Press, Berkeley (1958)
6. Matthews, D.R., Stimson, J.A.: Decision-making by US representatives: a preliminary model. Polit. Decis.-Mak. 14–43 (1970)
7. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)

Stop Words Are Not "Nothing": German Modal Particles and Public Engagement in Social Media

Fabian Rüsenberg[1](✉), Andrew J. Hampton[2], Valerie L. Shalin[3], and Markus A. Feufel[1]

[1] Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany
fabian.ruesenberg@campus.tu-berlin.de

[2] Institute for Intelligent Systems, University of Memphis, Memphis, TN 38152, USA

[3] Department of Psychology and Kno.e.sis, Wright State University, Dayton, OH 45435, USA
valerie@knoesis.org

Abstract. Social media research often exploits metrics based on frequency counts, e.g., to determine corpus sentiment. Hampton and Shalin [1] introduced an alternative metric examining the style and structure of social media relative to an Internet language baseline. They demonstrated statistically significant differences in lexical choice from tweets collected in a disaster setting relative to the standard. One explanation of this finding is that the Twitter platform, irrespective of disaster setting, and/or specifics of the English language, is responsible for the observed differences . In this paper, we apply the same metric to German corpora, to compare an event-based (the recent election) with a "nothing" crawl, with respect to the use of German modal particles. German modal particles are often used in spoken language and typically regarded as stop words in text mining. This word class is likely to reflect public engagement because of its properties, such as indicating common ground, or reference to previous utterances (i.e. anaphora) [2, 3]. We demonstrate a positive deviation of most modal particles for all corpora relative to general Internet language, consistent with the view that Twitter constitutes a form of conversation. However, the use of modal particles also generally increased in the three corpora related to the 2017 German election relative to the "nothing" corpus. This indicates topic influence beyond platform affordances and supports an interpretation of the German election data as an engaged, collective narrative response to events. Using commonly eliminated features, our finding supports and extends Hampton and Shalin's analysis that relied on pre-selected antonyms and suggests an alternative method to frequency counts to identify corpora that differ in public engagement.

Keywords: Big data · Text mining · Common ground · Collective narrative

1 Introduction

Social media research has demonstrated the capacity to assist in the identification of public response to products and entertainment [4], disaster [5, 6], social phenomena such as gender-based-violence [7] and political events such as elections [8]. Most of these efforts exploit metrics based on frequency counts, sometimes scaled with respect to the

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 89–96, 2018.
https://doi.org/10.1007/978-3-319-93372-6_11

corpus in question. For example, a sentiment analysis might compare the number of positive sentiment words to the number of negative sentiment words in a corpus, to determine net sentiment [9, 10]. Machine learning techniques assist in the identification of diagnostic items [11]. However, a non-existing reference distribution limits interpretation.

1.1 Introducing Relative Metrics

Hampton and Shalin [1] introduced an alternative metric, where observed frequency counts are scaled according to a population-based standard. Exploiting pre-selected lexical items corresponding to physical properties such as size and numerosity, and social properties such as cooperation, Hampton and Shalin demonstrated changes in lexical choice in social media collected during a disaster setting, relative to an Internet language baseline. Such a metric contributes to the development of social alarms, potentially assisting in the management of finite disaster response resources. Similar to sentiment analysis, their metric aims to be domain independent, but focuses more on style than content. Unlike sentiment analysis, and of particular relevance to the present paper, Hampton and Shalin included so-called stop words in their analysis, that is, words that are typically excluded in text mining due to their high prevalence. Consistent with Purohit et al.'s [5] claim that social media behave like a group conversation, Hampton and Shalin interpreted the departures from baseline word patterns as sentinels of breach, emergent from a corpus of tweets and constituting a collective narrative.

Hampton and Shalin's interpretation is not without criticism. In particular, it may be that social media messages are fundamentally different from other forms of Internet material, as Purohit et al.'s conversation analysis indicates. If so, all social media corpora would differ from a broadly composed standard. Moreover, the Hampton and Shalin study was confined to an analysis of English corpora, with its specific linguistic and cultural properties. This paper addresses some of these criticisms. It uses metrics inspired by Hampton and Shalin, applied to German social media corpora drawn from events related to the recent election in Germany, compared to a *Seinfeld*-ian "nothing" corpora concerning Maslow type needs about water, food, shelter, and sleep that were unassociated with any particular event. Continuing in the effort to develop domain-independent metrics, most of the words we examine are common stop words. Continuing in the effort to assess public response to an event, we examine engagement in the recent German elections relative to a "nothing" crawl without an event indicator, by analyzing tweet style and structure rather than its content.

1.2 Modal Particles

The German language contains a unique word class, known as modal particles. These are uninflected words characteristic of spoken language. We choose this particular word class because of its potential to reflect public engagement, indicating the speaker's attitude, referring to common ground, assumptions and expectations of the speaker or receiver, or referencing previous utterances (anaphora) [2, 3]. Linguists identify a core class of 15 modal particles [3, 12–15]: *aber, auch, bloß, denn, doch, eben, eigentlich,* *etwa, halt, ja, mal, nur, schon, vielleicht, wohl.* According to [16], all of these words except for *halt* and *mal* are unanalyzed stop words in conventional social media analysis. Examples for the use of modal particles appear in Table 1.

Table 1. Examples of the potential occurrence of a modal particle in a tweet.

Modal particle	Examples
denn/doch	Aber wer soll **denn** in den BT [Bundestag] einziehen? Die FDP besteht **doch** nur aus Lindner, oder? *But who should **"then"** move into the BT (Bundestag)? The FDP consists **"just"** only of Lindner, no?*
nur	Wie kann man **nur** AfD wählen? *How can one **"just"** vote for the AfD?*
vielleicht	Klar, aber wir sollten uns **vielleicht** mit Dingen beschäftigen, die bei uns passieren *Sure, but we should **"maybe"** care about things that happen here.*

German has another welcome property. Relative to English, the use of German is largely confined to Germany and nearby Austria and Switzerland whose citizens are presumably less engaged in the German election. This suspends the need to rely on location metadata for message source and maximizes location-specific data collection.

We demonstrate that the prevalence of modal particle increases for an event of national significance relative to "nothing", suggesting public engagement. We use the method of Hampton and Shalin as described in Sect. 2.1.

2 Methods

2.1 Dataset

Several thousand unique tweets in the German language were collected for three events related to the German Elections 2017 relative to a "nothing" crawl referring to basic human needs (see Sect. 1.1). Table 2 gives an overview of the events, the time frame in which data were collected and the keywords used. Relative to the event corpora, which by definition include election keywords, tweets in the "nothing" corpus do not contain election keywords. Tweets were obtained either through an online semantic web application, Twitris [17], or via the Twitter Search API in R using RStudio (Version 1.0.136). From the original data set, unique tweets were obtained by eliminating retweets and using the `unique()` command in R.

Table 2. Overview of events including start and end date and crawling word set.

Event	Start	End	Crawling word set
German Elections 1 $N = 601{,}498$	2017 Sep 20	2017 Sep 26	spd, cdu, fdp, afd, npd, grüne, gruene, linke, union, #btw17, #btw2017, bundestagswahl, bundestag, wahlen, deutschland, land, partei, merkel, schulz, stimme, demokratie, wahlkampf, hochrechnung
German Elections 2 $N = 183{,}861$	2017 Oct 11	2017 Oct 23	spd, cdu, fdp, afd, npd, grüne, gruene, linke, #btw17, #btw2017, bundestagswahl, bundestag, wahlen, partei, stimme, groko, demokratie, wahlkampf, hochrechnung, #groko, jamaika, #jamaika, wahlergebnis
German Elections 3 $N = 85{,}689$	2018 Mar 02	2018 Mar 06	spd, cdu, csu, union, groko, #groko, grogo, #grogo, #nogroko, #spderneuern, merkel, neuwahlen, regierung, koalition, koalitionsvertrag, minderheitsregierung, jamaika, mitgliedervotum, bundestagswahl
Nothing $N = 61{,}831$	2018 Mar 01	2018 Mar 06	wasser, getränk, trinken, frühstück, mittagessen, abendessen, brunch, snack, essen, haus, wohnung, appartment, schlaf, schlafen

Word Frequency Norms. Baseline frequencies of words in German Internet language can be found in a collection of linguistically processed web corpora called DECOW [18, 19]. In order to standardize comparison between pairs with different absolute frequencies, the following approach was used, exploiting the inability to interlink modal particles, i.e., to combine them by using the German words *und* and *oder* (and, or) [2] (see Eq. 1). We compare the observed proportion in the data to the proportion in a collection of Internet language.

$$Proportion = \frac{Count\,Modal\,Particle}{Count\,und\,and\,oder + Count\,Modal\,Particle} \tag{1}$$

An example illustrates the approach: *bloß* (mere) appears 387,392 times in the DECOW14AX corpus, while *und* and *oder* appeared in sum 308,628,935 times. Thus, a proportion of 0.0013 results. The idea is that when less common ground is present fewer modal particles will be used. The amount of *und* and *oder* in this case can therefore either increase or stay the same. In both cases the proportion will decrease. If people are referring to a common ground and thus use more modal particles, the frequencies of *und* and *oder* can either decrease or stay the same. In both cases the proportion will increase. We note the small resulting proportions, which result in very small standard errors and hence narrow confidence intervals.

2.2 Tabulation

The occurrences of the modal particles and the words *und* and *oder* in the different data sets were counted using R. The proportions were then calculated for each modal particle in each data set as described in Sect. 2.1. This led to 15 proportions per corpus, 60 respectively, to compare with the baseline proportions.

2.3 Statistical Analysis

Because we are dealing with big data, each modal particle proportion in a data set was compared to the baseline proportion using effect size metrics. Effect sizes and their surrounding 95% confidence interval were calculated in R using the Cox Logit method [20]. An effect size became significant if the 95%-CI excluded 0.

Furthermore, resulting effect sizes for the elections were compared to effect sizes of the comparison group. Deviations in the *d*-values were considered significant in cases where the lower bound of the 95%-CI for higher *d*-values and the 95%-CI upper bound for lower *d*-values did not overlap. Relative to significance testing, this approach is rather conservative.

3 Results and Discussion

As in Hampton and Shalin [1], the observed proportions of the modal particles in the event corpora as well as in the "nothing" corpora are highly correlated, despite adjustments for the influence of a common baseline (see Table 3). This result fails to distinguish language style during the election from language usage during "nothing".

Table 3. Partial spearman rank correlations between proportions controlling for normative influence.

	Elections 2	Elections 3	Nothing
Elections 1	.82*	.68*	.68*
Elections 2		.73*	.70*
Elections 3			.43

Note. $N = 15$ for all comparisons. *$p < .05$.

Effect size analyses are more informative. All calculated effect sizes deviate significantly from the baseline (see Table 4). Positive *d* indicate an increase in the modal whereas negative *d* indicate a decrease. Table 4 also shows the effect sizes in the election that differ significantly from the "nothing" corpus effect sizes. Moreover, effect sizes increase for the three election related events relative to the "nothing" data in 29 out of 45 cases ($P(K \geq 29, n = 45, p = 0.5) = 0.036$). Another 7 comparisons go in the same direction but were not significant, i.e., CIs overlapped. Just 6 modal particles decreased and are thus contradictory. Twenty four percent of the 45 effect sizes in the election corpora are small—below 0.20. Nearly 50% of the 15 effect sizes in the "nothing" corpus are below 0.20. This provides evidence for engagement specific to the election corpora.

Table 4. Effect size departures from norm by election event and control group.

Modal particle	Significant effect sizes *d*			
	Elections 1	Elections 2	Elections 3	Nothing
aber	0.20	0.22	**0.14**	0.19
auch	**0.04**	**0.10**	**0.02**	−0.02
bloß	**0.67**	**0.71**	**0.62**	0.35
denn	**0.13**	**0.11**	**0.14**	−0.19
doch	**0.40**	**0.51**	**0.45**	0.18
eben	**0.40**	**0.35**	**0.29**	0.14
eigentlich	**0.45**	0.44	**0.49**	0.39
etwa	−0.44	−0.44	−0.51	−0.50
halt	**0.61**	0.76	**0.53**	0.78
ja	**0.51**	**0.59**	**0.72**	0.43
mal	**0.43**	**0.43**	**0.37**	0.50
nur	**0.35**	**0.40**	**0.31**	0.21
schon	**0.34**	**0.39**	**0.34**	0.24
vielleicht	0.14	0.13	0.12	0.12
wohl	**0.48**	**0.57**	**0.48**	0.12

Note. All shown *d* are significant relative to norms. Positive *d* indicate an increase in the modal particle whereas negative *d* indicate a decrease. Bold *d* for the elections 1–3 are significantly different from the "nothing" *d* as indicated by non-overlapping confidence intervals.

4 Contributions

Twitter exchange, as measured by the presence of conversational modal particles, does differ from broad Internet language. The preponderance of significant differences in all corpora using conversational words is consistent with the view that Twitter constitutes a form of conversation [5]. However, modal particles in the three 2017 German election events also generally increased relative to the "nothing" corpus. Thus, these departures from Internet standards are not simply an artifact of Twitter. We interpret these indicators of common ground, point of view and anaphora as a measure of public engagement in a common event. Based on the observed differences, exchange regarding the German elections constitutes a collective narrative relative to an exchange with respect to "nothing".

Moreover, typically unexploited stop words are surely not nothing. In lieu of computational data driven methods, we employ linguistic, psycholinguistic and psychological theory to pre-select (and therefore interpret) our feature set. Our analysis of stop words expands the metrics of general social media analysis, providing a general feature that, now identified, could be combined with more conventional computational text mining. Stop words contain meaning; they need not be ignored. As in [5], focusing also on style and structure rather than only content can provide a first step of data analysis, of relevance to mining public opinion regarding virtually any consequential topic such as harassment, immigration, global warming or disaster response. Like sentiment, our metric is domain independent. Unlike sentiment analysis, our approach comes with an underlying statistical and social science rationale that assists in interpretation, facilitating the comparison of engagement between events.

Acknowledgments. The first author recognizes support from the German Academic Exchange Service (DAAD). The third author’s participation in this research was partially supported by NSF grants CNS 1513721 and EAR 1520870.

References

1. Hampton, A.J., Shalin, V.L.: Sentinels of breach: lexical choice as a metric of urgency. Hum. Factors **59**(4), 505–519 (2017). Davis, K. (ed.) Special Issue on Big Data/Winner of the 2016 Human Factors Prize
2. Thurmair, M.: Modalpartikeln und ihre Kombinationen. De Gruyter, Berlin (1989)
3. Bross, F.: German modal particles and the common ground. Helikon. Multidiscip. Online J. **2**, 182–209 (2012)
4. Pang, B., Lee, L.: Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124. Association for Computational Linguistics, Stroudsburg (2005). https://doi.org/10.3115/1219840.1219855
5. Purohit, H., Hampton, A., Shalin, V.L., Sheth, A.P., Flach, J.M., Bhatt, S.: What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination. Comput. Hum. Behav. **29**(6), 2438–2447 (2013)
6. Purohit, H., Hampton, A., Bhatt, S., Shalin, V.L., Sheth, A.P., Flach, J.M.: Identifying seekers and suppliers in social media communities to support crisis coordination. Comput. Support. Coop. Work (CSCW) **23**(4–6), 513–545 (2014)
7. Purohit, H., Banerjee, T., Hampton, A., Shalin, V.L., Bhandutia, N., Sheth, A.: Gender-based violence in 140 characters or fewer: a #BigData case study of Twitter. First Monday **21**(1–4) (2016)
8. Ebrahimi, M., Yazdavar, A.H., Sheth, A.: On the challenges of sentiment analysis for dynamic events. IEEE Intell. Syst. **32**(5), 70–75 (2017)
9. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., Kappas, A.: Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol. **61**(12), 2544–2558 (2010)
10. Thelwall, M., Buckley, K., Paltoglou, G.: Sentiment in Twitter events. J. Am. Soc. Inform. Sci. Technol. **62**(2), 406–418 (2011)
11. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends® Inf. Retr. **2**(1–2), 1–135 (2008)
12. Weydt, H.: Abtönungspartikel: die deutschen Modalwörter und ihre französischen Entsprechungen. Gehlen (1969)
13. Weydt, H., Hentschel, E.: Kleines Abtönungswörterbuch. In: Weydt, H. (ed.) Partikel und Interaktion, pp. 3–24. Niemeyer, Tübingen (1983)
14. Helbig, G.: Lexikon deutscher Partikeln. Verlag Enzyklopädie (1988)
15. Diewald, G.: Abtönungspartikel. In: Hoffmann, L. (ed.) Handbuch der deutschen Wortarten, pp. 117–142. De Gruyter, Berlin, New York (2007)
16. Götze, M., Geyer, S.: https://solariz.de/de/downloads/6/german-enhanced-stopwords.htm. Accessed 10 Mar 2018

17. Sheth, A., Jadhav, A., Kapanipathi, P., Lu, C., Purohit, H., Smith, G.A., Wang, W.: Twitris: a system for collective social intelligence. In: Alhajj, R., Rokne, J. (eds.) Encyclopedia of Social Network Analysis and Mining, pp. 2240–2253. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-6170-8_345
18. Schäfer, R.: Processing and querying large web corpora with the COW14 architecture. In: Bański, P., Biber, H., Breiteneder, E., Kupietz, M., Lüngen, H., Witt, A. (eds.) Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3). IDS, Lancaster (2015)
19. Schäfer, R., Bildhauer, F.: Building large corpora from the web using a new efficient tool Chain. In: Calzolari, N., Choukri, K., Declerck, T., Doğan, M.U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012), pp. 486–493. European Language Resources Association (ELRA), Istanbul (2012)
20. Lipsey, M.W., Wilson, D.B.: Practical Meta-Analysis. Sage Publications Inc., Thousand Oaks (2001)

Beaten Up on Twitter? Exploring Fake News and Satirical Responses During the *Black Panther* Movie Event

Matthew Babcock(✉), David M. Beskow, and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{mbabcock,dbeskow}@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract. The use of user-generated online satire, itself a form of fake news, may be one strategy used to highlight and shame fake news stories and promoters. Here, we begin to explore the differences between non-satire and satire fake posts by looking at Twitter data related to false stories of racially-motivated attacks during the *Black Panther* movie opening. Overall, we found that very few fake tweets of either type had high levels of replies or retweets. We found some evidence that the satire responses were supported and shared to a greater extent than the original non-satire tweets, which leaves open the possibility that satire may have been helpful in calling out the fake attack posts. We also found some evidence that the satire responses fooled some users into believing them to be real stories.

Keywords: Fake news · Satire · Twitter

1 Introduction

The intentional and unintentional spread of false information on the internet has been the subject of continual and increasing public discussions, policy debates, and academic research. Twitter in particular has been studied as a medium in which "fake news" stories and campaigns can find footing and flourish [1–3].

With the growing amount of research and public debate has come an increased interest in whether and in which way policy makers, institutions, and the public should respond to the spread of false information. A recent Policy Forum article in the March 9, 2018 issue of Science summarizes possible interventions into two types: empowering individuals and platform-based detection and intervention [4].

A type of intervention that does not fall neatly into either of the two categories involves community-based intervention and correction. Such community-based interventions involve going beyond empowering individuals to correctly evaluate false information. It additionally involves having those individuals act to call out, mitigate, or otherwise attempt to control the spread of false information, whether on their own or in concert with others. Community-led efforts can help highlight and correct, and thus perhaps control the spread of false or misleading information [5, 6]. If independent, non-government and non-platform-directed communities are involved in the calling out, halting, and/or correcting of fake news cascades, it will be important to describe the advantages and disadvantages of different types of responses communities can engage in.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 97–103, 2018.
https://doi.org/10.1007/978-3-319-93372-6_12

One possible community-based intervention involves the use of satire. Satire itself is a form of disinformation that seeks to expose and/or ridicule its target. Satire can be produced and has been studied at both the level of professional mass media and at the level of user-generated content in a specific community or social media ecosystem. Much of the recent research on satire has focused on professional mass media satire such as *The Daily Show*¸ *The Colbert Report*, and *The Onion* and how such satire impacts knowledge and perceptions of individual issues (for example see [7]).

The use of user-generated satire to specifically constrain the spread of other disinformation online has not been as thoroughly studied. The use of satire may assist in the control of the spread of other disinformation by either increasing the number of people who are exposed to the false information within the context of ridiculing it or by shaming those that spread disinformation to halt or constrain their activities. Being false information by design, satire may also be at a disadvantage as a tool to combat the spread of fake news. In fact, a recent study by Horne and Adali suggests that fake news articles are more closely related in complexity and style to satire articles than to "true" news articles [8]. If satire aimed at fake news is itself considered fake news it may only serve to spread additional false information or drain resources from attempts to control "real" fake news. User-generated satire may be susceptible in different ways from professional satire to this issue.

The preliminary research presented in this paper is therefore focused on exploring the differences and relationships between non-satirical "fake news" and satirical responses on Twitter. Using Twitter data related to the release of the Marvel comic book movie *Black Panther*, we specifically explored the retweeting and reply activity related to both types of fake posts, the presence of bots who tweet fake posts, and the network created between Twitter users responsible for such posts.

1.1 Event Background

Marvel Studios' *Black Panther* movie opened to on February 16, 2018 and tells the story of the Marvel Comics superhero of the same name, who becomes the king and protector of the hidden and technologically-advanced fictional African nation of Wakanda. *Black Panther* was the first movie in the Marvel Cinematic Universe series (and first superhero comic book movie in general) to have a predominately African and African-American cast and creative team, a fact promoted both by Marvel's parent company, Disney (who intentionally released the movie during Black History Month in the United States), and on social media prior to and during the release.

Early showings of the film began the evening of February 15. On the morning of February 16, it was reported by Buzzfeed that there had been a series of twitter posts claiming the user or their friends or family had been physically attacked attempting to see *Black Panther* [9]. Buzzfeed also reported that other Twitter users had quickly posted replies proving that images used in the original posts had been taken from other news and entertainment media. These response tweets called out the original posts as fake stories aimed at stoking racial conflict (most depicted white family members being attacked by black moviegoers, and some depicted the opposite) and tarnishing the film's reputation. Later the same day, Vox reported that in addition to posts debunking and calling out the original false beating tweets directly, some Twitter users were also mocking the original tweets by posting their own versions using either more clearly unrelated photos or additionally unbelievable language [10]. Additional news reports mentioned that some of these satire posts were being treated as if they were examples of the original fake posts.

2 Methods and Results

2.1 Data Collection and Analysis

Online news articles discussing the non-satire and satire tweets were collected using Google search, with the search terms, “Black Panther Fake News”. A preliminary set of non-satire and satire tweets were identified from these articles.

We collected all tweets containing “#BlackPanther” that were posted from February 8 to February 23. We additionally collected tweets containing phrases found in the tweets mentioned in the news articles. By searching our combined collected tweets for those that contained such phrases (e.g. “black youths”, “MAGA hats”, “cracker”) but not response phrases (e.g. “fake”, “troll”, “racist”) and then reviewing our search results, we identified a total of 249 distinct fake tweets (from 238 distinct screen names), 178 which we labeled as satire and 71 which we labeled as non-satire. We then additionally collected tweets that replied to or were retweets of any of the 238 “fake” tweeters.

Satire posts were distinguished from non-satire by manual review. Posts containing images from cartoons, movies, and classical art that depicted unrealistic violence or unrelated content were labeled as satire. Posts containing text describing unrealistic events (e.g. atomic bombs) and stories that started in a similar fashion to the fake beatings but ended positively (e.g. “we were approached by black youths…who then proceeded to give us high-fives”) were also labeled as satire.

The combined dataset contains a total of 5,151,935 individual tweets. We created a subset of 291,111 tweets that included all fake posts (non-satire and satire), all retweets and replies to those posts, and all retweets and replies to any other posts by the same users who posted the fake stories.

2.2 Retweets and Replies of Satire and Non-satire Tweets

We counted the retweets and replies in our dataset for each of the 178 satire and 71 non-satire tweets. Due to account suspension and/or tweet deletion, we were unable to identify the number of retweets and replies for 28% of non-satire tweets and 7% of the satire tweets. The satire tweets were in total retweeted 47,512 times and replied to 709 times. The non-satire tweets were in total retweeted 1,916 times and replied to 2,983 times. Table 1 shows that the percentages of retweets of satire and non-satire tweets are relatively similar within each class, except for the case of class 1 in which a larger percentage of the satire tweets are categorized. The same is true of the percentages of replies to satire and non-satire posts. It should be noted that less than 5% of the fake tweets were either retweeted or replied to more than 100 times.

Table 1. Percentage of total retweets and replies that fall within each count class for satire (n = 178) and non-satire (n = 71) tweets. Classes are defined as 1 (0 counts), 2 (1–10), 3 (11–100), 4 (101–1,000), 5(1,001–10,000), and 6 (10,001–100,000).

	Class						
	1	2	3	4	5	6	Unknown
RTs: non-satire	25%	32%	10%	4%	0%	0%	28%
RTs: satire	40%	46%	6%	1%	0%	1%	7%
Replies: non-satire	34%	27%	7%	3%	1%	0%	28%
Replies: satire	62%	29%	1%	1%	0%	0%	7%

We plotted the cumulative sum of replies and retweets over time for individual satire and non-satire tweets focusing on tweets that were mentioned in the news media and/or were representative of classes 3 through 6. For the top non-satire tweet, Fig. 1 (left plot) shows that the replies outweigh the retweets by an order of magnitude. Other non-satire tweets show similar patterns in that the growth in replies to the tweet outpace retweets (though not always by orders of magnitude) and in that the retweets level off sooner.

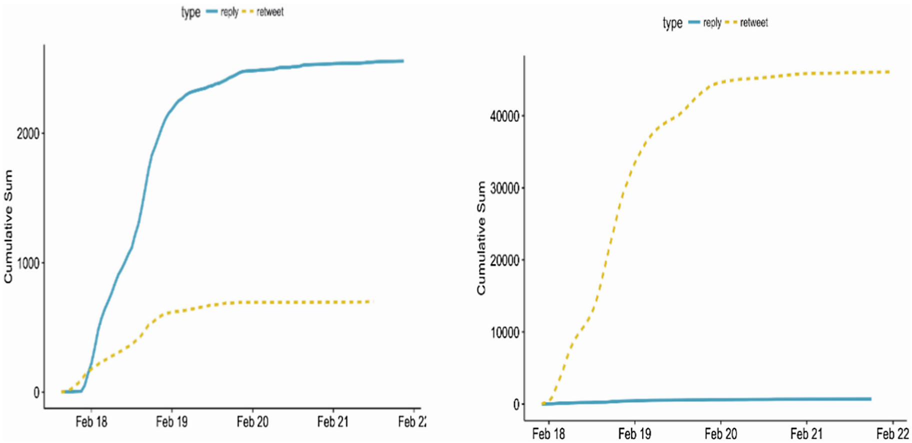


Fig. 1. Cumulative sum of retweets and replies to top non-satire tweet (left plot) and top satire tweet (right plot).

For the top satire tweet, Fig. 1 (right plot) shows a different relationship, where the retweets outweigh the replies by two orders of magnitude. Other top satire tweets show a similar pattern, with the growth in retweets outpacing replies and the replies leveling off sooner. We manually read through the response thread to the top non-satire tweet and a large majority of the replies were calling out the non-satire user for posting “fake” information. This appeared to be the case with other non-satire tweet threads that we looked at. We also read through the responses to the top satire tweet and found the replies to be a mix of supporters acknowledging the satire and some responses that attacked the satire tweet as if it was one of the non-satire tweets.

2.3 Bot Detection

We used CMU Bothunter, an integrated ML approach, to categorize the 238 individual user accounts who posted a fake tweet as bots or not. Due to accounts being deleted and/ or suspended prior to our running the bot detection algorithms, only 187 user accounts could be categorized. A total of 12 (6.7%) of the satire accounts and 2 (2.3%) of the non-satire accounts were categorized as bots. Within classes 3–6 from Table 1 there was only one satire account and no non-satire accounts that were categorized as bots.

2.4 Network of Satire and Non-satire Posters Over Time

We created a dynamic network with the 238 fake posting accounts as the nodes and where an edge exists between nodes if either of the users retweeted, replied, or otherwise mentioned the other. Figure 2 shows the cumulative progression of the network at each of 4 days.

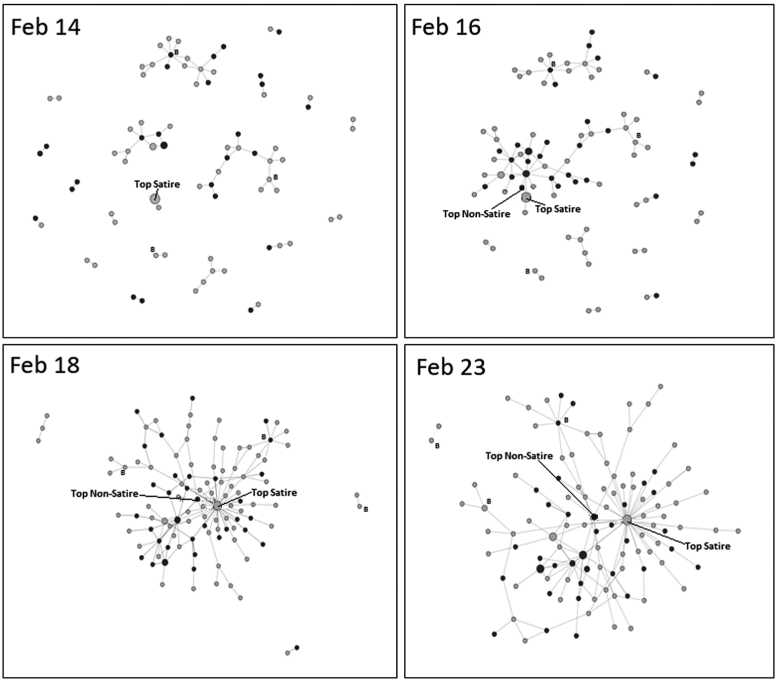


Fig. 2. Network of users who posted satire (grey) and non-satire (black) tweets. Isolates have been removed for clarity. "B" represents the three users in this network that were classified as bots (other bots were isolates). On February 14th, prior to the posting of any fake attack posts, there were three main components, and the top satire user existed in the network but was only connected to one other user. On February 16th, opening day and the day of the first news media reports, two of the main components were connected to each other and to both the top satire and top non-satire user. By February 18th, after the top satire tweet was posted, that user has become the center of the main component of the network. Between February 18th and the 23rd, only 12 additional network connections were made.

3 Discussion and Conclusions

Our preliminary results show that the fake stories of racially-motivated attacks – including both satire and non-satire versions – make up a small fraction of the overall conversation surrounding the Black Panther movie. This is even though many online news and social media outlets covered the story. The sets of identified satire and non-satire tweets were found to be similar in that only a small percentage of each type had high levels of retweeting and reply activity.

The comparison of retweets and replies made in response to individual satire and non-satire tweets suggests that in general the satire tweets were supported and spread by the community while the non-satire tweets were mostly called out and the posters shamed or attacked. This may be an indication of community peer pressure successfully mitigating the spread of non-satire fake news. The fact that response posts began before the first news story ran is also indicative of community self-correction. On the other hand, the high retweeting of specific satire posts may be leading to confusion for those that don't get the pop-culture jokes at the heart of many of such posts. This may also be making the overall "fake attack" story appear larger than it is.

The fact that the top satire tweet becomes the central network node of the main discussion amongst those that post either kind of fake story is interesting in part because it connects both satire and non-satire tweeters. The three bot accounts do not appear to have played a large role in making connections in this network. Further exploring the directionality of the network and expanding the analysis to the larger one-hop network (including all additionally mentions, retweets, and replies to the 249 fake posts) will help to describe how top satire posts may be bringing parts of the networks together. Future work will also include exploring the non-satire responses to the original fake stories as such responses started earlier and have the advantage of being less likely to be confused for the story they are attacking - though it remains to be seen if they spread as fast and/or deep as the satire responses. Exploring a more detailed timeline may provide an indication as to whether the news drove additional activity or active posts gained the attention of the news media.

There is uncertainty in the total number of fake posts of both kinds in our dataset due to our use of keyword searches based on the news articles. This is somewhat mitigated by the fact that many of the false posts that are worth exploring further due to their number of retweets and mentions are ones that the news media picked up. There is also some uncertainty in the labeling of satire posts, as we could not confirm the intent of such posts. This preliminary work is additionally limited in that we only currently have access to Twitter data and therefore are missing network connections between users on other social media. Deleted tweets and suspended accounts also inhibited some of the data collection and bot categorization.

Acknowledgements. This work was supported in part by the Office of Naval Research (ONR) under grants - N000141812108 and N00014-17-1-2605/25323818CMU and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

References

1. Chamberlain, P.: Twitter as a Vector for Disinformation. School of Computer & Security Science, Edith Cowan University, Australia (2009)
2. Shao, C., Ciampaglia, G.L., Varol, O., Flammini, A., Menczer, F.: The spread of fake news by social bots. arXiv preprint arXiv:1707.07592 (2017)
3. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**, 1146–1151 (2018)
4. Lazer, D.M., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D.: The science of fake news. Science **359**, 1094–1096 (2018)
5. Arif, A., Robinson, J.J., Stanek, S.A., Fichet, E.S., Townsend, P., Worku, Z., Starbird, K.: A closer look at the self-correcting crowd: examining corrections in online rumors. Presented at the Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (2017)
6. Dailey, D., Starbird, K.: Visible skepticism: community vetting after Hurricane Irene. In: ISCRAM (2014)
7. Day, A.: Breaking boundaries—shifting the conversation: colbert's super PAC and the measurement of satirical efficacy. Int. J. Commun. **7**, 414–429 (2013)
8. Horne, B.D., Adali, S.: This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. arXiv preprint arXiv:1703.09398 (2017)
9. Silverman, C.: Trolls are Posting Fake Claims of Being Assaulted at Showings of "Black Panther". https://www.buzzfeed.com/craigsilverman/trolls-are-posting-fake-claims-of-being-assaulted-at
10. Romano, A.: Racist trolls are saying Black Panther fans attacked them. They're lying. https://www.vox.com/culture/2018/2/16/17020230/black-panther-movie-theater-attacks-fake-trolls

#metoo Through the Lens of Social Media

Lydia Manikonda(✉), Ghazaleh Beigi, Subbarao Kambhampati, and Huan Liu

Arizona State University, Tempe, AZ, USA
{lmanikon,gbeigi,rao,huan.liu}@asu.edu

Abstract. Sexual abuse – a highly stigmatized topic in the society has spurred a revolution in the recent days especially through the shared posts on social media platforms via attaching the hashtag #metoo. Individuals from different backgrounds and ethnicities began sharing on the online venues about their personal experiences of getting sexually assaulted. This paper makes an initial attempt to asses the public reactions and emotions by utilizing the publicly shared #metoo posts by performing a comparative analysis of the tweets shared on Twitter as well as on Reddit. Though nearly equal ratios of negative and positive posts are shared on both platforms, Reddit posts are focused on the sexual assaults within families and workplaces while Twitter posts are on showing empathy and encouraging others to continue the #metoo movement. The data collected in this research helps in the preliminary analysis of the user engagement, discussion topics, word connotations and sentiment with respect to the #metoo movement.

1 Introduction

Sexual abuse has been traditionally buried due to the fear of shame, retribution and retaliation. Sexual abuse, and abuse in general is a very difficult topic for individuals to talk about, irrespective of an online or an offline setting [9,10]. In United States itself, on an average there are 321,500 victims (age 12 or older) of rape and sexual assault each year where, ages 12–34 are the highest risk years. The trauma of the sexual abuse has resulted in the long-term negative impacts such as anxiety, suicidal behavior, PTSD, panic disorder, mood and behavioral disorder problems [3].

This Stigmatized topic – sexual abuse, has gained a lot of attention recently especially with individuals self-disclosing their personal experiences on online social media platforms. While many personal stories have gained attention from the media and general public as the *#metoo* movement[1] where, most of the reports that gained attention were only the experiences of few people. With the online venues enabling individuals to maintain privacy and to self-disclose

[1] Although the term *#metoo* was originally coined in 2006 by social activists to raise awareness about sexual abuse, it became viral in October 2017, following the alleged sexual misconduct in the Hollywood.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 104–110, 2018.
https://doi.org/10.1007/978-3-319-93372-6_13

their true feelings [5–8], numerous individuals joined this movement through sharing their personal experiences and opinions about sexual abuse. The shared posts include different types of experiences related to sexual abuse as well as the opinions about how to bring awareness in the society to combat such issues. It is not very clear what the individuals are sharing through these posts as majority of the media coverages are about few individuals who are popular. Through a comparative analysis of posts shared on Twitter and Reddit, this paper provides a preliminary analysis on the demographics, user engagement, discussion topics, word connotations and sentiment.

2 Data

We obtained two sets of data from Twitter and Reddit using their corresponding python APIs – https://goo.gl/P6GoFy and https://goo.gl/F3981i respectively. We collect 620,348 posts from 205,489 users on Twitter and 190 posts from 70 users on Reddit. On Twitter, we crawl the public posts (from October 2017 to January 2018) that are attached with the #metoo hashtag where as for Reddit, we crawl all the self posts shared on */r/metoo* subreddit. The data includes all the meta-data associated with the post.

3 Social Engagement

Table 1. Statistics about the engagement attributes

Eng. Att.	Mean	Min	Max	Std
Favorites	5.69	0	104464	229.14
Retweets	2.38	0	22893	71.79
Mentions	1.13	1	25	0.81
Hashtags	1.93	1	36	2.12

Due to the sensitivity of the topic and specifically the viral nature of the #metoo hashtag, we first want to investigate how the tweets shared on this topic engage other users on Twitter. We compute statistics about the engagement attributes that include – number of favorites these tweets received, number of times a tweet is retweeted, number of mentions in the tweets and the number of hashtags attached to these tweets. Table 1 shows that on an average these tweets receive atleast 5 favorites and 2 retweets which is relatively more engaging compared to general tweets [12]. This might be due to other users endorsing the tweets. On the other hand, it is surprising to see that users tend to engage in conversations with other users or atleast mention them more prominently.

Figure 1 shows the log-log plot of these engagement attributes shedding light on the favorites and retweets received by these posts. All the engagement attributes follow a power-law distribution showing that there exist few posts which are very highly engaging relative to the majority of the remaining posts. Complementing these observations, Reddit posts which can receive both up votes and down votes, receive 2.26 up votes (standard deviation = 2.03) on an average. None of the *self* posts shared on #metoo subreddit received down votes suggesting that posts shared on Reddit are positively engaging.

4 Linguistic Themes to Understand the Content

Since #metoo related posts are socially engaging, it is important to understand the content of these posts. We first extract the latent topics present in these posts and then focus on how users label sexual abuse through their vocabulary usage.

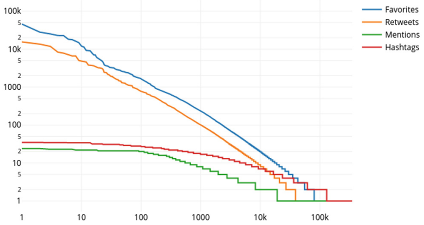


Fig. 1. Log-Log plot for engagement attributes on Twitter

4.1 Latent Topic Extraction

We extract the latent topics from the corpus containing all the *self* posts shared on #metoo subreddit as well as the corpus of all posts attached with the #metoo hashtag on Twitter. Topic analysis helps us understand the aspects of sexual abuse that the individuals are focusing on the two platforms. We use LDA (Latent Dirichlet Allocation) topic modeling technique [1] to extract the latent topics shown in Table 2.a and .b for Reddit and Twitter respectively.

On both these platforms, people share their experiences of getting assaulted (topic 0 in Reddit and topic 2 in Twitter). Users also encourage each other to be strong and fight against harassment by contributing to the movement (topic 2 in Reddit and topic 0 in Twitter). However, we notice two significant differences in

Table 2. Topic vocabulary.

(a) Reddit		(b) Twitter	
Topic	Top words	Topic	Top words
0	[Experience and memory, emotions] emotional, response, attacker, unwelcome, crime, severe, mugged, notmeanymore, starting, threat	0	[Fight against harassment] #metoo, harassment, movement, assault, campaign, silence, violence, teaching, abuse, business, #timesup, schools, workplace
1	[Story Details] date, shoulder, apartment, squad, strange, morning, van, club, escape, partner	1	[Sharing news] stories, weinstein, damon, harvey, share, backlash, mcgowan, allegations, news, misconduct, hollywood, accused, pbs
2	[Fight against harassment, being strong] movement, victims, survivor, abused, metooers, damage, accusations, battle, strength, suffer, wounds, trigger	2	[Sharing support by posting hashtags] #metoo, #timesup, #goldenglobes, black, #oprah, #resist, winfrey, #oprah2020, hollywood, president, #millennials, #veterans
3	[Story Details] home, remember, hell, car, boyfriend, hotel, officer, realtor, banker, fuck, weekend, conversation	3	[Real story sharing] lewinsky, bill, clinton, monica, #trump, accuser, video, #feminism, #rape, black, #maga, simmons, russell
4	[Offenders affiliation] head, older, house, attorney, military, hand, working, sally, face, army, black	4	[Discussing news] witch, hunt, social, campaign, harassment, world, woody, allen, reckoning

the types of posts shared on Reddit and Twitter. On Reddit, survivors mainly share the details of the story (e.g. how and when that happened to them) and how they were hurt emotionally (topics 0, 1 and 3). They also mention the affiliation of offenders (topic 4). While On Twitter, people do not expose the details and mainly focus on supporting victims of sexual violence by just posting relevant hashtags (topic 2 summarizes the mostly used hashtag during the movement), sharing relevant news (topic 1 and 3) and urls of related news articles (topic 4). On Twitter, users share their stories of being harassed at workplace and how they fear being retaliated for complaining about the harassment (topic 0). These differences in terms of being descriptive between Reddit and Twitter might be because of the character limit enforced by these platforms. In particular, Reddit has allowed users to share more details and thus users might be able to reveal their true feelings easier than Twitter. But it is interesting to see that this movement became viral due to the posts shared on Twitter[2].

4.2 Labeling Sexual Abuse

n-gram Analysis. To obtain a basic understanding of the content shared, we extract n-grams. Table 3 shows the *bi*-grams and *uni*-grams extracted from Twitter and Reddit posts. Bigrams show that majority of the Reddit posts focus on individual experiences about sexual harassment for example: *years old*, *Im sorry*, *one day*, etc., where as Twitter posts focus on the existing sexual assault stories and opinions about how to address these issues (*metoo movement*, *say metoo*, *social media*, etc.). Unigrams also highlight similar set of observations. Using the most frequently occurring keywords in these text corpuses, we dig a little deeper to understand how users label sexual abuse through the words associated with these keywords.

Table 3. Top-10 *uni*-grams and *bi*-grams

Twitter bigrams	metoo movement; sexual harassment; metoo timesup; metoo campaign; metoo moment; say metoo; metoo story; social media; witch hunt; sexual misconduct
Reddit bigrams	sexual harassment; dont want; sexual assault; years old; dont think; Im sorry; one day; metoo movement; first time; will never
Twitter unigrams	metoo; women; movement; sexual; men; harassment; now; assault; time; hollywood
Reddit unigrams	men; like; me; im; women; dont; people; know; time; sexual

Considering Syntactic as Well as Semantic Relationships. To ensure that both the syntactic and semantic relationships are captured, we represent the vocabulary of the corpus in a Word2Vec space and measure their similarities [4]. Through the pairwise word relationships shown in Table 4, the most frequently occurring keywords suggest that some of the terms such as *men* is associated with *aggressive* and *violate certain aspects* where as, *woman* is associated with *humiliated publicly* and *intimidated.* Whenever users mention their personal experiences (for example the term *story*), it is highly correlated with words such as *heartbreaking, frightening, terrifying, horrifying, awful,* etc. Alongside, most of the other keywords (such as *sex*, *rape*, *victims*, etc.) are similarly

[2] http://time.com/5051822/time-person-year-alyssa-milano-tarana-burke/.

Table 4. Semantic and syntactic co-occurrence patterns from tweets. Keywords are the 12 most frequent words. The right column shows the most co-occurred words associated with left.

Keyword	Most co-occurring words
men	aggressive, pigs, socialized, violate, proclaim, educated
story	heartbreaking, frightening, terrifying, horrifying, awful, painful, triggering, insightful
assault	prevention, policing, devaluing, mishandling, payouts, regrettable
sex	perform, oral, consensual, date, violent, nonconsensual
harassment	misconduct, rampant, ubiquity, experiencing, assaults, secrecy
#metoo	#spite, #mentalhealth, #gossip, #sexpredator, #activism, investigative
movement	travesty, witchhunt, hysterical, concerns, nonsense, ridiculously, damaging
timesup	oprahs, deathknell, globes, gowns, attendees, #golden, staged
woman	single, unconscious, humiliated, dragged, publicly, qualified, intimidated, backed
abuse	exploitation, stigma, secrecy, psychological, admitting, harassment, severity
victims	survivors, condemning, offenders, assistance, minimize, pedophiles, bystanders, prevent, suffering
rape	attempted, kits, marital, molestation, hookup, aggression, shame

associated with a vocabulary that is mostly negative. However, the users are also recognizing that these issues should be addressed immediately (see keyword *timesup*) and is slightly on an encouraging side compared to other keyword relationships. Words such as *movement* is co-occurring with words such as *hysterical*, *nonsense* and it is not very clear if users are mocking the #metoo movement that may require further analysis. Due to the limited set of posts crawled from Reddit, we didn't find any significant co-occurring patterns.

5 Individual Emotions Through Linguistic Markers

5.1 Emotion Attributes

We use psycholinguistic lexicon LIWC[3] to characterize and compare the type of emotions expressed on both the platforms. We obtain measures of the attributes related to user behavior: emotionality (how people are reacting that includes *sad*, *anger*, *anxiety*, *positive* and *negative* emotions), social relationships (*family*), and individual differences (*work*, *bio*,

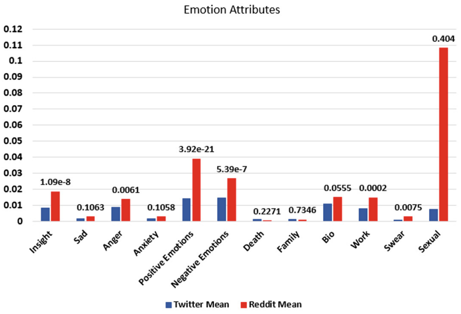


Fig. 2. Emotion attributes for Reddit and Twitter posts. Numbers on the bar show the p-value.

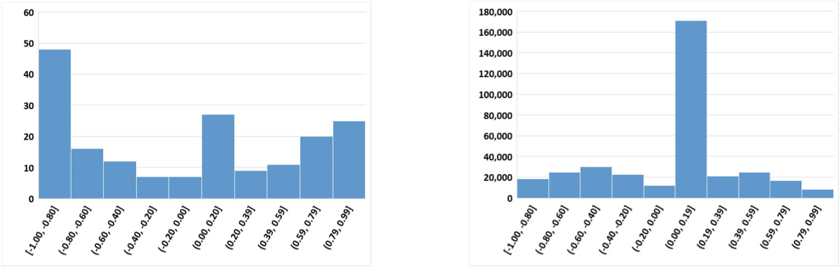
[3] http://liwc.wpengine.com/.

death, *swear*, *sexual*). For each attribute, we use the statistical *t*-test to check if the Twitter distribution is not significantly different from those of Reddit. Null hypothesis is rejected if p-value ≤ 0.05.

Figure 2 shows that the distribution of insight, anger, work, swear and positive and negative emotions attributes in Twitter are significantly different from those of Reddit. In contrast, posts on both platforms have the same distribution for *sadness* and *anxiety* attributes.

5.2 Sentiment Extraction

To measure the type of sentiment on both platforms, we use Vader [2] – a sentiment analysis tool designed specifically to extract sentiments from social media posts. Results are shown in Fig. 3 with the following observations. Reddit posts are generally more negative than Twitter posts. This might be because people have no limitation on their posts lengths and thus can easily share their feelings about their stories. However, few posts on Reddit express positive sentiment emphasizing to support the movement. Considering these platforms exclusively, the ratio of positive to negative posts on these platforms are equal to each other showcasing the presence of positivity towards the movement.



(a) Reddit **(b) Twitter**

Fig. 3. Sentiment distribution for Reddit and Twitter posts

6 Conclusions

In this paper, we focus on the shared posts of users attached with the #metoo hashtag on these platforms (i.e. Reddit and Twitter). The insights obtained from this research reveal the fundamental differences in the behaviors of individuals on these two platforms. Some of the key findings are: (1) users share their personal stories in details on Reddit while on Twitter, they tend to pursue other users to continue the #metoo movement; (2) Reddit posts are more negative while positive posts on Twitter showcase the presence of positivity towards the movement. These differences show that Twitter is a venue for sparking the movement while Reddit provides the chance for the people to share personal moments. We hope that our findings shed light on the important aspects associated with the sexual abuse which could initiate discussions between the individuals in the society as well as researchers and lawmakers.

Acknowledgements. This research is supported in part by a Google Faculty Research Award, an AFOSR grant FA9550-18-1-0067, the ONR grants N00014161-2892,N00014-13-1-0176,N00014-13-1-0519, N00014-15-1-2027, and the NASA grant NNX17AD06G.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
2. Gilbert, C.J.H.E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of Eighth International AAAI Conference on Weblogs and Social Media (2014)
3. Petrak, J., Hedge, B.: The Trauma of Sexual Assault: Treatment, Prevention and Practice (2002)
4. Tomas, M., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems (2013)
5. Goffman, E.: Stigma: Notes on the Management of Spoiled Identity. Simon and Schuster, New York (2009)
6. Altman, I., Taylor, D.A.: Social Penetration: The Development of Interpersonal Relationships. Holt, Rinehart & Winston, New York (1973)
7. Cozby, P.C.: Self-disclosure: a literature review. Psychol. Bull. **79**(2), 73 (1973)
8. Joinson, A.N., Paine, C.B.: Self-disclosure, privacy and the Internet. In: Oxford Handbook of Internet Psychology (2007)
9. Finkelhor, D., Gerald, H., Lewis, I.A., Smith, C.: Sexual abuse in a national survey of adult men and women: prevalence, characteristics, and risk factors. Child Abuse Negl. **14**(1), 19–28 (1990)
10. Coffey, P., Leitenberg, H., Henning, K., Turner, T., Bennett, R.T.: Mediators of the long-term impact of child sexual abuse: perceived stigma, betrayal, powerlessness, and self-blame. Child Abuse Negl. **20**(5), 447–455 (1996)
11. McClain, N., Amar, A.F.: Female survivors of child sexual abuse: finding voice through research participation. Issues Mental Health Nurs. **34**(7), 482–487 (2013)
12. Manikonda, L., Meduri, V.V., Kambhampati, S.: Tweeting the mind and instagramming the heart: exploring differentiated content sharing on social media. In: ICWSM, pp. 639–642 (2016)
13. Andalibi, N., Haimson, O.L., De Choudhury, M., Forte, A.: Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 3906–3918. ACM (2016)

An Agent-Based Model for False Belief Tasks: Belief Representation Systematic Approach (BRSA)

Zahrieh Yousefi[1](✉), Dietmar Heinke[1], Ian Apperly[1], and Peer-Olaf Siebers[2]

[1] School of Psychology, University of Birmingham, Birmingham, UK
z_yousefi@yahoo.com
[2] School of Computer Science, University of Nottingham, Nottingham, UK

Abstract. A meaningful social life relies on understanding others' minds and behaviours. The ability to reason about an individual's mental states such as beliefs and desires, and to understand and predict how these mental states shape an individual's behaviour is called theory of mind (ToM). In order to examine an individual's ToM ability, false belief tasks have been used widely in the literature.

This research is a novel attempt to clarify the basic cognitive processes shared across the different varieties of false belief tasks. For this purpose, an agent-based model has been implemented to evaluate how agents' achievement of goals in a social context is dependent on the ability to understand others' beliefs.

In our study, we offer a methodological framework for many belief-reasoning tasks called Belief Representation Systematic Approach (BRSA). BRSA is a simple and robust approach that breaks down false belief tasks into four fundamental cognitive phases, including Perception, Memory, Reasoning beliefs and desires, and Expressing others' beliefs and desires in an action. BRSA identifies a network of indispensable resources for belief representation. It also clarifies that there is a difference between 'understanding' others' beliefs and 'using' that understanding. BRSA demonstrates that false belief tasks, as a common decisive methodology for theory of mind competence, might involve more than understanding others' beliefs. Moreover, the model demonstrates that agents' understanding of others' beliefs on the micro level will lead to significant improvements in their performances on the macro level.

Keywords: False belief task · Theory of mind · Belief representation
Agent-based model

1 Introduction

Humans are a social species; they understand their own and others' behaviour in their day-to-day life. Their capability to make spontaneous inferences about the invisible thoughts and feelings of others enables them to communicate. Typically, humans have theory of mind ability; the ability to take others' mental states such as beliefs and desires into account and apply this coherent information to infer their actions (Frith 2012). People distinguish others' beliefs, desires and goals from their own. They understand and predict others' actions by reasoning about their beliefs and desires. From its inception, theory of mind research has evolved with various experiments to assess different

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 111–126, 2018.
https://doi.org/10.1007/978-3-319-93372-6_14

characteristics of theory of mind abilities in non-human and human children. However, one type of task, the false belief task, has been the most frequently used and has widely stimulated developmental research (e.g. Apperly 2011; Doherty 2009; Bloom and German 2000). Baron-Cohen et al. (1997) introduced the 'Sally and Ann false belief task' which is identified as the standard false belief task to test children's ability in understanding others' beliefs. The scenario of the task includes two puppets: Ann and Sally, a ball, a basket and a box. The subject child being tested for belief representation watches as Sally puts the ball in the basket. Then Sally leaves the room and when Sally is out, Ann moves the ball from the basket into the box. Sally returns, and the child is asked where Sally will look for the ball. The correct answer to this question is the basket where Sally put the ball. The child needs to answer the question correctly to pass the false belief test. The general design of false belief tasks solves the behavioural aspect of understanding others' beliefs by introducing two separate beliefs about the location of an object (the ball); one is the real location of the object and another is the protagonist's (Sally's) perspective, which is a false belief about the location of the object.

The results of verbal false belief tasks, for example Sally and Ann false belief task, show that children under the age of 4 years fail the verbal false belief tasks (e.g. Wellman et al. 2001; Call and Tomasello 2008; Wellman 2014). Verbal (explicit) false belief tasks are naturally complex, as they require the integration of linguistic information. In addition, in the process of tracking events from the protagonist's point of view, these tasks may cause disruption for children (Rubio-Fernández 2013). Thus, researchers started to design non-verbal (implicit) false belief tasks with less cognitive demands to test infants' belief representation competence. Onishi and Baillargeon (2005) developed a non-verbal false belief task for 15-month-old infants concerning the change of location of an object and measuring their looking time. The findings from this experiment and others (e.g. Southgate et al. 2007; Surian et al. 2007; Baillargeon et al. 2010; Kovács et al. 2010; Southgate and Vernetti 2014) indicate that infants are able to pass non-verbal false belief tasks as young as 7 months, and in any case well below 4 years old. This contradiction has been a pivotal debate in developmental literature, producing fruitful research.

Historically, the literature has expanded with the diversity of false belief tasks; however, there has been very little consensus on core principles. The design of false belief tasks sometimes contains ambiguity or complexity, which makes it difficult to accurately interpret the experimental results. Despite an increasing number of studies, false belief literature lacks a systematic approach to its basic processes, leading to confusion in many of the experiments and the results. Our study therefore seeks to clarify the processes of understanding others' false beliefs and addresses a key question: Which sets of basic processes are shared across the different varieties of false belief tasks?

The main objective of false belief tasks was to examine children's ability of recognizing the perspectives of others in contrast to a differing real world state. This might be the underlying reason why understanding others' false belief is considered as an acid test for the presence of theory of mind ability (e.g. Wellman and Bartsch 1988; Doherty 2009; Workman and Reader 2014). In contrast, Bloom and German (2000) explain two reasons why the false belief task needs to be abandoned as a test for theory of mind. The first reason given is that to successfully pass a false belief task requires abilities other than ToM. The second is that ToM ability does not require the ability to reason about false beliefs. This discrepancy in the literature drives us to explore the association between ToM and false belief tasks in more detail.

Our study presents a computational model for false belief tasks to address the above inconsistencies in the literature. It also explores some of the advantages and costs of understanding others' beliefs. For this purpose, an agent-based model called 'Belief Representation Model' (BRM) was designed to shed light on false belief's processes at both micro and macro levels. On the micro level, BRM examines the concept of belief representation, procedures and the minimum resources it might require in a dynamic environment. On the macro level, the aggregated results of BRM are comparable with the empirical effects of passing or failing false belief tasks in a virtual society; the BRM simulation results reflect the effect of understanding the beliefs of others in agents' performances.

The underpinning premise of BRM was inspired by the study of Martin and Santos (2014). In their experiment the participants, rhesus macaque monkeys, saw scenarios in which a human mediator was watching an apple moving between two boxes. They provided different scenarios of true and false beliefs about the final location of the apple for both the monkeys and the human mediator by occluding parts of the apple's movement from either the monkey or the mediator. The results suggest that monkeys fail to represent others' beliefs whereas human infants pass the experiment's test and demonstrate belief representation (Martin and Santos 2014). Martin and Santos (2016) argue that primates' belief representation is limited to the relations between mediators and information that is true and they are unable to represent relations between mediators and untrue information. They suggest that belief representation may be unique to humans as part of their core knowledge systems with automatic process that enable human infants to make sense of their physical and social environments (Martin and Santos 2014).

In our study motivated by the experiment of Martin and Santos, we offer a methodological framework for many belief-reasoning tasks called Belief Representation Systematic Approach (BRSA). BRSA is a simple and robust approach that breaks down false belief tasks into four fundamental cognitive phases, including Perception, Memory, Reasoning beliefs and desires, and Expressing others' beliefs and desires in an action. These collective phases identify a network of indispensable resources for belief representation. BRSA clarifies the difference between 'understanding' others' beliefs and 'using' that understanding. BRSA also demonstrates that false belief tasks, as a common decisive methodology for theory of mind competence, might involve more than understanding others' beliefs. In addition, the model demonstrates that agents' understanding of others' beliefs on the micro level will lead to significant improvements in their performances on the macro level. To the best of the authors' knowledge, our paper is the first attempt to explain underlying processes of understanding the beliefs of others through an agent-based model.

2 The BRM Implementation

BRM is implemented in the Repast Simphony (Repast 2017), an integrated open source Java-based modelling platform. The Unified Modelling Language (UML) class diagram of BRM is similar to the Stupid Model (Bersini 2012) with different objectives and functions. Also, there are common features between BRM and the predator-prey model and the SugarScape model (Epstein and Axtell 1996), but they differentiate in their aims and domains. For example, BRM as a generative psychology model does not benefit from the growth or decline of the agents' populations. Moreover, statistical methods are usually required if results are very noisy and effects are not clear-cut. The parameters of BRM settings have been systematically changed and have provided clear results. Hence, the application of statistical methods is not required when interpreting the BRM simulation results.

3 BRM Methodology

BRM consists of two types of reactive agents interacting within the environment: Monkey and Infant agents. The names are based on the experiment by Martin and Santos (2014) indicating two different capabilities in regard to understanding others' false beliefs; Infant agents represent the ability to understand others' false beliefs whereas Monkey agents are able to remember and track information from the past but lack the ability to understand others' beliefs.

Agents in BRM are randomly placed in a toroidal grid space of 50 by 50 and the goal of agents is to consume food. The first neighbourhood of agents refers to the Moore neighbourhood consisting of 8 cells around them. The second and third neighbourhoods are an expansion of the Moore neighbourhood to 24 and 48 cells respectively. The agents' field of view defines the agents' visual perception in the simulation. The agents' field of movement, the area in which an agent can move, consists of the cells within the first neighbourhood. The unit of time is called the 'tick' and it is the time in which each agent's action is scheduled. Each simulation by default consists of 1,000 ticks. Throughout each simulation, the number of food in the environment remains constant in every time step.

3.1 Monkey Agents' Strategy

Monkey agents observe, collect and store information about the location of food in their first neighbourhood. They are able to remember the location of food from the previous time step. They are egocentric and lack the ability to consider others' perspectives. Monkey agents' strategy comprises three phases: collecting information, recording information and action, which is illustrated in Fig. 1. In the collecting information phase, Monkey agents observe the environment and collect information about the location of the food. In the recording information phase, they store the location of the food. In the action phase, Monkey agents randomly choose food and consume it. When there is no food available, they move towards the location of the food which has been stored in their memory from the previous time step.

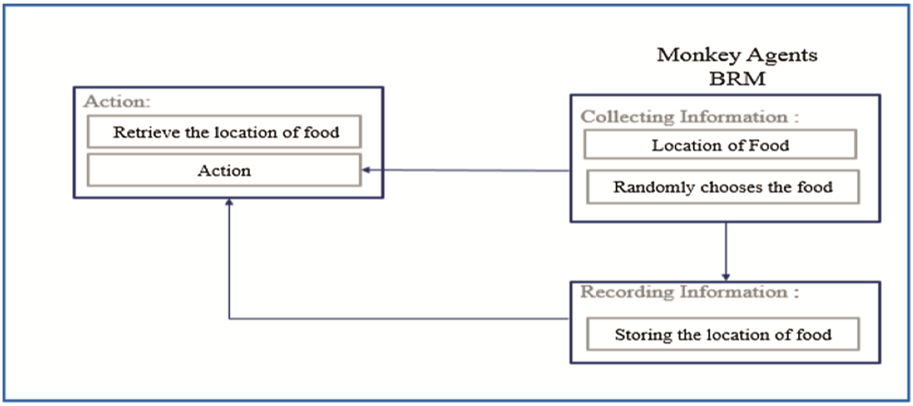


Fig. 1. Arrow and box diagram for Monkey agents' strategy

3.2 Infant Agents' Strategy

Infant agents recognize Monkey agents' beliefs regarding the location of the food. They identify all Monkey agents within their field of view that have access to the same potential food as them. Infant agents can observe up to their third neighbourhood as their field of view and only use their second and third neighbourhoods to identify the Monkey agents' perspective. Thus, Infant agents' field of view is limited to the first neighbourhood in the absence of Monkey agents. Infant agents perceive the area of Monkey agents' field of view which is shared with their own. They are able to store each Monkey agent's perspective as long as that Monkey agent exists in their field of view. Infant agents track Monkey agents' field of view and store their perspectives to create false beliefs for Monkey agents. Figures 2 and 3 illustrate two main scenarios of Infant agents' strategy regarding Monkey agents' false beliefs. The Infant agent stores this perspective. Thus, the Infant agent is able to identify the Monkey agent's false belief when another agent consumes the food registered by the Monkey agent.

In situations where there is more than one source of food available, the Infant agents' strategy is to choose the food that creates a false belief for the Monkey agents. Infant agents prioritize acquisition of the food based on two conditions. Firstly, that the location of the food has previously been stored in the Monkey agent's memory and in the Monkey agent's perspective, is still there. Secondly, that the Monkey agent has no alternative food in its field of view. This is called the priority function through which Infant agents apply belief representation abilities.

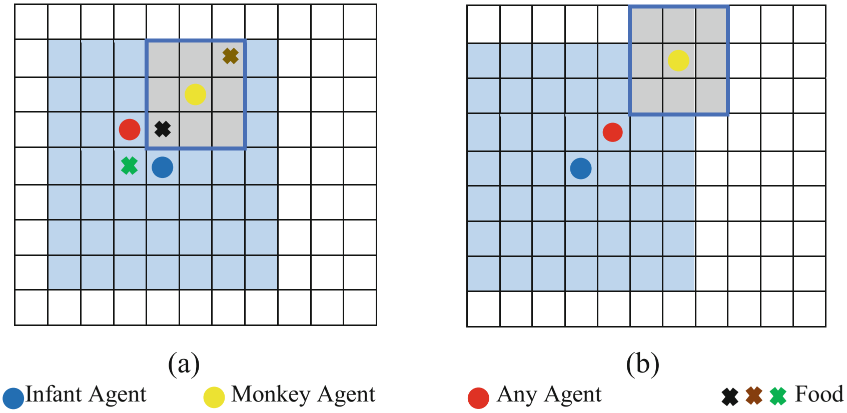


Fig. 2. Infant agents' strategy regarding Monkey agents' false beliefs. (a) The Infant agent (blue circle) encounters green and black food whereas the Monkey agent (yellow circle) encounters the black and brown food in their field of view in time step t. The red agent (red circle) is another agent that has access to the black and green food. (b) The Monkey agent, the red agent and the Infant agent respectively consume the brown, black and green food. In the Monkey agent's perspective, the black food is still in its previous location. (Color figure online)

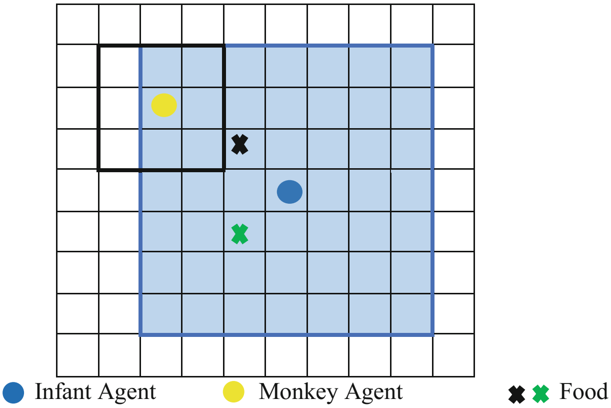


Fig. 3. Infant agents' strategy. The Infant agent (blue circle) encounters the green and black food. Note that the Infant agent has already stored the Monkey agent's perspective regarding the location of the black food. Thus, the Infant agent's priority is to move towards the black food to create a false belief for the Monkey agent (yellow circle). (Color figure online)

The Infant agents' strategy, consisting of collecting information, recording information, reasoning process and expressing (using this understanding of) others' belief-desire phases, is shown in Fig. 4 (IAD). IAD shows that Infant agents collect and reason which information to choose from their field of view, including information about the location of the food and other agents' beliefs regarding the location of the food.

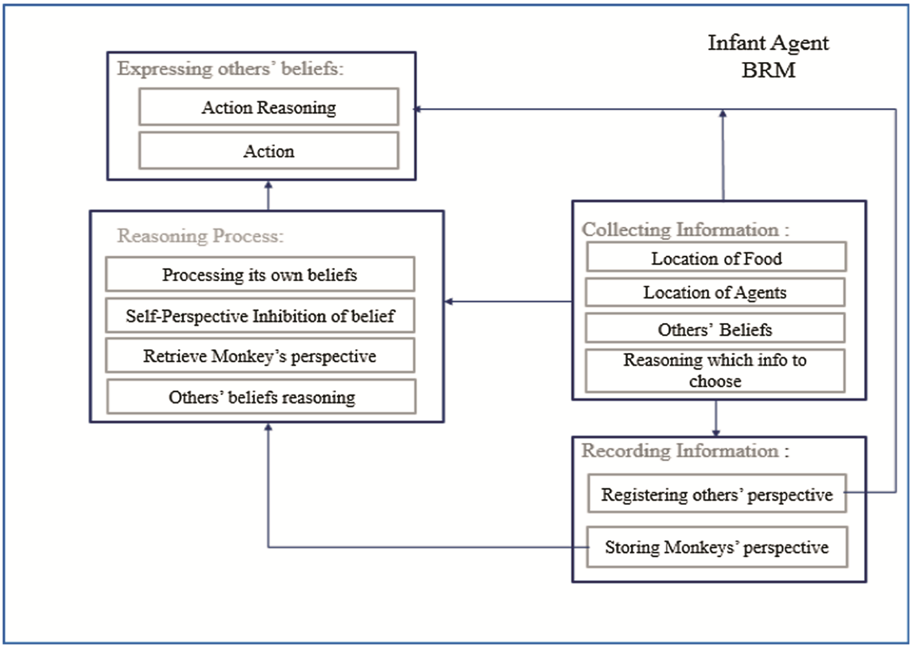

Infant Agent
BRM
Expressing others' beliefs:
Action Reasoning
Action
Reasoning Process:
Processing its own beliefs
Self-Perspective Inhibition of belief
Retrieve Monkey's perspective
Others' beliefs reasoning
Collecting Information :
Location of Food
Location of Agents
Others' Beliefs
Reasoning which info to choose
Recording Information :
Registering others' perspective
Storing Monkeys' perspective

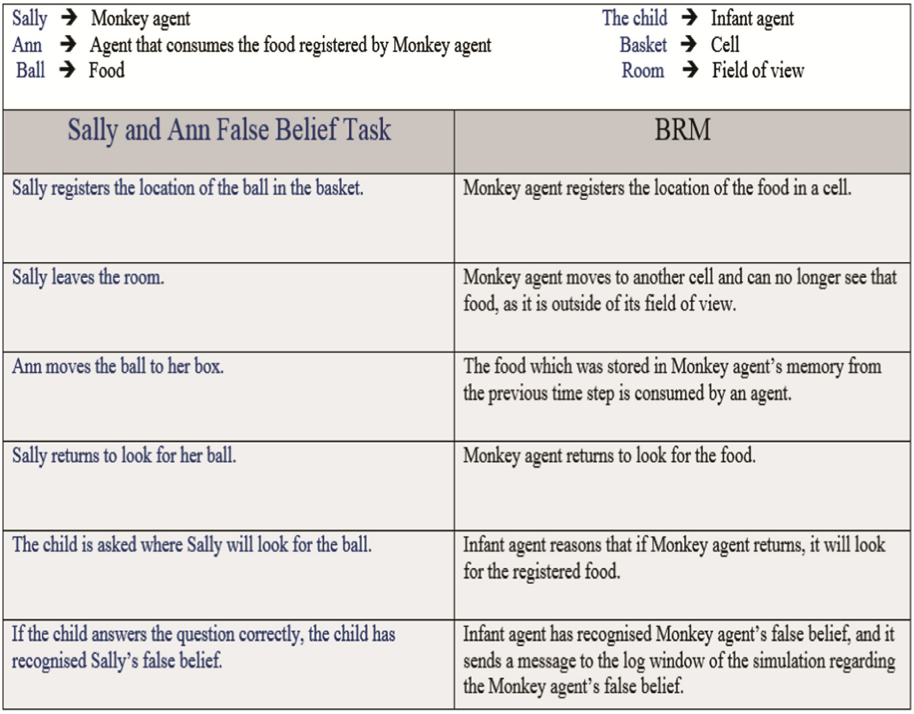
Fig. 4. Infant Agents' arrow and box Diagram (IAD)

In the recording information phase of IAD, Infant agents store the information of Monkey agents' perspective (time step t − 1) and register the current location of the food which is outside the Monkey agent's field of view (time step t). Note that in registration, the access to the information is only possible at the current time step while in recording the information, it is possible to store the information for using in future time steps. The belief representation in BRM hinges on three different perspectives: the Infant agent's perspective, the Monkey agent's belief and the Infant agent's perspective of the Monkey agent's belief. The concept of time is critical in false belief scenarios in BRM; in the previous time step (t − 1), all of these perspectives are identical. However, in the current time step (t), there is a contradiction between the Infant agent's perspective about the location of the food and the Infant agent's perspective of the Monkey agent's belief. In the reasoning process of IAD, Infant agents' beliefs about the location of the food is the same as the real information of the world. Nevertheless, they temporarily inhibit their own perspective, and retrieve the stored information about Monkey agents' perspectives which are in their field of view. This is considered as self-perspective inhibition of Infant agents. Given the Monkey agent's perspective regarding the location of the food, Infant agents reason about Monkey agents' desire towards the food. Finally, in the expressing others' beliefs phase, Infant agents reason and use the understanding of Monkey agents' belief in an action.

4 The Analogy Between the Standard False Belief Task and BRM

The Sally and Ann false belief task is considered as the standard false belief task. The fit between the Sally and Ann false belief task and BRM is measured by comparing their corresponding critical features, which are shown in Table 1.

Table 1. A comparison between the Sally and Ann false belief task and BRM



Sally	➔	Monkey agent	The child	➔	Infant agent
Ann	➔	Agent that consumes the food registered by Monkey agent	Basket	➔	Cell
Ball	➔	Food	Room	➔	Field of view

Sally and Ann False Belief Task	BRM
Sally registers the location of the ball in the basket.	Monkey agent registers the location of the food in a cell.
Sally leaves the room.	Monkey agent moves to another cell and can no longer see that food, as it is outside of its field of view.
Ann moves the ball to her box.	The food which was stored in Monkey agent's memory from the previous time step is consumed by an agent.
Sally returns to look for her ball.	Monkey agent returns to look for the food.
The child is asked where Sally will look for the ball.	Infant agent reasons that if Monkey agent returns, it will look for the registered food.
If the child answers the question correctly, the child has recognised Sally's false belief.	Infant agent has recognised Monkey agent's false belief, and it sends a message to the log window of the simulation regarding the Monkey agent's false belief.

More specifically, Infant agents are analogous to the participant child, while the Monkey agents are analogous to Sally. Any agent which consumes the food that the Monkey agent registered earlier acts as ‘Ann’ in the task. Similar to Sally’s registration regarding the location of the ball in the basket, the Monkey agent registers the location of the available food in its field of view. When Sally leaves the room, it is similar to when the Monkey agent moves, causing the food to no longer be in its field of view; both unintentionally create environments which have the potential for a false belief scenario. Moreover, similar to the child who is capable of passing the Sally and Ann false belief task, an Infant agent is able to recognize the Monkey agent’s perspective and predict its desire to move towards a registered food in its memory.

There is a perspective difference between the child and Sally regarding the location of the ball. Accordingly, the perspective differences between Infant and Monkey agents are related to the location of food. The real location of the food is not the same as that in the Monkey agent’s perspective because the Monkey agent is unable to update its belief about the current location of the food. In contrast, the Infant agent has access to the real information as well as the Monkey agent's perspective, both of which provide key information for the false belief scenarios. Hence, the Infant agents' belief attribution, similar to the child's belief attribution in the standard false belief task, is represented by BRM. As the interactions between agents create a number of false belief scenarios within the same time step, these belief attributions occur simultaneously for a number of Infant agents in the environment. The impact of belief attribution on social performance naturally emerges within the dynamics of BRM's virtual society. Moreover, a variety of true and false belief scenarios develop through the simulation, which is far beyond the scope of the isolated Sally and Ann false belief task.

5 The BRM Results

The setup of simulations consists of Infant agents and Monkey agents with two parameters including the number of food sources and the number of agents. The agents' average performance is calculated by running the simulation four times for each combination of the parameter values. The BRM sensitivity analysis to its initial conditions are examined by altering the parameter values in the setup. Table 2 shows the chosen parameter values. The reason for choosing these is that they correspond to a high number of Monkey agents' false belief scenarios, which is critical for evaluating the agents' performance in the context of false beliefs.

Table 2. Parameter values

Parameters	Values
Number of food (Food)	500, 600,700,800
Number of agents	400, 500, 600,700

5.1 Agents' Performances

One of the objectives of BRM is to analyse the effects of parameters on agents' performances. The number of food sources consumed by agents is used as a measurement to evaluate the agents' performances. Therefore, by comparing the performances of agents, it would be possible to identify some patterns between their performances and the concepts behind their abilities.

The difference between Infant agents and Monkey agents' performances, as illustrated in Fig. 5, demonstrates that Infant agents perform significantly better than Monkey agents. The most salient difference in performance occurs when the number of each type of agent is equal to 700; nevertheless, it is primarily subject to the number of food sources. For example, when the number of food sources is 800, the differences decrease. The main reason is that the high availability of food enables the agents to consume food without the occurrence of false belief scenarios. The prominent pattern here is that differences in performance increase as the number of agents increases. However, the number of food has a great impact on this pattern.

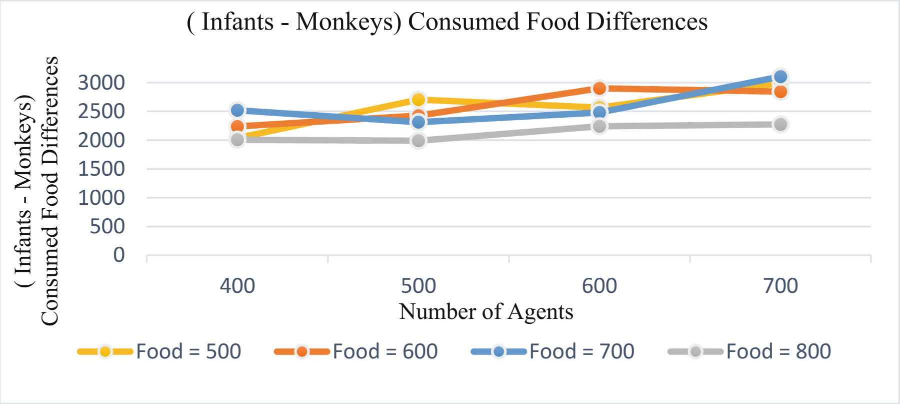


Fig. 5. The performance differences of Infant agents and Monkey agents

5.2 The Number of False Beliefs of Monkey Agents

Infant agents send a message to the log window of the simulation as an output when there is a false belief scenario regarding any specific Monkey agent. At the end of each simulation run, the total number of Monkey agents' false beliefs which occurred in that run is displayed on the basis that only one false belief is recorded for each Monkey agent in each time step. The number of Monkey agents' false beliefs in the simulation starts to increase as the number of agents in the simulation setup increases (see Fig. 6). The reason is that more Monkey agents are able to register the food and pursue it in the next time step. The number of Monkey agents' false beliefs is negatively correlated with the number of food sources but it is positively correlated with the number of agents. These results are consistent with agents' performance results.

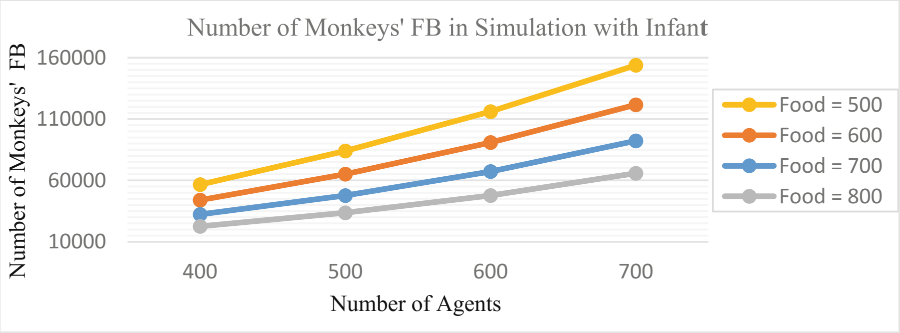


Fig. 6. The number of False Beliefs (FB) experienced by Monkey agents in the simulation with Infant agents, Food = 500, 600, 700, 800.

6 Discussion

The Belief Representation Model (BRM) consists of two types of agents: Infant and Monkey agents. Infant agents are capable of reasoning about Monkey agents' desires and beliefs, and recognize Monkey agents' false beliefs; they track Monkey agents' field of view, and register and store Monkey agents' perspectives regarding the location of food. They are also able to inhibit their own perspective regarding the location of the food and consider Monkey agents' perspectives. Thus, Infant agents are able to understand Monkey agents' beliefs about the location of the food. In contrast, Monkey agents remember and track the food from the previous time step but they lack the ability to take into account the perspective of others. Agents' decisions are influenced by their capability to consider the perspective of others and the information they perceive from their neighbourhoods. Infant agents utilize their belief representation ability and store Monkey agents' perspectives to choose food which creates false belief scenarios for Monkey agents.

6.1 Belief Representation Systematic Approach (BRSA)

One methodological approach in agent-based models is to present diagrams that illustrate the control flow and the underlying logic of the complicated and interconnected

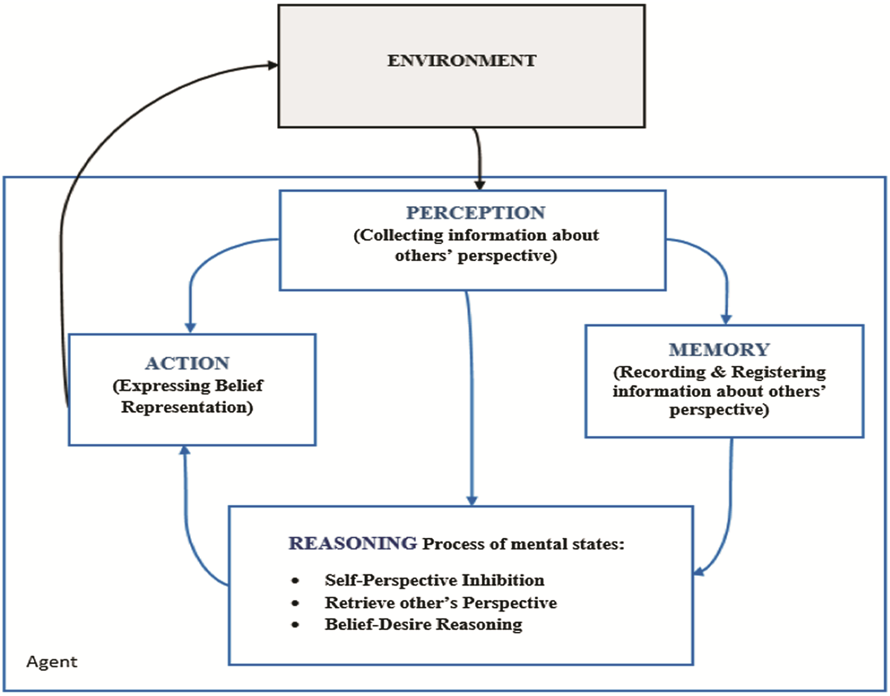

ENVIRONMENT
PERCEPTION
(Collecting information about others' perspective)
ACTION
(Expressing Belief Representation)
MEMORY
(Recording & Registering information about others' perspective)
REASONING Process of mental states:
• Self-Perspective Inhibition
• Retrieve other's Perspective
• Belief-Desire Reasoning
Agent

Fig. 7. Belief Representation Systematic Approach (BRSA)

procedures of agents' actions. Figure 4, IAD, illustrates the basic underlying phases that occur for an agent with belief representation competence. The concept behind the phases in IAD provides a structured and coherent approach to belief representation processes. The collective phases are derived from the examining the behaviour and the dynamic processes of decision trees of Infant agents and is called the Belief Representation Systematic Approach (BRSA). BRSA classifies the belief representation procedure into four basic underlying phases: Collecting information, Recording information, Reasoning process of beliefs and desires and finally, Expressing mental states of others by an action. The diagram of BRSA is shown in Fig. 7.

Phase 1 of BRSA: Collecting Information. Infant agents collect information from their field of view in each time step including information about the location of food, the location of other agents and particularly information relating to the Monkey agents' perspective of the location of the food. Infant agents reason about the information they need to collect. They are interested in the perspectives of Monkey agents which have access to the same food as them. The collecting information phase is a central online phase which feeds other phases in BRM. Similar to the Sally and Ann false belief task question, Infant agents must answer the following questions correctly to pass the false belief test:

- Where was the location of food registered in the Monkey agent's memory (which cell)?
- Is the food still in the Monkey agent's field of view?
- Can the Infant agent consume the food which was stored in Monkey agents' memory? (Has the food been eaten by other agents?)
- Where will the Monkey agent search for the food when it returns to its previous position?

The answers to these questions will be collected through the collecting information phase in different time steps. There is a dynamic link between the collecting information phase and the other phases in regard to each false belief scenario. The collecting information phase is parallel to the time and dynamics of the world; this means that the collecting information phase is a continuous process corresponding to the time steps and environmental changes. The online raw information becomes available from the collecting information phase, which can then feed other phases, enabling them to complete their related processing simultaneously.

Phase 2 of BRSA: Recording Information. BRSA demonstrates that memory plays a crucial role in belief representation. Memory is indispensable for Infant agents in order to pass the false belief test. Infant agents record information about the location of food from two different perspectives: Monkey agents' perspectives and their own. Infant agents are able to switch from one perspective to another.

Phase 3 of BRSA: Reasoning Process of Beliefs and Desires. The reasoning process of the beliefs and desires phase involves complex information processing. This phase demonstrates the capability of the Infant agents to understand the false beliefs of others. Conceivably, there are two different versions of beliefs about the location of the food in false belief scenarios; the Infant agents' own perspective, which is the last updated version of the reality, and the Monkey agents' perspective which is not updated from the previous time step. By default, agents use the updated information about the location of food due to the dynamics of the environment. However, Infant agents inhibit their own beliefs about the current location of the food temporarily and restore the Monkey agents' beliefs, which have already been stored. Infant agents take into consideration that other agents have a common desire towards the food. The procedure of the reasoning process of beliefs and desires includes:

- Self-perspective inhibition
- Retrieving the protagonist's perspective data from memory
- Selective processing of protagonist's (Monkey agent's) belief and desire based on its own belief and desire.

At this stage of the BRM, the Infant agent's recognition of the Monkey agent's belief is complete.

Phase 4 of BRSA: Expressing Beliefs and Desires of others. There is a critical difference between having a competence and using it. This phase represents the agents' actions based on their understanding of other agents' beliefs and desires. Noticeably, expressing (using their understanding of) the beliefs and desires of others is analogous with the measurement test in false belief tasks. Infant agents utilize their understanding of Monkey agents' false beliefs in their actions. First, when the Monkey agent's belief is true, the Infant agent prioritizes consuming the food which creates a false belief for Monkey agents. Second, once the Monkey agent's belief is false, then Infant agents express their understanding of the Monkey agent's false belief by sending a message and at the end of the simulation, a message shows the total number of Monkey agents' false beliefs recognized by Infant agents.

6.2 In Which Conditions Is False Belief Task a Decisive Test for ToM Based on BRSA?

BRSA demonstrates the necessary resources for each phase of the Infant agent's understanding of others' beliefs. Firstly, the Reasoning phase demonstrates the demands on cognitive resources in the false belief task include reasoning, inhibition, recording and retrieving information about others' perspective from memory, as well as the required interconnection between the resources. Therefore, having these resources is a precondition for success in false belief tasks. Secondly, the Expressing phase demands more than understanding others' beliefs. These two reasons validate the point that false belief tasks require sufficient cognitive resources as a precondition to act as a decisive test for theory of mind. This is compatible with the literature which suggests that false belief tasks involve challenging actions, engaging with complex reasoning, intellectual connections and skills such as linguistic abilities, which might be more demanding than understanding others' beliefs.

6.3 The Effects of Belief Representation on Agents' Performances

The simulation results show that Infant agents perform consistently better than Monkey agents. Infant agents' efficiency is due to three factors. First, Infant agents' recognition of Monkey agents' false beliefs. Second, their ability to apply this understanding and information into a plan that enhances their chances of achieving their goals. Third, performing an action by employing the plan (creating more false belief scenarios for Monkey agents) increases the successful performance of Infant agents. Therefore, belief representation ability is a fundamental element in higher performance, along with other factors of reasoning, planning and contributing an action in achieving a goal.

6.4 The Network of Resources in BRSA

BRSA presents the key components of the belief representation processes, which consist of perception, memory, inhibitory control and selective process reasoning, in addition to complex reasoning resources, which are essential for the phase of expressing others' beliefs and desires. Together, these components represent a network of resources that shapes the individual's ability to understand others' beliefs. This network is compatible with the developmental literature underpinning the theory of mind network (Mohnke et al. 2016; Gallagher and Frith 2003; Carrington and Bailey 2009).

7 Conclusion

BRM is an original model which illustrates the underlying processes of understanding others' false beliefs in a structured and coherent approach by classifying this procedure into four phases called the Belief Representation Systematic Approach (BRSA). The first phase involves agents collecting information, particularly in relation to other agents' perspectives on the location of the food. The second phase, the recording information phase, is when agents store the collected information, including the perspectives of other agents, in their memory. This phase highlights the role of memory and time traveling in belief representation. The third phase, the reasoning process of beliefs and desires phase, is the main phase for processing information about others' perspectives. In this phase, agents inhibit their own beliefs temporarily and restore others' beliefs. This phase also involves critical reasoning about the beliefs and desires of others. The fourth phase, expressing others' beliefs and desires phase, is concerned with deciding on an action by considering others' beliefs and desires. This phase identifies the subtle difference between having the ability to understand the beliefs of others and using this understanding in agents' actions. BRSA validates that false belief tasks require sufficient resources as a precondition to act as a decisive test for theory of mind. Furthermore, BRSA identifies the key components of belief representation processes consisting of perception, memory, inhibitory control and selective process reasoning. These components represent a network of indispensable resources for belief representation. In addition, the performance of agents capable of belief representation is consistently higher in achieving their goals. The main factors in producing a more efficient performance in agents include a combination of understanding others' beliefs and implementing this understanding by taking action.

References

Apperly, I.: Mindreaders: The Cognitive Basis of "Theory of Mind". Psychology Press, Hove (2011)

Baillargeon, R., Scott, R.M., He, Z.: False-belief understanding in infants. Trends Cogn. Sci. **14**, 110–118 (2010)

Baron-Cohen, S., Jollife, T.M., Robertson, M.: Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome. J. Child Psychol. Psychiatry **38**, 812–822 (1997)

Bersini, H.: UML for ABM. J. Artif. Soc. Soc. Simul. **15**(1), 9 (2012)

Bloom, P., German, T.: Two reasons to abandon the false belief task as a test of theory of mind. Cognition **77**, B25–B31 (2000)

Call, J., Tomasello, M.: Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn. Sci. **12**(5), 187–192 (2008)

Carrington, S.J., Bailey, A.J.: Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum. Brain Mapp. **30**(8), 2313–2335 (2009)

Doherty, M.J.: Theory of Mind: How Children Understand Others' Thoughts and Feelings. Psychology Press, Hove (2009)

Epstein, J.M., Axtell, R.L.: Growing Artificial Societies: Social Science from the Bottom Up. The MIT Press, Cambridge (1996)

Frith, C.D.: The role of metacognition in human social interactions. Philos. Trans. R. Soc. B **367**, 2213–2223 (2012)

Gallagher, H.L., Frith, C.: Functional imaging of 'theory of mind'. Trends Cogn. Sci. **7**(2), 77–83 (2003)

Kovács, Á.M., Téglás, E., Endress, A.D.: The social sense: susceptibility to others' beliefs in human infants and adults. Science **330**(6012), 1830–1834 (2010)

Martin, A., Santos, L.R.: The origins of belief representation: monkeys fail to automatically represent others' beliefs. Cognition **130**, 300–308 (2014)

Martin, A., Santos, L.R.: What cognitive representations support primate theory of mind? Trends Cogn. Sci. **20**(5), 375–382 (2016)

Mohnke, S., Erk, S., Schnell, K., Romanczuk-Seiferth, N., Schmierer, P., Romund, L., Walter, H.: Theory of mind network activity is altered in subjects with familial liability for schizophrenia. Soc. Cogn. Affect. Neurosci. **11**, 299–307 (2016)

Onishi, K.H., Baillargeon, R.: Do 15-month-old infants understand false beliefs? Science **308**, 255–258 (2005)

Rubio-Fernández, P.: How to pass the false-belief task before your fourth birthday. Psychol. Sci. **24**(1), 27–33 (2013)

Southgate, V., Vernetti, A.: Belief-based action prediction in preverbal infants. Cognition **130**, 1–10 (2014)

Southgate, V., Senju, A., Csibra, G.: Action anticipation through attribution of false belief by 2-year-olds. Psychol. Sci. **18**, 587–592 (2007)

Surian, L., Caldi, S., Sperber, D.: Attribution of beliefs by 13-month-old infants. Psychol. Sci. **18**, 580–586 (2007)

Repast: The Repast Suite (2017). https://repast.github.io/. Accessed 05 Apr 2018

Wellman, H.: Making Minds. How Theory of Mind Develops. Oxford University Press, New York (2014)

Wellman, H.M., Bartsch, K.: Young children's reasoning about beliefs. Cognition **30**, 239–277 (1988)

Wellman, H.M., Cross, D., Watson, J.: Meta-analysis of theory-of-mind development: the truth about false belief. Child Dev. **72**, 655–684 (2001)

Workman, L., Reader, W.: Evolutionary Psychology: An Introduction. Cambridge University Press, New York (2014)

Information, Systems, and Network Science

Similar but Different: Exploiting Users' Congruity for Recommendation Systems

Ghazaleh Beigi(✉) and Huan Liu

Arizona State University, Tempe, AZ, USA
{gbeigi,huan.liu}@asu.edu

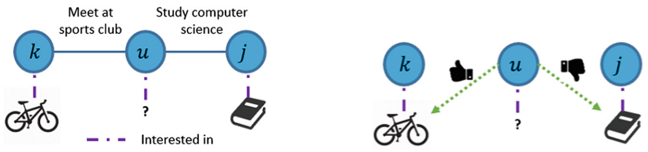
Abstract. The pervasive use of social media provides massive data about individuals' online social activities and their social relations. The building block of most existing recommendation systems is the similarity between users with social relations, i.e., friends. While friendship ensures some homophily, the similarity of a user with her friends can vary as the number of friends increases. Research from sociology suggests that friends are more similar than strangers, but friends can have different interests. Exogenous information such as comments and ratings may help discern different degrees of agreement (i.e., congruity) among similar users. In this paper, we investigate if users' congruity can be incorporated into recommendation systems to improve it's performance. Experimental results demonstrate the effectiveness of embedding congruity related information into recommendation systems.

1 Introduction

Recommender systems play an important role in helping users find relevant and reliable information that is of potential interest [12]. The increasing popularity of social media allows users to participate in online activities such as expressing opinions and emotions [4] (via commenting or rating), establishing social relations [3,5,6] and communities [1]. Extracting these additional information (e.g., social relations) from social networks in favor of the task of recommendation, has attracted increasing attentions lately [10,16]. In particular, homophily [17] which states that friends are more likely to share similar preferences with each other than strangers, is the backbone paradigms of recommendation systems that exploit social relations.

Despite the close-knit interests between friends, their friendship shall not always be treated as if they are completely alike. Naturally, as the number of friends of a user grows, it is inevitable that her friends' preferences diverge [22,23]. For example, a user's friend circles are constituted of different people with various backgrounds and interests, ranging from her family to her schoolmates or co-workers. Research findings from sociology suggest that friends can make different decisions and many a time, these decisions can be very different from each other [7]. Furthermore, although individuals have the tendency to become similar within a friendship, this should be considered as an effect of

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 129–140, 2018.
https://doi.org/10.1007/978-3-319-93372-6_15



(a) u**'s local network.**

(b) u**'s opinion about others' interest.**

Fig. 1. The role of opinion agreement in inferring user u's interests.

friendship, not a constitutive of it [7]. Sociologists have also shown that the level of similarity among online users is much lower than that of actual friends in the real world [2]. Thus, considering similarity at friendship level alone could be too coarse for recommendation tasks and might degrade the performance.

To give a palpable understanding of the above scenario, let us take a look at a toy example in Fig. 1. Assume user u is connected to user j since they both study computer science, and is connected to user k because they usually meet at the same sports club (see Fig. 1(a)). Our goal is then to infer the interests of user u, given the information about interests of users j and k. Social relations alone as representative of shared preferences would suggest that user u is interested in both Machine Learning books and biking. However, from the opinions that user u has expressed towards others' interests (see Fig. 1(b)), we can infer that she seems to be only interested in what user k is interested in. Therefore, exogenous information such as the opinions of users regarding each other's interests can help inferring varying interests between them.

In this study, we use the term *congruity* defined as a *degree of agreement* between people [21], to refer to such a *degree of match* among users' opinions. In other words, according to the sociology, congruity shall be treated as a perceptual concept that captures consensus between people [8,20]. Augmenting the recommendation systems with congruity might help precise inferring of the users' preferences. None of the existing recommendation systems until now have taken this into account. The merit of exploiting the congruity obtained from users' interactions is that we can capture their preferences more accurately with potentials in improving the performance of recommendars, while it also poses new challenges. First, users' congruity information is not always readily available and extra effort is required to extract users' opinions towards each other's interests– this is in contrast to the ideal scenario in the example shown in Fig. 1. Second, it's challenging to wisely incorporate congruity in recommendation systems.

The abundant information about users' behaviors and interactions is a rich source of users' congruity as most social media websites allow for free interaction and exposing viewpoints between users. This information has also potentials in distinguishing between congruity and social relations. In this paper, we seek to answer the following research questions: (1) What is the relationship between users' social relations (or friendship) and congruity? how different are social relations and congruity? (2) Why is it sensible to integrate congruity in recommendation systems? and (3) How can we mathematically obtain users' congruity from social data? We then propose a novel framework based on congruity for recommendation systems (CR).

2 Data Analysis

We use two large online product-review websites, Epinions and Ciao, where users can establish friendship links toward each other from which we can construct the user-user social relations matrix $\mathbf{G}$. We denote by $\mathbf{G}_{ij} = 1$, if u_i and u_j are friends, and $\mathbf{G}_{ij} = 0$ otherwise. Different products are given ratings of 1 to 5 by users. From these ratings, we build our user-item rating matrix $\mathbf{R}$ where $\mathbf{R}_{ij}$ is the rating score that user u_i has given to the item v_j. Users are also allowed to write reviews and can express their opinions toward each other by rating how helpful their reviews were from 1 to 5. Some key statistics are shown in Table 1. We perform some standard preprocessing by filtering out items and users with less than 3 ratings and users without social relations.

Table 1. Statistics of the preprocessed data.

Name	Epinions	Ciao
Users	22,264	6,852
Items	35,040	16,202
Ratings	577,692	159,615
Friendships	292,345	111,672
Pairs of users with congruity	621,327	575,414

2.1 Congruity and Social Relations Difference

Recall that congruity is defined as a degree of agreement between people, which captures the socially defined levels of consensus between them [8,20], and can be gleaned from users' interaction data. To illustrate this, let us glance at Fig. 2 which demonstrates the typical users' interactions on websites such as Epinions and Ciao. Note, this is an extension to our toy example in the previous section, in that we have added another user u' which is **not** connected to the existing users. In this example, users u and u' could rate the helpfulness of reviews written by users k and j on the bike and Machine Learning book. The high helpfulness rating that user u has given to k's review demonstrates the high level of opinion agreement and congruity between them, while the low helpfulness rating given to j's review, by user u, implies lower congruity between them. Likewise, user u' has similar congruity levels with users k and j, however, we cannot infer the congruity level between u and u', given this information.

Accordingly, we construct the user-user congruity matrix from users positive and negative interaction matrices, $\mathbf{P}$ and $\mathbf{N}$, which are obtained from helpfulness

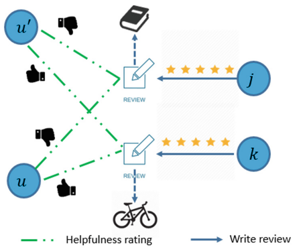


Fig. 2. An illustration of users' interaction in product review sites.

ratings as follows. First we consider high helpfulness ratings $\{4, 5\}$ as positive user interaction, low helpfulness ratings $\{1, 2\}$ as negative interaction and rating $\{3\}$ as neutral. Then, for each pair of users $\langle u_i, u_j\rangle$, we count the number of positive and negative interactions, p_{ij} and n_{ij}, between u_i and u_j. We calculate the positive interaction strength $\mathbf{P}_{ij}$ as a function of p_{ij}, i.e., $\mathbf{P}_{ij} = g(p_{ij})$ where $\mathbf{P}_{ij} \in [0, 1]$. Therefore, we need the function $g(x)$ to have the following properties: (1) $g(0) = 0$, (2) $\lim_{x\to\infty} g(x) = 0$, and (3) be an increasing function of x. One choice could be $g(x) = 1 - \frac{1}{\log(x+1)}$ for $x \neq 0$ and $g(x) = 0$ otherwise.

Likewise, we construct the user-user negative interaction matrix $\mathbf{N}$. Ultimately, we create the user-user congruity matrix $\mathbf{C}$ by utilizing user-user positive and negative interaction matrices. Positive interactions imply more congruity between users while negative interactions imply the opposite. Matrix $\mathbf{C}$ is then built from the linear combination of $\mathbf{P}$ and $\mathbf{N}$ as $\mathbf{C} = \mathbf{P} - \mathbf{N}$. Note that there might be other ways to construct $\mathbf{C}$, $\mathbf{P}$ and $\mathbf{N}$, which we leave to future work.

Next, we further dig into our preprocessed data. As we see from Table 2, there are four possible types of pairs of users in our data: (1) users who are friends with each other and are congruent, (2) users who are friends with each other but are incongruent ($\langle i, j\rangle$ are incongruent if $\mathbf{C}_{ij} \leq 0$), (3) users who are strangers but are congruent, and (4) strangers who are also incongruent. These statistics suggest that, not all friends are always congruent. In particular, 24% of friends in Ciao and 43% of friends in Epinions are not congruent at all. Another interesting observation is that, 85% and 73% of pairs of congruent users in Ciao and Epinions are not friend with each other. Consequently, there might be some users with a degree of match in their preferences, who are not necessarily within their friend circles of each other. On the other hand, the number of congruent users in both datasets are much more than that of friends, which results in significantly different sets of users. These all, ultimately motivate us to exploit congruity for recommendation tasks rather than merely using social relation.

2.2 Analysis of Users' Congruity

Before leveraging users' congruity for recommendation tasks, we would like to conduct a sanity check to see if this concept is applicable to social media data. We first study if social relations between users correspond to their congruency or in other words, if all friends are congruent with each other or not. Then, we

Table 2. Number of pairs of users with different properties.

(a) Ciao

	Congruent	**Incongruent**
Friends	84,063	27,609
Strangers	491,351	$\sim$ 46 M

(b) Epinions

	Congruent	**Incongruent**
Friends	163,985	128,360
Strangers	457,342	$\sim$ 494 M

investigate the correlation between users' congruity and preferences. Specifically, we verify two questions: (1) Are all friends congruent?, and (2) Does congruity among users imply a higher chance of sharing similar preferences between them?

To answer the first question, for each user u_i, we consider all of her friends. Then, we compute the minimum c^i_{min} and maximum c^i_{max} values of congruity between user u_i and her friends. Two vectors $\mathbf{c}^{min}$ and $\mathbf{c}^{max}$ are obtained by computing c^{min}s and c^{max}s for all users. We conduct a two-sample t-test on $\{\mathbf{c}_{min}, \mathbf{c}_{max}\}$ where the null hypothesis H_0 is that friends are all congruent, i.e. there is no significant difference between minimum and maximum value of users' congruity. The H_1 is also that friends are not all congruent:

$$H_0 : \mathbf{c}_{min} = \mathbf{c}_{max}, \qquad H_1 : \mathbf{c}_{min} \neq \mathbf{c}_{max}. \tag{1}$$

The null hypothesis is rejected at significance level $\alpha = 0.01$ with p-value shown in Table 3. This suggests a negative answer to the first question.

A similar procedure we did for the first question can be followed to answer the second question. Consider the pair of user $\langle u_i, u_j \rangle$ with positive value of congruity ($\mathbf{C}_{ij} > 0$). We randomly select user u_k who is incongruent with u_i. Users similarities cp^{ij} and cr^{ik} have been calculated for $\langle u_i, u_j \rangle$ and $\langle u_i, u_k \rangle$, respectively. Finally, two vectors $\mathbf{c}_p$ and $\mathbf{c}_r$ are obtained where $\mathbf{c}_p$ is the set of all cp's for pairs of users with congruity; while $\mathbf{c}_r$ is the set of cr's for pairs of users without congruity. We use cosine similarity over the item-rating entries to find the similarities between users. We conduct a two-sample t-test on $\{\mathbf{c}_p, \mathbf{c}_r\}$ where the null hypothesis H_0 is that users without congruity are more likely to share similar preferences:

$$H_0 : \mathbf{c}_p \leq \mathbf{c}_r, \qquad H_1 : \mathbf{c}_p > \mathbf{c}_r. \tag{2}$$

The null hypothesis is rejected at significance level $\alpha = 0.01$. Thus, users with congruity are more likely to share preferences than those without.

The corresponding p-values for the above t-tests are summarized in Table 3 for both datasets. The results from these analyses: (1) demonstrate that although friends share similar interests, but friendship relations are not good measure of congruency as friends are not always congruent with each other and thus considering similarity at friendship level alone degrade the performance of recommendation tasks, and (2) confirm the importance of deploying a measure of users' congruity in computing users' similarity other than social relations.

Table 3. p-values of t-test results corresponding to analysis tests.

	Ciao: $\{\mathbf{c}_{min}, \mathbf{c}_{max}\}$	Ciao: $\{\mathbf{c}_p, \mathbf{c}_r\}$	Epinions: $\{\mathbf{c}_{min}, \mathbf{c}_{max}\}$	Epinions: $\{\mathbf{c}_p, \mathbf{c}_r\}$
p-value	3.09e–6	1.72e–5	6.17e–5	4.81e–4

3 Congruity-Based Recommendation

We begin this section by introducing matrix factorization based collaborative filtering technique which we chose as basis of CR. Matrix factorization based techniques have been widely used for building recommender systems [13,16] and the basic assumption is that a small number of factors influence user rating behavior and maps both user and item to a joint latent factor space with dimensionality d. Assume that $\mathbf{U}_i \in \mathbb{R}^{1\times d}$ and $\mathbf{V}_j \in \mathbb{R}^{1\times d}$ are the user preference vector for u_i and item characteristic vector for v_j, respectively. The rating score given by u_i to v_j is modeled as $\mathbf{R}_{ij} = \mathbf{U}_i\mathbf{V}_j^\top$. Matrix factorization seeks to find $\mathbf{U} = [\mathbf{U}_1, ..., \mathbf{U}_n]$ and $\mathbf{V} = [\mathbf{V}_1, ..., \mathbf{V}_m]$ by solving the following problem:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i\mathbf{V}_j^\top)^2 + \lambda(\|\mathbf{U}\|_{\mathbf{F}}^2 + \|\mathbf{V}\|_{\mathbf{F}}^2) \tag{3}$$

where $\lambda(\|\mathbf{U}\|_{\mathbf{F}}^2 + \|\mathbf{V}\|_{\mathbf{F}}^2)$ is added to avoid over-fitting and $\mathbf{I}_{ij}$ controls the contribution from $\mathbf{R}_{ij}$. A typical choice of $\mathbf{I}$ is $\mathbf{I}_{ij} = 1$ if $\mathbf{R}_{ij} \neq 0$ and $\mathbf{I}_{ij} = 0$, otherwise. The observations in the previous section demonstrated that although friends are similar, they are not all congruent and we need to integrate congruity to measure the similarity of shared preferences between users. Moreover, users with congruity are more likely to share similar preferences compared to those without either of them. These findings provide the groundwork for us to model users' congruity to measure the users' preferences closeness.

Let $\mathbf{L} \in \mathbb{R}^{n\times n}$ be the preference closeness matrix where $\mathbf{L}_{ik}$ denotes the preference closeness strength between u_i and u_k. The motivation behind preference closeness strength is that users are more/less likely to share similar preferences when they establish higher/lower level of congruity with each other. Following this idea, $\mathbf{L}_{ik}$ could be calculated as the congruity $\mathbf{C}_{ik}$ between them. The closeness of u_i and u_k user preference vectors is then controlled by their preference closeness strength,

$$\min \sum_{i=1}^{n}\sum_{k\in\mathcal{T}_i} \mathbf{L}_{ik}\|\mathbf{U}_i - \mathbf{U}_k\|_2^2 \tag{4}$$

where $\mathcal{T}_i = \{u_k|\mathbf{C}(\mathbf{i},\mathbf{k}) \neq 0\}$. In Eq. 4, a larger value of $\mathbf{L}_{ik}$ indicates the strong association between u_i and u_k; hence u_i's preference vector $\mathbf{U}_i$ is more likely to be close to u_k's preference vector $\mathbf{U}_k$– this makes the distance between $\mathbf{U}_i$ and $\mathbf{U}_k$, smaller. While a smaller value of $\mathbf{L}_{ik}$ indicates weak association between $\mathbf{U}_i$ and $\mathbf{U}_k$; therefore their distance is larger. Having introduced our solutions to model users' congruity, our framework, congruity based recommendation system (CR), is to minimize the following problem,

$$\mathcal{J} = \min_{\mathbf{U},\mathbf{V}} \sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i\mathbf{V}_j^{\top})^2 + \gamma \sum_{i=1}^{n}\sum_{k\in\mathcal{T}_i} \mathbf{L}_{ik}\|\mathbf{U}_i - \mathbf{U}_k\|_2^2 + \lambda(\|\mathbf{U}\|_{\mathbf{F}}^2 + \|\mathbf{V}\|_{\mathbf{F}}^2) \quad (5)$$

where γ is used to control contributions of the users' preferences closeness strength. We use gradient descent method to solve Eq. 5, which has been proven to gain an efficient solution in practice. The partial derivations of $\mathcal{J}$ with respect to $\mathbf{U}_i$ and $\mathbf{V}_j$ are as follows,

$$\frac{1}{2}\frac{\partial \mathcal{J}}{\partial \mathbf{U}_i} = -\sum_{j} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i\mathbf{V}_j^{\top})\mathbf{V}_j + \lambda\mathbf{U}_i + \gamma\sum_{k\in\mathcal{T}_i}\mathbf{L}_{ik}(\mathbf{U}_i - \mathbf{U}_k) \quad (6)$$

$$\frac{1}{2}\frac{\partial \mathcal{J}}{\partial \mathbf{V}_j} = -\sum_{i} \mathbf{I}_{ij}(\mathbf{R}_{ij} - \mathbf{U}_i\mathbf{V}_j^{\top})\mathbf{U}_i + \lambda\mathbf{V}_j \quad (7)$$

We use Eqs. 6 and 7 to update $\mathbf{U}$ and $\mathbf{V}$ until convergence. After learning the user preference matrix $\mathbf{U}$ and the item characteristic matrix $\mathbf{V}$, an unknown score $\hat{\mathbf{R}}_{i'j'}$ from the user $u_{i'}$ to the item $v_{j'}$ will be predicted as $\hat{\mathbf{R}}_{i'j'} = u_{i'}^{\top}v_{j'}$.

4 Experiments

In this section, we conduct experiments to answer the following two questions: (1) Does leveraging users' congruity help recommendation?, and (2) How does integration of users' congruity with social relations improve recommendation performance? and which one of the congruity and social relations contribute most to the performance improvement?

We use Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to evaluate the performance and smaller values indicate the better performance. Note that *small improvement in RMSE or MAE terms could result in a significant impact on the quality of top few recommendations* [11]. In this work, we randomly select $x\%$ of the ratings as training and treat the remaining $(100-x)\%$ as test ratings to be predicted. We vary x as $\{40, 50, 70, 90\}$.

4.1 Performance Comparison

To answer the first question, we compare the proposed framework CR with the following recommender systems:

- **MF**: It performs basic matrix factorization on the user-item rating matrix to predict the new ratings by only utilizing the rating information [19]
- **SMF**: Similarity based matrix factorization method is a variation of our method which uses user-user similarity matrix $\mathbf{S}$ for calculating $\mathbf{L}$. We use cosine similarity over the item-rating entries to find the users similarities.

- **SoReg**: It performs matrix factorization while exploiting social regularization defined based on both user-item matrix and positive social relations [16].
- **DualRec**: It integrates both review and rater roles of each user and uses item and review helpfulness ratings to learn reviewer and rater roles, respectively.

It's notable to say that other social relation based recommenders such as [14,15] have comparable results with [16]. Note that *CR* incorporates users' congruity while all three baselines *SoReg*, *DualRec* and *SMF* methods use social relations, helpfulness ratings and user-user rating similarity, respectively. This results in a substantially different method in terms of both key ideas and techniques. For all baselines with parameters, we use cross-validation to determine their values. For the proposed framework we set the parameters for Epinions and Ciao $\{\lambda = 0.01, \gamma = 100, d = 15\}$, and $\{\lambda = 0.01, \gamma = 10, d = 20\}$, respectively. Since the test set is selected randomly, the final results are reported by taking the average of 20 runs for each method. We also conduct a t-test on all comparisons, and the results are significant. The comparisons on Epinions are shown in Table 4. We also conduct experiments on Ciao and observe very similar trends. Due to lack of space, we leave the results out. We have the following observations,

- All methods outperform *MF*, suggesting the importance of leveraging exogenous information (e.g. users' congruity, social relations and rater roles) for improving the performance of recommendation systems.
- *SMF* fails to demonstrate comparable results compared to all other methods. This indicates that users' rating similarity cannot capture their shared preferences as good as social relations and congruity.
- The proposed framework *CR* always obtains the best performance. The reason is that using social relations or rater roles of users does not capture the closeness of users' preferences. This confirms the effectiveness of users' congruity in learning their preferences and improving the performance of recommenders.

Table 4. Performance comparison of different methods.

Training	Metrics	MF	SMF	SoReg	DualRec	CR
90%	MAE	0.9768 ± 0.0027	0.9578 ± 0.0028	0.9352 ± 0.0030	0.9231 ± 0.0026	**0.9136** ± **0.0029**
	RMSE	1.1687 ± 0.0029	1.1476 ± 0.0027	1.1294 ± 0.0030	1.1167 ± 0.0028	**1.1041** ± **0.0032**
70%	MAE	0.9848 ± 0.0029	0.9611 ± 0.0031	0.9417 ± 0.0027	0.9387 ± 0.0030	**0.9252** ± **0.0034**
	RMSE	1.1776 ± 0.0030	1.1597 ± 0.0029	1.1356 ± 0.0031	1.1253 ± 0.0028	**1.1172** ± **0.0030**
50%	MAE	0.9921 ± 0.0026	0.9702 ± 0.0027	0.9539 ± 0.0029	0.9471 ± 0.0029	**0.9335** ± **0.0031**
	RMSE	1.1894 ± 0.0031	1.1655 ± 0.0030	1.1478 ± 0.0031	1.1416 ± 0.0030	**1.1339** ± **0.0034**
40%	MAE	0.9969 ± 0.0029	0.9783 ± 0.0030	0.9582 ± 0.0029	0.9506 ± 0.0029	**0.9400** ± **0.0035**
	RMSE	1.1932 ± 0.0025	1.1761 ± 0.0028	1.1531 ± 0.0029	1.1446 ± 0.0029	**1.1378** ± **0.0031**

4.2 Integrating Congruity with Social Relations

Here, we investigate the impact of integrating congruity along with social relations and then study the effect of each to answer the second question. To achieve this goal, we define **CSRR** as a variation of our proposed method in which the preference closeness matrix $\mathbf{L} \in \mathbb{R}^{n \times n}$ is updated as follows:

$$\mathbf{L}_{ik} = \delta \mathbf{G}_{ik} + (1 - \delta)\mathbf{C}_{ik}. \tag{8}$$

As discussed earlier, $\mathbf{G}_{ik} \in [0, 1]$ and $\mathbf{C}_{ik} \in [-1, 1]$, therefore, to reduce the further complexities in our model, we replace $\mathbf{C}_{ik}$ by $\frac{\mathbf{C}_{ij}+1}{2} \in [0, 1]$. This further makes $\mathbf{L}_{ik}$ to be in $[0, 1]$. Also δ controls the contributions of $\mathbf{G}_{ik}$ and $\mathbf{C}_{ik}$. Here, we set $\delta = 0.3$. We now define the following variants of CSRR as follows:

- *CSRR–S*: Eliminates the effect of social relations by setting $\delta = 0$ in Eq. 8. This variation is equal to the **CR** method described earlier;
- *CSRR–C*: Eliminates the effect of congruity by setting $\delta = 1$ in Eq. 8;
- *CSRR–CS*: Eliminates the effects of both social relations and congruity by setting $\gamma = 0$ in Eq. 5. This variation is equal to the **MF** method.

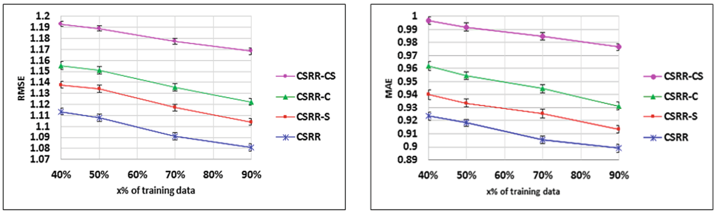
The results are shown in Fig. 3 for Epinions. The results for Ciao have very similar trends but they are omitted due to space limit. We observe the following:

- When we remove the effect of congruity, the performance of CSRR–C degrades compared to CSRR. We have the similar observations for the elimination of social relations. Those results support the importance of integrating congruity as well as social relations information in a recommender system.
- We note that the performance degrades more by eliminating congruity information, CSRR–C compared to eliminating social relations, CSRR–S. This is because the congruity information is much denser than social relations.
- Removing the effects of social relations and congruity, the performance of CSRR–CS reduces compared to CSRR–C and CSRR–S. This suggests that incorporating users' congruity along with social relations are important and have a complementary role to each other.

To recap, users' congruity and social relations are two different sets of information. Exploiting congruity information has potentials in more accurate measuring of users' opinions degree of match.

5 Related Work

Collaborative filtering methods are categorized into the neighborhood-based and model-based models. The low-rank matrix factorization methods are one example of model-based methods which estimate the user-item rating matrix using low-rank approximations method to predict ratings [13,19]. The increasing popularity of social media encourages individuals to participate in various activities which could provide multiple sources of information to improve recommender systems. Some algorithms incorporate user profile [18,24,26]. For example, the work of [18] constructs tensor profiles of user-item pairs while the method in [26] makes a profile for users using the initial interview process to solve the cold-start problem in the recommendation. Another method [25] considers both rater and



(a) RMSE

(b) MAE

Fig. 3. Effect of social relations and congruity in Epinions.

reviewer roles of each user to improve recommendation. It uses item ratings and helpfulness ratings to obtain the reviewer and rater roles, respectively.

Social relations also provide an independent source of information which brings new opportunities for recommendation [3,5,6,9,14,16,22]. The work of [16] incorporates user social relations which force a user's preferences to be close to her friends' and is controlled by their similarity which is measured based on item-ratings. The work of [9] proposes a random walk model based on users' friendship relations and item-based recommendation. The length of the random walk which is based on both item ratings and social relations. Another work [14] proposes a probabilistic framework which assumes that individuals preferences could be influenced by their friends' tastes. It fuses both users' preferences and their friends' tastes together to predict the users' favors on items. The importance of exploiting heterogeneity of social relations [23] and weak dependency connections for recommendation systems has been shown in [22]. To capture the heterogeneity of social relations and weak dependency connections, it adopts social dimensions by finding the overlapped communities in the social network.

The difference between CR and the above models is that we investigate the role of users' congruity as social relations alone do not demonstrate the degree of opinion match between users. Moreover, congruity is determined independently from social relationship information and is obtained from users' interactions to further capture different degrees of match between their opinions.

6 Conclusion and Future Work

In this paper, the concept of congruity, a degree of agreement and appropriateness between people, borrowed from sociology is tailored to discern different degrees of match between their opinions and enhance the performance of recommendation systems. To overcome the challenge that users' congruity is not readily available, we leverage the available users' interaction data and capture the congruity between users from data. We propose the framework CR, which predicts unknown user-item ratings by incorporating congruity information. We conduct experiments on real-world data, and the results confirm the efficiency of congruity for inferring users' opinions degree of match. In future, we would like to incorporate temporal information to study the dynamics of users' congruity in recommendation systems. Also, the findings of this work may be helpful for other tasks such as friend recommendation in social networks.

Acknowledgments. This material is based upon the work supported by, or in part by, Office of Naval Research (ONR) under grant number N00014-17-1-2605.

References

1. Alvari, H., Hajibagheri, A., Sukthankar, G., Lakkaraju, K.: Identifying community structures in dynamic networks. Soc. Netw. Anal. Min. **6**(1), 77 (2016)
2. Antheunis, M.L., Valkenburg, P.M., Peter, J.: The quality of online, offline, and mixed-mode friendships among users of a social networking site. Cyberpsychology **6**(3) (2012). https://cyberpsychology.eu/article/view/4272/3312. Article no. 6
3. Beigi, G., Jalili, M., Alvari, H., Sukthankar, G.: Leveraging community detection for accurate trust prediction. In: ASE International Conference on Social Computing, Palo Alto, CA, May 2014, June 2014
4. Beigi, G., Hu, X., Maciejewski, R., Liu, H.: An overview of sentiment analysis in social media and its applications in disaster relief. In: Pedrycz, W., Chen, S.-M. (eds.) Sentiment Analysis and Ontology Engineering. SCI, vol. 639, pp. 313–340. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30319-2_13
5. Beigi, G., Tang, J., Liu, H.: Signed link analysis in social media networks. In: Tenth International AAAI Conference on Web and Social Media (2016)
6. Beigi, G., Tang, J., Wang, S., Liu, H.: Exploiting emotional information for trust/distrust prediction. In: Proceedings of the 2016 SIAM International Conference on Data Mining, SIAM 2016, pp. 81–89 (2016)
7. Cocking, D., Kennett, J.: Friendship and the self. Ethics **108**(3), 502–527 (1998)
8. Enz, C.A.: The role of value congruity in intraorganizational power. Adm. Sci. Q. **33**(2), 284–304 (1988). ERIC
9. Jamali, M., Ester, M.: Trustwalker: a random walk model for combining trust-based and item-based recommendation. In: Proceedings of SIGKDD (2009)
10. Jiang, M., Cui, P., Liu, R., Yang, Q., Wang, F., Zhu, W., Yang, S.: Social contextual recommendation. In: Proceedings of CIKM (2012)
11. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: Proceedings of ACM SIGKDD (2008)
12. Koren, Y.: Collaborative filtering with temporal dynamics. Commun. ACM **53**, 89–97 (2010)
13. Koren, Y., Bell, R., Volinsky, C., et al.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
14. Ma, H., King, I., Lyu, M.R.: Learning to recommend with social trust ensemble. In: Proceedings of ACM SIGIR (2009)
15. Ma, H., Lyu, M.R., King, I.: Learning to recommend with trust and distrust relationships. In: Proceedings of RecSys, pp. 189–196. ACM (2009)
16. Ma, H., Zhou, D., Liu, C., Lyu, M.R., King, I.: Recommender systems with social regularization. In: Proceedings of ICWSM, pp. 287–296 (2011)
17. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: homophily in social networks. Ann. Rev. Sociol. **27**, 415–444 (2001)
18. Park, S.-T., Chu, W.: Pairwise preference regression for cold-start recommendation. In: Proceedings of RecSys, pp. 21–28. ACM (2009)

19. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems, vol. 20 (2008)
20. Schein, E.H.: Organizational Culture and Leadership. Wiley, Hoboken (2010)
21. Simpson, J., Weiner, E.S.C.: Oxford English Dictionary Online, vol. 6 (1989)
22. Tang, J., Wang, S., Hu, X., Yin, D., Bi, Y., Chang, Y., Liu, H.: Recommendation with social dimensions. In: Proceedings of AAAI (2016)
23. Tang, L., Liu, H.: Scalable learning of collective behavior based on sparse social dimensions. In: Proceedings of Conference on Information and Knowledge Management (2009)
24. Wang, J., Zhao, W.X., He, Y., Li, X.: Leveraging product adopter information from online reviews for product recommendation. In: Proceedings of ICWSM (2015)
25. Wang, S., Tang, J., Liu, H.: Toward dual roles of users in recommender systems. In: Proceedings of CIKM (2015)
26. Zhou, K., Yang, S.-H., Zha, H.: Functional matrix factorizations for cold-start recommendation. In: Proceedings of SIGIR (2011)

Mining Help Intent on Twitter During Disasters via Transfer Learning with Sparse Coding

Bahman Pedrood[1(✉)] and Hemant Purohit[2]

[1] Computer Science Department, George Mason University, Fairfax, VA, USA
bpedrood@gmu.edu

[2] Information Sciences and Technology Department,
George Mason University, Fairfax, VA, USA
hpurohit@gmu.edu

Abstract. Citizens share a variety of information on social media during disasters, including messages with the intentional behavior of seeking or offering help. Timely identification of such help intent can operationally benefit disaster management by aiding the information collection and filtering for response planning. Prior research on intent identification has developed supervised learning methods specific to a disaster using labeled messages from that disaster. However, rapidly acquiring a large set of labeled messages is difficult during a new disaster in order to train a supervised learning classifier. In this paper, we propose a novel transfer learning method for help intent identification on Twitter during a new disaster. This method efficiently transfers the knowledge of intent behavior from the labeled messages of the past disasters using novel Sparse Coding feature representation. Our experiments using Twitter data from four disaster events show the performance gain up to 15% in both F-score and accuracy over the baseline of popular Bag-of-Words representation. The results demonstrate the applicability of our method to assist realtime help intent identification in future disasters.

Keywords: Help seeking · Help offering · Intent · Transfer learning Crisis

1 Introduction

Social media and Web 2.0 provides a platform for citizens to discuss real-world disaster events, where they share a variety of information including messages to seek help [3]. A recent survey [27] shows the expectation of citizens to get a response from emergency services on social media. As a result, resource-constrained response agencies have started to leverage social media platforms to both communicate and monitor data to enrich their situational awareness for disaster response coordination [10,20,26,32]. Timely extraction of relevant social

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 141–153, 2018.
https://doi.org/10.1007/978-3-319-93372-6_16

Table 1. Examples of messages with intent to seek or offer help during recent disasters.

Social media message	Intent
I wanna give #blood today to help the victims #sandy	*Offering help*
Does anybody know a direct place to send clothing you want to donate to #HurricaneHarvey victims??	*Offering help*
OU students are seeking donations to provide relief and help victims of the Hurricane Harvey disaster: _URL_	*Seeking help*
You can help us send clean water, food, and shelter to those impacted by #HurricaneHarvey! #HoustonStrong	*Seeking help*

media messages allows a better understanding of needs in the affected community. Mining messages with intent to seek or offer help provides specific insights to assist resource coordination by bringing the awareness of actors seeking and offering resources to help. The uncoordinated efforts between such actors and the response organizations can lead to a second disaster of resource mismanagement[1]. Table 1 provides some examples of messages with help intent. However, the intent is not always explicitly expressed in social media messages due to a variety of ways to communicate intentionality. Actionable information can be identified by understanding intent during disasters, likewise, research on modeling intent in user queries on search engines improved the retrieval of relevant results [2].

Our problem of intent identification from social media messages is a form of text classification, however, intentional behavior is focused on future action (e.g., an act of offering donation) in contrast to topic or sentiment/emotion. Prior works on intent classification during disaster events [8,16,21,22,24,34] have developed event-specific supervised learning models using labeled dataset of the corresponding event. However, there are two key limitations of the prior research. First, supervised learning models developed for specific past events do not generalize due to differences in the distributions of the training event data and the testing event data. Second, developing a new supervised learning model onset a future disaster event would require preparing a large labeled dataset quickly to perform well. Therefore, we propose a novel method for the intent classification using transfer learning approach. Transfer learning focuses on leveraging knowledge gained while solving one problem (e.g., identifying intent in a past disaster) and applying it to a different but related problem (e.g., identifying intent in a future disaster). We study four disaster event datasets for the experimental evaluation with an intent class set of {*seeking*, *offering*, *none*} and present result analysis for different transfer learning settings of single and multiple disaster data sources. Our specific contributions are the following:

[1] https://www.npr.org/2013/01/09/168946170/thanks-but-no-thanks-when-post-disaster-donations-overwhelm.

- We present the first study of realtime intent identification for help-seeking and help-offering messages on social media in a future disaster event by leveraging past event datasets via transfer learning.
- We demonstrate the efficacy of a data-driven representation of Sparse Coding in contrast to the popular Bag-of-Words (BoW) model, to efficiently learn and transfer the knowledge of intentional behavior from past events.
- We evaluate the proposed method in transferring knowledge from both single and multiple past events to a future disaster event, and show performance up to 80% for F-score, indicating good prediction ability in general.

The rest of the paper is organized as follows. We first discuss related work on mining social media for crisis informatics and mining help intent in Sect. 2. We describe our proposed approach in Sect. 3 and experimental setup in Sect. 4. Lastly, we discuss the results and future work directions in Sect. 5.

2 Related Work

In the last two decades, there has been extensive research in the area of crisis informatics for the use of social media in all phases of pre, during, and post disasters (*c.f.* [3,11]). Among different types of social media analytics for disasters, the content-driven analyses are focused on the nature of social media messages including topics such as damage [12,35] and behaviors such as help-seeking [23,24]. User-based analyses include modeling of user attributes, such as trustworthiness [1] and types such as government agencies [17]. Network-based analyses are focused on information diffusion and user engagement patterns, such as communication of official response agencies [33] and retweeting [31].

The proposed research is closely related to content-based analysis of modeling help-seeking behavior on social media. Prior work by [24] identified actionable message category of seeking or offering help while studying Yushu Earthquake in 2010. [21,34] proposed supervised machine learning classifiers to identify and match messages with the complementary intent of request-offer during the events of Great East Japan Earthquake 2011 and Hurricane Sandy 2012 respectively. [22,23] proposed methods for supervised learning and linguistic rule-based classifiers to identify messages with seeking and offering help intent, however, without studying the generalization of methods to leverage the labeled data from past event datasets to mine intent in the future events. [16] proposed a system to classify requests for help during Hurricane Sandy 2012 using n-grams and context-based features. [8] studied the dynamics of messages coordinated by a common hashtag #*PorteOuverte* with the intent to seek or offer help during 2015 Paris Attack and developed an automated classifier for such messages. In the aforementioned works, there was a focus on developing event-specific methods to identify relevant messages with help intent. There is a lack of investigation on how to leverage and transfer the knowledge of intent behavior observed in the past events to help quickly identify relevant messages in the future events and thus, we propose to study transfer learning [18] techniques for mining intent.

3 Approach

We propose a novel approach of transfer learning method for the problem of real-time intent identification in future disasters. This problem is challenging due to differences in the probability distributions of source and target event datasets, the imbalance of intent classes across past and future events, and the lack of effective data representation for inferring intent from the short text.

To address the representation challenge in machine learning and natural language understanding, there is a growing interest in exploring Sparse Coding representation [6,13] in contrast to the popular BoW model. Sparse Coding provides a succinct representation of training text instances using only unlabeled input data, by learning basis functions (e.g., latent themes in the message text) to constitute higher abstract-level features. Given the complexity of expressing intent by multiple combinations of word senses in the text, we hypothesize the use of Sparse Coding to efficiently capture, learn, and transfer intent behavior from past disasters. Next, we describe dataset preparation for past disasters and the features for the proposed model of transfer learning with Sparse Coding.

3.1 Dataset Preparation

This study is based on Twitter messages ('tweets') collected during the past large-scale disasters with a focus on hurricanes and typhoons. We acquired two datasets of tweets annotated for help intent classes of {*seeking*, *offering*, *none*} from our past work [22], for two events – Hurricane Sandy 2012 and Supertyphoon Yolanda 2013. We also collected tweet datasets during the recent two disasters in 2017 – Hurricane Harvey and Hurricane Irma, which caused extensive devastation in the United States. We used Twitter Streaming API to collect English language tweets using 'filter/track' method for a given set of keywords (Harvey: {#harvey, hurricane harvey, Harvey2017, HurricaneHarvey, harveyrelief, houstonflood, houstonfloods, houwx}, Irma: {hurricane irma, hurricaneirma, HurricaineIrma, Hurricane Irma, HurricaineIrma, #hurricaneirma2017, #irma, #hurcaneirma}). We collected 8,342,404 tweets from August 29 to September 15 for Hurricane Harvey and 861,503 tweets between September 7 to September 21 for Hurricane Irma. For labeling of help intent classes {*seeking*, *offering*, *none*} in each event dataset, we employed a biased sampling approach to increase the coverage of help intent messages given the sparse distribution of intent classes observed in the first two datasets. First, we randomly sampled 2000 tweets from the full dataset of an event and second, we randomly sampled 2000 tweets from the donation-classified subset of messages that provide context for expressing intent. We used the related work's donation topic classifier [21]. We asked three human annotators (no author was involved) to label a tweet into the three exclusive help intent classes for each event and chose the majority voting scheme for finalizing the labels. The resulting labeled class distribution for both the acquired and the collected datasets[2] is shown in Table 2.

[2] Datasets are available upon request, for research purposes.

3.2 Feature Representation

Prior work in crisis informatics for social media text classification has extensively used BoW model for representing text messages (*c.f.* survey [11]). However, the BoW representation model limits capturing context and semantics of the text content [5,6], which is essential for inferring intent. Since BoW model loses ordinal information of text content, one direction of research to tackle this challenge is to enhance the text representation in a way that can preserve some ordered information of the words, such as N-grams [19]. Because N-gram representation adds extra terms to the word vocabulary, the problem of curse of dimensionality gets worse. The increase in the feature space, especially for the tasks of small-scale datasets and short text, causes the loss of generalization of the training models and results in a negative effect in the prediction capability.

Table 2. Labeled data distribution for help intent classes across four disaster events, ordered by time of occurrence.

Event	Month	Seeking	Offering	None
Hurricane Sandy	*2012 (October)*	1626	183	1326
Supertyphoon Yolanda	*2013 (November)*	197	91	475
Hurricane Harvey	*2017 (August)*	816	331	2853
Hurricane Irma	*2017 (September)*	530	244	3226

Sparse Coding is an effective approach for reducing dimensionality of feature space that generally assumes an over-complete basis set for the input data and it is capable of a complete description and reconstruction of the input data. Also, every input data point can be described using a linear combination on a small number of the new basis vectors. According to the theory of compressed sensing, when the data are distributed on an underlying manifold with the characterizing bases, only a small number of bases are required to fully describe any arbitrary point on the manifold [4]. Sparse Coding representation has shown significant improvements in transfer learning in the recent years. While most of the works addressed the challenges of image processing and computer vision [7,9,15,25], there are few research studies for text analytics [14]. However, the previous works primarily required a large amount of data in either source or both source and target domains to be effective, and cannot be applied in our crisis informatics case directly. Because both source and target sets are not only at the small scale but also contain redundant and incomplete information in short text.

For feature extraction, we first perform standard text-preprocessing on the text message to remove stop-words, replace numbers (e.g., money donation mentions), user mentions, and URLs with constant characters as well as lower-casing the tokenized text of messages. After constructing a vocabulary of the extracted tokens, every message is represented by a real-valued vector of vocabulary entries, where each component holds the *tf-idf* value of that entry [30]. After the text samples are transformed to numeric vectors of tf-idf values, a sparse representation of the vectors in a feature space with a significantly reduced dimension is learned and explained in Sect. 3.3.

3.3 Learning Model: Transfer Learning with Sparse Coding

To tackle the intent mining problem when both source and target data are small, we propose a novel transfer learning approach. Transfer learning has become a popular solution to bring information from past experiences to better characterize the data and class distribution of them. Generally, there exists a large number of labeled data samples from past relevant experiences as well as a large set of unlabeled samples for the new problem, where the two sets are assumed to share common information. The goal of transfer learning is to find better transformations to map the distribution of one of the datasets to the other and learn a predictive model using the transformed data to find a stronger prediction of the new data. Fig. 1 shows an abstract workflow of our proposed method.

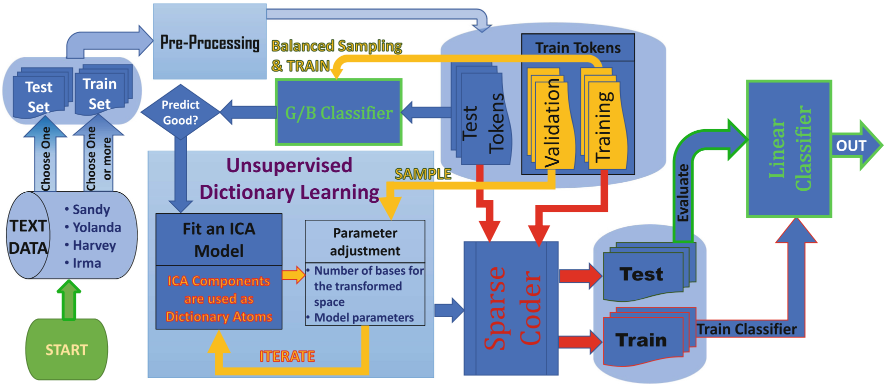


Fig. 1. The sequence of steps in the proposed model. On the top, after pre-processing of train and test data, test set is used for unsupervised learning of *dictionary atoms*. The optimal parameters of the model are learned by cross-validation on the training data. Then a linear classifier is learned from the coded samples in the sparse space.

To formalize the approach, assume a set of message text data, where each data point is coming from a *Domain* $\mathcal{D}$. The given points are usually assumed to be sampled independently and randomly from the domain $\mathcal{D}$. A domain $\mathcal{D}_l = (\mathcal{X}_l, \mathcal{Y}_l)$ is defined as a distribution of pairs of $\{x_{il}, y_{il}\}$, where $\forall\{x_{il}, y_{il}\} \in \mathcal{D}_l : x_{il} \in \mathcal{X}_l, y_{il} \in \mathcal{Y}_l$ and the domain assumes a joint probability distribution $p\left(\mathcal{X}_l, \mathcal{Y}_l|\theta_l\right)$. In transfer learning and domain adaptation, we have a labeled *Source* dataset D_S, where data is given as N_S pairs of observations and labels, $\{X_S, Y_S\}$, and those pairs are assumed as i.i.d. samples from the joint probability distribution of the source domain $p(\mathcal{X}_S, \mathcal{Y}_S|\theta_S), \mathcal{D} = \{\mathcal{X}_S, \mathcal{Y}_S\}$. The data of interest is defined as *Target* dataset D_T, where only X_T is given, but Y_T is unknown. The target data points are also assumed as N_T i.i.d. samples from the target domain $\mathcal{D}_T = \{\mathcal{X}_T, \mathcal{Y}_T\}$, which along with their *latent* labels are jointly distributed as $p\left(\mathcal{X}_T, \mathcal{Y}_T|\theta_T\right)$. For a domain $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, a *task* is defined as a function $f(.)$ on the attribute space $\mathcal{X}$ of that domain to predict the labels $\mathcal{Y}$. In transfer learning, the task is to find a $f(x_{it})$ that approximates y_{it}. In other words, a function $f(x_{it})$ must be learned to predict the values of $\hat{y}_{it}$ for every $x_{it} \in X_T$ that minimizes $\sum_{i=1}^{N_T} d(\hat{y}_{it}, y_{it})$, where $d(.,.)$ is some distance measure of interest. Considering the target domain distribution of $p\left(\mathcal{X}_T, \mathcal{Y}_T|\theta_T\right)$, in order to have the most accurate predictions, ideally $f(x_{it})$ must become the closest possible to $p(y_{it}|x_{it}, \theta_T)$. The posterior distribution of target data is unlikely to estimate without having any labels for this data. Since the task in our problem is semantically a complete equivalence in both source and target data, the basic assumption of transfer learning for homogeneous datasets is valid in our context and the joint probability of the labels is assumed to be similar in both source and target domains. Formally:

$$p(X_S) \times p(Y_S|X_S, \theta_S) = p(X_T) \times p(Y_T|X_T, \theta_T) \tag{1}$$

Note that although the joint distribution is assumed as equal for the two datasets, the conditional probabilities are generally not. If the conditional probabilities were equal, then the key difference lies in the marginal distributions. Geometrically, source and target samples are assumed as being scattered in different regions in the feature space. Since $P(X_S) \neq P(X_T)$, the parameters θ_S and θ_T should also be different to compensate the difference of marginal distributions in the conditional probabilities. To fill the gap between the datasets, most of the research in transfer learning is about finding a good transformation $T_{TS}(.) : \mathcal{X}_T \rightarrow \mathcal{X}_S$ from the target to source domain, then using a classifier trained on the source data to predict the labels for the transformed target data: $\hat{Y_T} = f(X_T^{ts}|\hat{\theta_S})$. While this approach works well in some cases, since none of the characteristics of the target data is incorporated in training the classifier, the generalization of the model for the target data reduces. A better approach is a reverse form of transformation: finding a transformation $T_{ST}(.) : \mathcal{X}_S \rightarrow \mathcal{X}_T$ from source data to make the marginal distribution of transformed source data similar to the target data.

Our Sparse Coding approach is also about using the unlabeled target data to find a mapping for the source data. The proposed method is distinctive in handling the limitations of intent mining in the specific case of crisis informatics by providing a solution to work with a small number of short-text samples. Instead of only transforming the data, the proposed model combines the domain transfer with the feature reduction step to make the representation generalize well. Fig. 1 shows the steps in the proposed learning model, where it shows how target data is used to build the model, while independent subsets of training data tailor the representation. Sparse Coding in general is about using the characteristics of input data $X \in \mathbb{R}^{N \times M}$ while finding a set of basis $B \in \mathbb{R}^{N \times K}$ that is over-complete on the underlying manifold of data and then, approximating each input instance as a linear combination of a small number of those bases (*atoms*) in the *dictionary*. If the manifold assumptions hold on the data, Sparse Coding guarantees a perfect reconstruction. However, the assumption cannot be easily confirmed, especially when dealing with a small number of data instances. Also, prior research strongly suggests using an under-complete basis for Sparse Coding in classification [28,29]. The general form of Sparse Coding is about minimizing $\|X - BA\|_2^2$ while, simultaneously approximating the bases and the sparse code vectors incorporates the following optimization problem:

$$\underset{B,a_i}{\text{argmin}}\ \frac{1}{2}\left\|\sum_{i=1}^{M} x_i - Ba_i\right\|_2^2 - \lambda \|a_i\|_1 \tag{2}$$

where $\|.\|_p$ is the p-norm. The main objective function that tries to minimize the information loss of coding is a loss function for reconstructing the input sample x_i from the transformed sample a_i, formulated as a Sum of Squared Error (SSE) function. In the regularization term, the parameter λ is introduced to control the sparsity of the coded samples and the *first norm* is used for efficient convergence to a sparse solution. It guarantees a more generalizable solution for the objective function by removing the coded coefficients with an energy level less than λ.

Prior to Dictionary Learning, a logistic classifier is learned on training data in order to improve quality of data for learning the representation, by contrasting the most important intent classes (*seeking*, *offering*) against the rest (*other* class). Then the dictionary is learned by applying Independent Component Analysis (ICA) method only on the target samples that are classified as relevant on this classifier. To find the best parameter setting for the dictionary, the parameters λ and K (the number of atoms) are selected with a focus on optimizing the resulting metric (F-score) using a 10−fold Cross-Validation over the important classes *seeking* and *offering* of the training data.

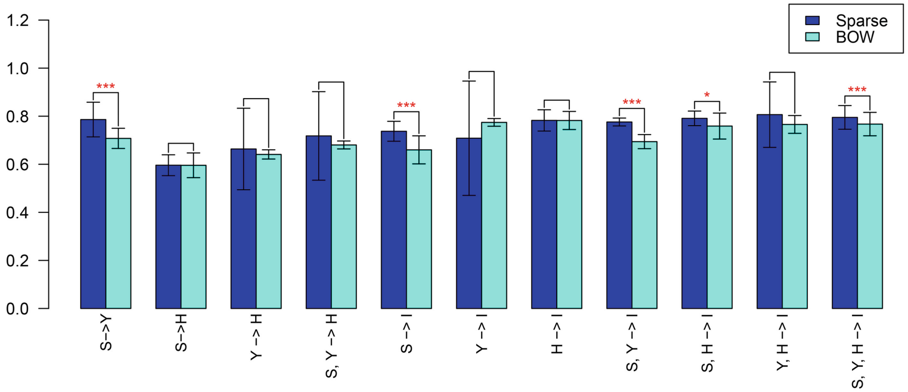


Fig. 2. Accuracy for predicting intent classes across different experimental settings. The X-axis represents the source to target events, where *S = Hurricane Sandy, Y = Super-typhoon Yolanda, H = Hurricane Harvey, I = Hurricane Irma.* Single *and* Triple Asterisks *show where the difference is significant at* 99% *and* 99.99%, *respectively.*

4 Experimental Setup

We employ two different experimental schemes for evaluating the performance of our proposed transfer learning method as follows:

- ***Single past disaster as source***: In this case, we use only one past disaster dataset to learn the features for the source training set and a future disaster event dataset is used as a target test set.
- ***Multiple past disasters as source***: In this case, we use more than one past disaster event datasets with different combinations as the source training set and a future disaster event dataset is used as a target test set.

Given the different possible selections of train/test set, we chose to consider only the case of predicting on the *future* disaster event given the source of *past* events on the timeline (*c.f.* Table 2). For evaluating the effectiveness of the proposed Sparse feature representation, we created a baseline of BoW representation for the features. The results are provided by repeating the experiments 20 times. A Wilcoxon's Signed-Rank test is used to find significant contrasts between the results of BOW and Sparse representation at 99% level and the significant cases are marked in the figures and tables by the *asterisks*.

5 Results and Discussion

This section discusses experimental results for the defined schemes of single and multi-event sources as well as the benefits of the Sparse Coding representation.

5.1 Performance Analysis

Figure 2 shows the results for both single and multi-source experiment settings. Overall, We note the following key observations from the results:

1. The accuracy was achieved close to 80%, which shows good predictability for the proposed model, given the complexity of detecting both explicit and implicit intent expressions from textual utterances.
2. We generally found the superior performance of the Sparse representation in contrast to the BoW representation across both single and multi-source experiments. In the setting of the source dataset as Yolanda and the target as Irma, the poor performance is likely due to the small size of the Yolanda dataset for training. We suspect this role of the small dataset in influencing the performance through the multi-source type experiments.
3. We note better performance in the experiments of leveraging multiple event datasets as the source in contrast to only the single event data source, which is likely contributed by the ability of Sparse representation to effectively capture intent cues across diverse disaster contexts.

We further observe the following points from Table 3:

Table 3. Details of the experimental setup and performance for F-score metric. K shows the number of atoms and λ holds for the regularization term in Sparse Coding.

Train	Test	Scheme	K	λ	F-score		
					Seeking	None	Offering
Sandy	Yolanda	Sparse	22	0.009	**0.744 ± 0.072***	**0.831 ± 0.092***	**0.682 ± 0.055***
		BOW			0.681 ± 0.037	0.758 ± 0.053	0.559 ± 0.035
Sandy	Harvey	Sparse	74	0.02	**0.529 ± 0.026***	0.719 ± 0.048	**0.187 ± 0.016**
		BOW			0.501 ± 0.027	**0.720 ± 0.060**	0.172 ± 0.017
Yolanda	Harvey	Sparse	37	0.0005	0.425 ± 0.107	**0.797 ± 0.235**	**0.174 ± 0.031***
		BOW			**0.515 ± 0.017***	0.769 ± 0.017	0.146 ± 0.013
Sandy	Irma	Sparse	17	0.007	**0.457 ± 0.040**	**0.847 ± 0.034***	**0.170 ± 0.015***
		BOW			0.443 ± 0.040	0.786 ± 0.056	0.146 ± 0.017
Yolanda	Irma	Sparse	12	0.0015	0.449 ± 0.154	0.837 ± 0.284	**0.188 ± 0.037**
		BOW			**0.535 ± 0.029**	**0.879 ± 0.010**	0.119 ± 0.022
Harvey	Irma	Sparse	55	0.0085	**0.605 ± 0.040**	**0.882 ± 0.034**	**0.276 ± 0.025**
		BOW			0.583 ± 0.026	0.879 ± 0.029	**0.316 ± 0.029***
Sandy, Yolanda	Harvey	Sparse	57	0.01	0.520 ± 0.126	0.768 ± 0.279	**0.172 ± 0.033***
		BOW			**0.546 ± 0.017**	**0.794 ± 0.014**	**0.128 ± 0.012**
Sandy, Yolanda	Irma	Sparse	78	0.0007	**0.571 ± 0.023***	**0.881 ± 0.012***	**0.159 ± 0.012***
		BOW			0.501 ± 0.032	0.815 ± 0.023	0.141 ± 0.016
Sandy, Harvey	Irma	Sparse	87	0.007	0.504 ± 0.026	**0.889 ± 0.022**	**0.303 ± 0.016***
		BOW			**0.527 ± 0.035***	0.866 ± 0.046	0.257 ± 0.023
Yolanda, Harvey	Irma	Sparse	90	0.0003	0.397 ± 0.085	**0.898 ± 0.133**	**0.297 ± 0.025**
		BOW			**0.585 ± 0.023***	0.870 ± 0.029	0.288 ± 0.025
Sandy, Yolanda, Harvey	Irma	Sparse	90	0.0015	**0.594 ± 0.041**	**0.890 ± 0.038***	0.273 ± 0.023
		BOW			0.568 ± 0.036	0.870 ± 0.040	**0.286 ± 0.019***

1. A general pattern for F-score indicates the better performance of Sparse representation than BoW in predicting important classes of help *offering* and *seeking*. The potential factor for the varied performance of Sparse Coding representation in Table 3 is the difference in the labeled sample size and class distribution of source datasets (as shown in Table 2).
2. We observe the low predictive power for the *offering* class, which is likely affected by the imbalanced class distribution and the lowest number of labeled instances for this class across each dataset.
3. We further note a direct relation between the size of source data and the optimal number of Atoms (K) that is discovered using Cross-Validation process as explained in Sect. 3.2 and shown in Fig. 1. Given a larger dataset may contain more information, it would likely require more bases for information representation.

5.2 Limitation and Future Work

We presented a novel approach of transfer learning for efficiently identifying help-seeking or offering intent in a future disaster event, in contrast to event-specific supervised learning methods requiring large labeled datasets for the future event. While we presented the experiments for disaster events of similar types (hurricane and typhoon), future work could investigate the performance of transfer learning of help intent across the types of disasters, such as earthquakes and hurricanes. We considered English language tweets due to the complexity of understanding intent from the short text and our natural next step is to leverage the proposed method for cross-language intent identification. Further, the proposed method provides a general framework for boosting any input text representation by adding a layer of Sparse Coding using the explained workflow in Fig. 1. For instance, a future study can enhance the learning performance while using more advanced input representations such as n-grams, word vectors, etc. for different types of learning tasks.

6 Conclusion

This paper presented a novel approach of transfer learning with Sparse Coding feature representation for the task of help intent identification on social media during disasters. We experimented with four disaster event datasets ordered over time and analyzed the performance of the proposed model for both single-event source and multi-events source schemes to predict in a target (future) event. Our results showed that using the Sparse Coding representation model in contrast to the baseline model of Bag-of-Words representation allows the efficient transfer of knowledge for intent behaviors from past disaster datasets. The application of the proposed approach can enhance real-time social media filtering tools during future disaster responses, when there would be insufficient event-specific labels to train a supervised learning model.

Acknowledgment. Authors thank Rahul Pandey for discussion and proofreading as well as acknowledge the partial support from U.S. National Science Foundation (NSF) grant IIS-1657379. Opinions in this article are those of the authors and do not necessarily represent the official position or policies of the NSF.

References

1. Abbasi, M.-A., Liu, H.: Measuring user credibility in social media. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 441–448. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37210-0_48
2. Ashkan, A., Clarke, C.L.A., Agichtein, E., Guo, Q.: Classifying and characterizing query intent. In: Boughanem, M., Berrut, C., Mothe, J., Soule-Dupuy, C. (eds.) ECIR 2009. LNCS, vol. 5478, pp. 578–586. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-00958-7_53

3. Castillo, C.: Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press, Cambridge (2016)
4. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
5. Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., Smith, N.A.: Retrofitting word vectors to semantic lexicons. In: Proceedings of NAACL (2015)
6. Faruqui, M., Tsvetkov, Y., Yogatama, D., Dyer, C., Smith, N.A.: Sparse overcomplete word vector representations. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (vol. 1: Long Papers), pp. 1491–1500 (2015)
7. Han, Y., Wu, F., Zhuang, Y., He, X.: Multi-label transfer learning with sparse representation. IEEE Trans. Circ. Syst. Video Technol. **20**(8), 1110–1121 (2010)
8. He, X., Lu, D., Margolin, D., Wang, M., Idrissi, S.E., Lin, Y.R.: The signals and noise: actionable information in improvised social media channels during a disaster. In: Proceedings of the 2017 ACM on Web Science Conference, pp. 33–42. ACM (2017)
9. Huang, Z., Pan, Z., Lei, B.: Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data. Remote Sens. **9**(9), 907 (2017)
10. Hughes, A.L., St Denis, L.A., Palen, L., Anderson, K.M.: Online public communications by police & fire services during the 2012 hurricane sandy. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 1505–1514. ACM (2014)
11. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: a survey. ACM Comput. Surv. (CSUR) **47**(4), 67 (2015)
12. Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., Meier, P.: Practical extraction of disaster-relevant information from social media. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1021–1024. ACM (2013)
13. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: Advances in Neural Information Processing Systems, pp. 801–808 (2007)
14. Lee, H., Raina, R., Teichman, A., Ng, A.Y.: Exponential family sparse coding with application to self-taught learning. In: IJCAI, vol. 9, pp. 1113–1119 (2009)
15. Maurer, A., Pontil, M., Romera-Paredes, B.: Sparse coding for multitask and transfer learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML 2013), pp. 343–351 (2013)
16. Nazer, T.H., Morstatter, F., Dani, H., Liu, H.: Finding requests in social media for disaster relief. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1410–1413. IEEE (2016)
17. Olteanu, A., Vieweg, S., Castillo, C.: What to expect when the unexpected happens: social media communications across crises. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 994–1009. ACM (2015)
18. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
19. Peng, F., Schuurmans, D.: Combining naive bayes and *n*-gram language models for text classification. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 335–350. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36618-0_24
20. Plotnick, L., Hiltz, S.R.: Barriers to use of social media by emergency managers. J. Homel. Secur. Emerg. Manag. **13**(2), 247–277 (2016)

21. Purohit, H., Castillo, C., Diaz, F., Sheth, A., Meier, P.: Emergency-relief coordination on social media: automatically matching resource requests and offers. First Monday **19**(1) (2013)
22. Purohit, H., Dong, G., Shalin, V., Thirunarayan, K., Sheth, A.: Intent classification of short-text on social media. In: 2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity), pp. 222–228. IEEE (2015)
23. Purohit, H., Hampton, A., Bhatt, S., Shalin, V.L., Sheth, A.P., Flach, J.M.: Identifying seekers and suppliers in social media communities to support crisis coordination. Comput. Support. Coop. Work (CSCW) **23**(4–6), 513–545 (2014)
24. Qu, Y., Huang, C., Zhang, P., Zhang, J.: Microblogging after a major disaster in china: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 25–34. ACM (2011)
25. Quattoni, A., Collins, M., Darrell, T.: Transfer learning for image classification with sparse prototype representations. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8. IEEE (2008)
26. Reuter, C., Ludwig, T., Kaufhold, M.A., Spielhofer, T.: Emergency services' attitudes towards social media: a quantitative and qualitative survey across europe. Int. J. Hum. Comput. Stud. **95**, 96–111 (2016)
27. Reuter, C., Spielhofer, T.: Towards social resilience: a quantitative and qualitative survey on citizens' perception of social media in emergencies in Europe. Technol. Forecast. Soc. Change **121**, 168–180 (2016)
28. Rodriguez, F., Sapiro, G.: Sparse representations for image classification: learning discriminative and reconstructive non-parametric dictionaries. Technical report, Minnesota Univ Minneapolis (2008)
29. Singh, O.P., Haris, B., Sinha, R.: Language identification using sparse representation: a comparison between GMM supervector and i-vector based approaches. In: 2013 Annual IEEE India Conference (INDICON), pp. 1–4. IEEE (2013)
30. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. **28**(1), 11–21 (1972)
31. Starbird, K., Palen, L.: Pass it on?: retweeting in mass emergency. In: Proceedings of the 7th International ISCRAM Conference-Seattle, vol. 1. Citeseer (2010)
32. Stephenson, J., Vaganay, M., Coon, D., Cameron, R., Hewitt, N.: The role of Facebook and Twitter as organisational communication platforms in relation to flood events in Northern Ireland. J. Flood Risk Manag. (2017)
33. Sutton, J.N., Spiro, E.S., Johnson, B., Fitzhugh, S.M., Greczek, M., Butts, C.T.: Connected communications: network structures of official communications in a technological disaster. In: Proceedings of the 9th International ISCRAM Conference (2012)
34. Varga, I., Sano, M., Torisawa, K., Hashimoto, C., Ohtake, K., Kawai, T., Oh, J.H., De Saeger, S.: Aid is out there: looking for help from tweets during a large scale disaster. In: ACL, vol. 1, pp. 1619–1629 (2013)
35. Vieweg, S.E.: Situational awareness in mass emergency: a behavioral and linguistic analysis of microblogged communications. Ph.D. thesis, University of Colorado at Boulder (2012)

People2Vec: Learning Latent Representations of Users Using Their Social-Media Activities

Sumeet Kumar(✉) and Kathleen M. Carley

Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA
{sumeetku,kathleen.carley}@cs.cmu.edu

Abstract. In most social network studies, it is assumed that nodes are simple and carry no information, and links are explicit ties such as friendship. Which nodes are in which group is determined as a function of these explicit ties. For example, given a set of random walks through the network, it is possible to learn a vector for each node which contains a latent representation of the node. These latent representations have useful properties that can be easily exploited by statistical models for tasks like identifying groups and inferring implicit links. However, most existing representation learning methods ignore node attributes. In many cases, there is a rich body of information and events associated with nodes that also can be used for node clustering and to infer ties. In social media, e.g., an explicit relationship is friendship, and another is the follower-followee relation. Besides, there are the set of messages passed by the users, as well as, their activities in the form of liking or mentioning. What is needed is a way of collectively using both the explicit ties and this rich body of additional information in learning these latent node representations. Combining such data should enable more effective link inference and grouping strategies. In this research, we propose *People2Vec* an algorithm to learn representations that takes into account proximity between users due to their social media activities. We validate our model by experiments on two different social-media datasets and find the model to perform better than prior state-of-the-art approaches.

1 Introduction

Online social media platforms are popular for sharing information and allowing users to network with each other. Analyzing such social networks is an active area of research. Significant problems in this field are finding communities, predicting interests of users, and recommending friends and content. Typical end goals are to identify groups, infer links and make network predictions. To achieve these goals, it is first necessary to determine for two social-media users, how socially proximate they are. Two individuals who are connected by a friendship tie or a follower-followee tie are more socially proximate than are those not connected. However, just examining the binary friendship or follower-followee ties is insufficient for assessing there overall social proximity. Users are more

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 154–163, 2018.
https://doi.org/10.1007/978-3-319-93372-6_17

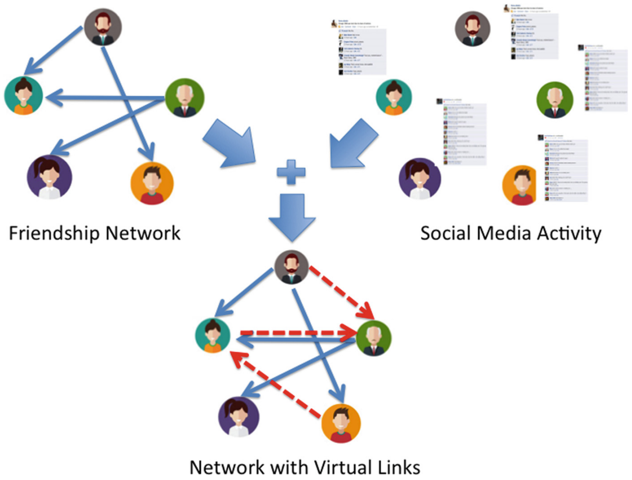


Fig. 1. Using just the binary friendship network, the only similarity between the orange and green person is that they are both friends of the brown person. Each person also has social media activity, e.g. a set of messages that they send. Combining these two types of data can generate new weighted virtual links (the red dashed lines) between nodes and reveal hidden connections. (Color figure online)

socially proximate vis-a-vis a topic, the more they have in common. Learned representations using random walks over the network links provide a better feature to find the social proximity of users, but they still miss important cues such as what they say or feel about a topic. We argue, that in addition to the explicit ties among users, the activities and preferences of the social media users could be used to find weighted *virtual links* that can be leveraged to learn more useful node representations (Fig. 1).

In this paper we present People2vec, a latent space representation of the user in context as a vector. This representation supports generalization and the identification of similar actors and so groups. This representation is learned from the data and captures the complexities of the situation in which the user is embedded. Our approach is inspired by an analogous problem in language technology which is to learn a representation of a word from the common context of the word in sentences. To that end we draw on the word2vec approach for identifying the context of the node by random traversals of the network similar to [4,8]. However, we move beyond this approach by bringing in user activities as node attributes, which enables the inference of additional links and solves the sparse network problem. Existing network analytic tools, like ORA [2], uses both the network and the attributes on the nodes. By exploiting attribute information to infer the network, People2vec supports a more detailed analysis.

The important contributions of this paper are:

- We propose People2vec, a model to learn node representations that captures similarity in users' attributes and activities, in addition to their friendship links.
- Our model extends the popular random walk approach of learning node representation and brings valuable improvements, yet preserves the simplicity of the approach.

The paper is outlined as follows. First, we discuss prior work in Sect. 2. Then, we introduce our 'People2Vec Model' in Sect. 3. In Sect. 4, we present our experiments on two different datasets, along with a discussion of the results. We finally conclude and discuss the future work.

2 Related Work

Some recent advancements in learning node representations are inspired from the improvement in natural language processing. Bengio et al. [1] proposed the distributed representations of words aka 'Word embeddings' which was later used by Collobert and Weston [3] to demonstrate their usefulness in many NLP tasks. Mikolov et al. [6] proposed Word2vec, a Skip-gram model for learning high-quality distributed vector representations using skip-gram model. Such representations capture many syntactic and semantic word relationships e.g. they predict: vec('Berlin') - vec('Germany') + vec('France') = vec('Paris'). The concept of learning representations using skip-gram could be applied to networks as well. These latent representations can be learned in a number of ways; e.g. (a) factorization of social network's adjacency matrix (b) learning functions to find better features. Recently, researchers have tried to train neural-networks to find efficient nonlinear transformations for learning node embeddings. However, unlike words in a language that has plenty of examples to get related contextual words, often networks are sparse. Besides, there is no clear notion of social-context in networks. To incorporate context in networks, researchers have tried random walks [4,8]. Random walks on links in a network help to generate contexts that consist of proximity nodes. Like in language models, such context is then used to build embedding vectors (representations) often by training a shallow neural network. Learning low dimensional representation of nodes in networks allow mapping local structural characteristics to a continuous space representation. Learning these representations of nodes have helped to improve performance in many tasks including node classification and link prediction. Though models proposed in [4,8,9] learn good representation on simple graphs, they mostly explore binary edges, so are best suited for social-networks that have clear links as in friendship and follower-followee relationship. They do not exploit node attributes and preferences which are often very relevant and strong indicators in social networks. This gap is the focus of this research.

3 People2Vec Model

We consider the problem of learning node representation in social-networks that captures users' preferences, in addition to their explicit links in the form of friendship or follower-followee relationship. We expect a good solution to have the following two properties:

(a) Users with direct links in a network should be closer in latent representation space. Many existing models exhibit this property.

(b) Users with similar preferences should be closer in latent representation space. The way to measure similarity in preferences should be flexible to allow the model to adapt to different formulations of preferences. For example, in one situation, two users discussing a topic could be similar, and in another situation, two users using same tag could be similar.

We formulate the problem as follows: Let $G(V, E)$ be a network where $v \in V$ are nodes (or users) and $e \in E$ are edges. Let $F : v \rightarrow \mathcal{R}^d$ be the function that learns a d dimensional representation (z) of a node. Let Y be a matrix of user preferences that contains a set of preferences for each user, where Y_i^j indicates preferences of node v_i towards jth item. The jth 'item' could be stance towards a topic or the count of a tag (as in hashtag used by a user). The goal of the algorithm is to learn z_i, a low-dimensional representation of user v_i that considers the explicit links in E and also the similarity in Y space.

As in Deepwalk [8], we follow the language modeling technique of generating latent representation of words from sentences. In language modeling, given some text corpus W = $(w_0,\ w_1, \ldots,\ w_n)$, the goal is to maximize the likelihood $Pr(w_i|w_0, w_1, \ldots w_{i-1}, w_{i+1}, \ldots, w_n)$ over the entire corpus. By analogy, in social networks, we define the likelihood of observing a node v_i given other nodes by $Pr(v_i|v_0, v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_n)$. In the latent space F of node representations, this can be formulated as maximizing the likelihood of

$$Pr\Big(v_i|\big(F(v_{i-k}), F(v_{i-k+1}), \ldots, F(v_{i-1}), F(v_{i+1}), F(v_{i+k-1}), F(v_{i+k})\big)\Big) \tag{1}$$

where we only consider $2k$ immediate neighbors on node v_i. To efficiently solve such a formulation, we use the skip-gram [6] approach. Taking log the optimization problem can be formulated as:

$$\min_F - \log Pr\Big(\big(v_{i-k}, v_{i-k+1}, \ldots, v_{i+k-1}, v_{i+k}\big)|F(v_i)\Big) \tag{2}$$

Since a node and its latent vector have symmetry in latent space, the conditional likelihood of a neighbor node v_j given by $Pr(v_j|F(v_i))$ can be approximated as similarity in latent space. As in Deepwalk [8], we use the stochastic gradient descent over neighbor nodes collection generated by random walk to optimize the final objective function. To speed up the training we used hierarchical soft-max [7]. In the proposed model, the neighborhood of a node v_i $(v_0, v_1, \ldots v_{i-1}, v_{i+1}, \ldots, v_n)$ is not limited to nodes reachable by explicit links, and is extended to nodes having similar attributes. We call such links '*virtual links*' (explained next).

3.1 Virtual Links from Users' Activities

We define *virtual links* as edges in social-networks that connect users with similar preferences as evident by their involvement on these platforms. These virtual links (like real links) can be used in random-walks to explore node neighborhood, thus enhancing the network information present in the original networks. Moreover, there virtual edges, based on nodes similarities, can also strength the existing linkage. Hence, virtual links enable to learn more meaningful node representations.

There are two possible formulations of virtual-links. A rigid link that is either present or absent, and a weighted link that that gives a probabilistic score of links being present. We go with the probabilistic version of virtual-links as it enables a more flexible learning framework.

The probability of having a virtual link between two users is based on similarity between their activity on social-media platform. As discussed earlier, let's model the activity profile of users as a matrix Y_i^j, where Y_i^j indicates preferences of node v_i towards jth item. Similarity between users is obtained by measuring similarity between vectors representing users preferences.

The similarity between two users is defined as:

$$Similarity\big(v_k, v_i\big) = Sim\big(Y_k, Y_i)\big) \tag{3}$$

There are a number of ways to measure such similarity. We tried three such similarity measures: (a) Cosine Similarity (b) Hamming Distance (c) Euclidean Distance. On our datasets, we find that 'Cosine Similarity' (CS) performs better than other measures. Hence, we use CS as the similarity measure in rest of the paper.

Random Walks. We use random walks to sample neighborhood nodes. The random walks approach provides a more flexible approach over known 'Depth First' and 'Breadth First' sampling as explained in [4]. A random walk starts from a node, say v_0, and uses node links to find the next node. In prior studies, the probability of transition from node v_0 to v_i is given by:

$$P(v_i|v_0) = \begin{Bmatrix} 0, \text{if}(v_0, v_i) \notin E \\ \frac{1}{\sum_k 1}, \text{otherwise} \end{Bmatrix} \tag{4}$$

where k = Number of links of v_0.

Random Walks over Virtual Links. In our model, we extend the random walk over nodes to include random walk over virtual links. Figure 2 explains the idea.

To weigh the relative importance of virtual links and to real links, let's introduce a hyper-parameter α, a weighing factor. This parameter is tuned based on characteristics on the network under consideration. For random walks over

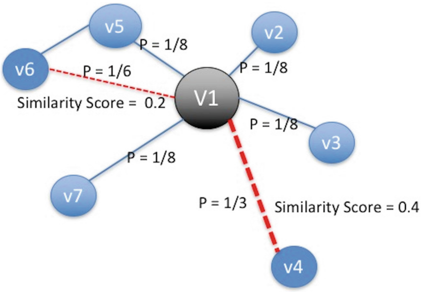


Fig. 2. An example of a random walk transition in People2vec. Real links are shown in blue and virtual links are shown in red. The walk originates from node v1. The probability of transition to other nodes is shown in text. Similarity scores for virtual links are obtained using *Sim* function. Here we consider the weighing factor $\alpha = 0.5$, i.e. we weigh the real links and the virtual links equally. For clarity, we have not shown virtual links for nodes already connected via real-links. (Color figure online)

virtual links (see Fig. 2), we use the probability of transition from node v_0 to v_i as:

$$P(v_i|v_0) = \left\{ \begin{array}{l} \alpha * Sim'\big(Y(v_0), Y(v_i)\big), \text{if}\,(v_0, v_i) \notin E \\ (1-\alpha) * \frac{1}{\sum_k 1} + \alpha * Sim'\big(Y(v_0), Y(v_i)\big), \text{if}\,(v_0, v_i) \in E \end{array} \right\} \quad (5)$$

where k = Number of real links of v_0, and Sim' is the normalized similarity score defined as $Sim' = \frac{Sim\big(Y(v_0),Y(v_i)\big)}{\sum_{v_i} Sim\big(Y(v_0),Y(v_i)\big)}$.

3.2 People2Vec Algorithm

People2Vec extends the original Deepwalk algorithm [8] by including virtual links. Like in Deepwalk, our algorithm uses random walk to learn node representation. However, in the process of generating walks, the algorithm uses 'virtual links' in addition to the real links, thus considers neighbors that were not accessible in plain random walks. The steps as described in Algorithm 1. Again, similar to Deepwalk, we use skip-gram [6] algorithm to efficiently learn the representations for each node.

4 Experiments and Results

In this section, we present the experimental evaluation of our model on two different datasets. The first dataset is a set of blogs and another is a sample of Flickr website. We also evaluate the impact of using different latent-space dimensions.

4.1 Datasets

We use two existing datasets (BlogCatalog and Flickr) to evaluate our algorithm. These datasets are publicly available and were used is earlier studies [5].

Algorithm 1. The People2Vec algorithm.

```
learnRepresentations
Graph G, VirtualGraph $G_v$
RepresentationDimension d
walks per node r
walk length l
window size w
$walks$ = []
for $iter \in \{1, \ldots, r\}$ do
   for all $node \in V$ do
      $walk$ = randomWalk(G, $G_v$, node, l)
      $walks$.append($walk$)
   end for
end for
SkipGram(F, $walks$, w) (see Ref. [6])
————————————————————————————
randomWalk(Graph G, VirtualGraph $G_v$, Start node u, Length l)
$walk$ = [u]
for $walk\_iter \in \{1, \ldots, l\}$ do
   $currentNode$ = walk[-1]
   $newNode$ = getNeighbor($currentNode$, G, $G_v$)
   $walk$.append($newNode$)
end for
————————————————————————————
getNeighbour(Start node $v_0$, Graph G, VirtualGraph $G_v$)
start at v
$N_r$ = getRealNeighbors(G, $v_0$)
$N_v$ = getVirtualNeighbors($G_v$, $v_0$)
pick neighbor $v_1$ from $[N_v + N_r]$ using $P(v_1|v_0)$ (See Eqn. 5)
```

Table 1. Dataset Description

Dataset	Nodes	Edges	Labels	Attributes
BlogCatalog	5,196	171,743	6	8,189
Flickr	7,575	239,738	9	12,047

BlogCatalog Dataset. BlogCatalog is an online community of bloggers. The dataset is created by including keywords used in blog description as attributes [5]. Using those keywords, we generate users preference matrix. In this dataset, the labels used for predicting the classification performance represent bloggers' interests (Table 1).

Flickr Dataset. Flickr is popular website that hosts videos and images. Users can follow each other, thus, forming a network. They can join different groups which is used as labels. To get the users' preference matrix, tags by users are used [5].

4.2 Baseline Algorithms and Model Optimization

We measure the performance of People2Vec against several state-of-the-art algorithms [4,8,9]. **DeepWalk** [8] uses uniform random walks on networks to learn embeddings as in language modeling techniques like word2vec. **LINE** [9] algorithm preserves both local and global network structures and uses edge sampling approach for optimization. **Node2Vec** [4] extends Deepwalk by combining depth-first and breadth-first search in their sampling strategy.

We use social-media tags to create a nodes' preference vectors. Preference vectors of different nodes are then compared using cosine similarity measures to create weighted virtual links which are later used to learn node representations. Because similarity between nodes generates $O(n^2)$ possible edges, which could result in a very dense graph so we use a threshold to reduce the number of virtual-edges used in learning embeddings. The exact threshold is of lesser importance as People2vec random-walk prefers node transitions to higher similarity nodes, and thus ignores less similar nodes more often. The dimensions of representations used are 64 and 128. For a fair comparison, all algorithms used $walklength = 10$ and $walkcount = 40$. For all models, we use one-vs-rest logistic regression as used in [8]. We trained all the models on an Ubuntu Linux machine with 64 GB ram and eight core Intel i7 processor with 4.00 GHz processing speed.

4.3 Experimental Results

We present the results of the experiments. Table 2 shows the top classification performance for the two datasets. Figure 3 shows the trend of F1-score (macro) for different train and test ratio. The plot show that most algorithms have a reasonable performance right from the smallest training percentage (10%). There are small improvements as we increase the training data percentages. Peopl2vec performed better than rest of the algorithms. In general, embeddings with 128 dimension perform better than 64 dimension embeddings. People2Vec is an exception for which scores were very similar. Figure 4 shows the results for the Flickr dataset. Again for most algorithms 128 dimensional embedding performed better than 64 dimensional. Like before, Peopl2vec performed better than all other algorithms.

Table 2. Best F1 macro score

Method	BlogCatalog	Flickr
DeepWalk	0.73	0.58
Node2vec	0.72	0.58
LINE	0.73	0.58
People2Vec	**0.83**	**0.75**

We don't compare our results with LANE [5] (BlogCatalog: 0.90 best F1 score, Flick: 0.90 best F1 score) as their implementation uses matrix factorization

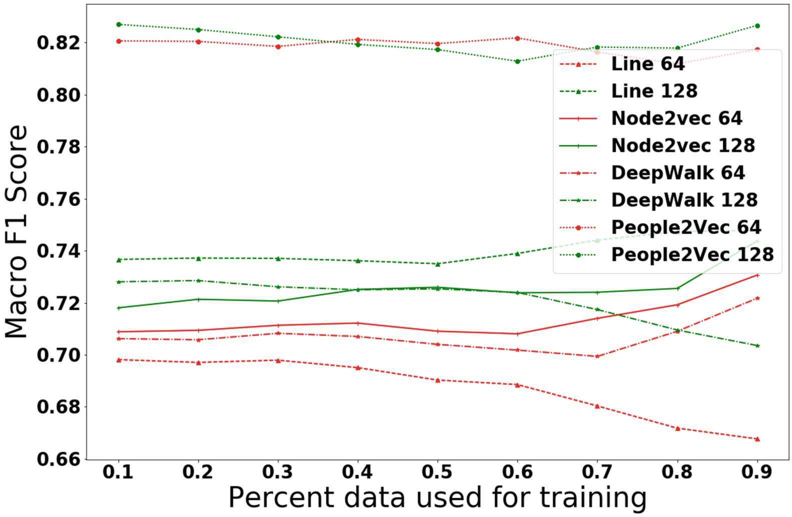


Fig. 3. We tried different algorithms to learn the class labels (a proxy of community) of nodes in BlogCatalog graph. In this plot, we compare mean F1 score for different algorithms.

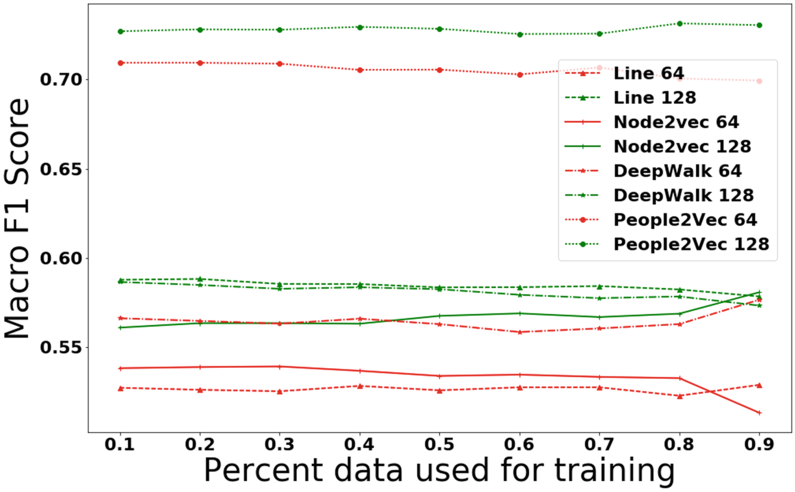


Fig. 4. We tried different algorithms to learn the class labels (a proxy of community) of nodes in Flickr graph. In this plot, we compare mean F1 score for different algorithms.

approach, which is a very different from random-walk based approaches compared in this work. However, we observe similar performance gains over the baselines e.g. on BlogCatalog dataset, LANE improved from 0.81 (Deepwalk) to 0.90 (LANE), which is similar to ours, which improved from 0.73 (Deepwalk) to 0.83 (People2vec).

5 Conclusions and Future Work

We proposed People2Vec, a model to learn the representation of nodes in complex networks. We used nodes similarity to construct virtual links between nodes. Virtual links enhance the original graph with additional information that uses users' attributes and preferences. People2Vec considers these virtual edges while building random-walk paths, thus, exploits similarity of nodes in addition to real links. Experiments on two real-world datasets reveal that using People2Vec to learn representations substantially improves node classification performance. On the BlogCataog dataset, we observe an improvement of 13% (F1 score) over other state-of-the-art algorithms to learn embeddings. On the Flickr dataset, the performance improved by 25% on F1-score.

People2Vec is straightforward and intuitive, yet learns better node representations. The approach is also very general, hence can easily be extended to other types of data on users. In future, we plan to investigate the usage of sentiment and emotions in social media posts, to learn node representations that consider the stance of users towards topics. We would also like to explore confidence levels on the results for sparse graphs.

Acknowledgments. This work was supported in part by the MURI Award No. N000140811186, MURI Award No. N000141712675 and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors only.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
2. Carley, K.M.: ORA: A Toolkit for Dynamic Network Analysis and Visualization. Springer, New York (2017). https://doi.org/10.1007/978-1-4614-7163-9_309-1
3. Collobert, R., Weston, J.: A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 160–167. ACM (2008)
4. Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 855–864. ACM (2016)
5. Huang, X., Li, J., Hu, X.: Label informed attributed network embedding. In: Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. pp. 731–739. ACM (2017)
6. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013) arXiv preprint arXiv:1301.3781
7. Morin, F., Bengio, Y.: Hierarchical probabilistic neural network language model. In: Aistats, vol. 5, pp. 246–252. Citeseer (2005)
8. Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 701–710. ACM (2014)
9. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1067–1077. International World Wide Web Conferences Steering Committee (2015)

Finding Organizational Accounts Based on Structural and Behavioral Factors on Twitter

Sultan Alzahrani[1,2](✉), Chinmay Gore[1], Amin Salehi[1], and Hasan Davulcu[1]

[1] CIDSE, Arizona State University, Tempe, AZ 85287, USA
{ssalzahr,cgore1,asalehi1,hdavulcu}@asu.edu
[2] King Abdulaziz City for Sciences and Technology (KACST), Riyadh 11442, Kingdom of Saudi Arabia
szahrani@kacst.edu.sa

Abstract. Various socio-political organizations, from activist groups to propaganda campaigners, create accounts on Twitter to reach out, influence and gain followers. In order to analyze the impact of these organizational accounts, the first step is to identify them. In this paper, we develop and experiment with a set of network-based, behavioral, temporal and spatial characteristics in these accounts, independent of domain or language, to identify features that can be useful in detecting organizational accounts. In order to assess this model, we experimented with a microblog corpus comprised of over 7 million tweets from 150,000 Twitter users in Bangladesh, tweeted between June and October 2016. We sampled 31,139 accounts using cold-start heuristics to locate and label nearly 200 organizational accounts, distributed as 68 NGOs, 62 news outlets, 35 political groups, and 17 public intellectual and iconic figures. The remaining accounts were labeled as individuals. Next, we developed a set of features and experimented with a set of linear and non-linear classifiers. The highest performing sparse logistic regression classifier achieved an accuracy of 68.2% precision and 64.4% recall leading to a 66.2% F1-score in detecting less than 1% rare organizational accounts using a set of content- and language-independent features.

Keywords: Social network · Social network analysis
Automatic identification

1 Introduction

Social media has emerged as an integral part of a connected lifestyle, in which people locate and interact with each other to stay informed on current issues and to help shape their own opinions. Twitter is a micro-blogging social network which allows users to post, like and share or retweet short messages or tweets. According to the most recent Twitter statistics[1], more than 330 million active

[1] Twitter Usage Statistics Report, Internet Live Stats, http://www.internetlivestats.com/twitter-statistics/.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 164–175, 2018.
https://doi.org/10.1007/978-3-319-93372-6_18

daily users post more than 500 million tweets daily, a quarter of which are tagged with hashtags. 80% of active Twitter users are outside the United States. Twitter supports over 40 languages and allows users to connect with their friends as well as organizational accounts such as religious, political and educational groups, NGOs, news outlets, public figures, and celebrities.

Several works focus on developing methods to predict a Twitter account's demographic attributes such as age [20], gender [10], location [18], and political affiliation [8]. Related works on individual vs. organization (IvO) detection include those by De Silva and Riloff, McCorriston et al. and Kim et al. [5,6,11]-bearing in mind the significant role of organizational accounts for being what is called "mouthpieces" of disseminating ideologies and aiding mass mobilization within both political and communication context [19]. IvO detection algorithms were previously used for ground truth labeling in community detection, where key organizational accounts, such as political parties and their leaders, were located and their followers, who like and share their messages, were grouped into different communities. Unlike Facebook pages and groups, Twitter does not explicitly support the notion of an organizational account, hence detecting them remains an open problem.

In order to calibrate and evaluate a classification model that could detect organizational accounts on Twitter, we experimented with a 5-month Twitter corpus comprising over 7 million tweets from Bangladesh. We sampled 31,139 accounts using cold-start heuristics presented in Sect. 3 to locate and label nearly 200 organizational accounts classified as 68 NGOs, 62 news outlets, 35 political groups, and 17 public intellectual and iconic figures. The remaining 30,957 accounts were labeled as individual user accounts. We then experimented with a set of network-based, behavioral, temporal and spatial features, all independent of domain and language, to identify relevant features useful in differentiating between IvO accounts. Organizational accounts[2] correspond to a rare category of accounts in this corpus, with less than 1% frequency. Following this, we experimented with a set of linear and non-linear classifiers. The highest performing sparse logistic regression classifier achieved an accuracy of 68.2% and 64.4% recall leading to a 66.2% F1-score in detecting organizational accounts using the set of domain- and language-independent features presented in Sect. 4.

The rest of the paper is organized as follows: Sect. 2 features a review of related works. Section 3 describes the Bangladesh tweet corpus that we used and our cold-start heuristic methods-based ground truth collection process. Section 4 explains the content-independent, network-based, spatio-temporal and behavioral features employed in our experiments. Section 5 presents experimental evaluations and findings. Finally, Sect. 6 concludes the paper and discusses future work.

[2] Since public intellectuals, leaders and celebrities share similar spatial, behavioral and connectivity related characteristics with group and organizational accounts, we labeled them together as "organizational" - as opposed to "individual" accounts.

2 Related Works in IvO Type Detection

Rao et al. [9] propose a method to infer a user's demographic account information, for e.g., age, gender, political affiliation and location, by detecting latent features in their messages. Recent work by Varol et al. [12] works on resolving a bot-detection issue by employing a diverse set of features categorized into user profile, followers, network, temporal, content and sentiment related ones. They identified at least 9% of accounts as bots. However, organizational accounts seem to be a rarer category than bots. A paper by Wu et al. [14] focuses on categorizing Twitter accounts and then analyzing the flow of information between them. Their focus is on gathering evidence to validate the two-step communication flow model developed by Katz [16]. In this model, elite users are divided into 4 categories - mass media, celebrities, organizations and bloggers. A set of discriminating keywords was generated and a score calculated for unseen accounts based on their tweet content to enable classification. Recent work by Savage et al. [22] develops anomaly detection methods suitable for use in social media to detect anomalous account patterns such as malicious spammers, fraudsters, cyberbullies and predators. Their methods also use language- and content-independent network and behavioral features for detecting anomalous accounts.

Three related studies examine the IvO account detection problem. De Silva and Riloff [5] propose a classification model that depends on multi-lingual content features tested on English and Spanish datasets. McCorriston et al. [6] employ a set of network (e.g., ratio of followers to friends, etc.), temporal (e.g. tweet volumes), spatial and content-related features to distinguish between individuals and organizations on Twitter. Kim et al. [11], utilizing network- and content-based features to identify organizational accounts, reported that content-based features were the key determiners of account types. The key contribution of our method is that, it employs a set of *content- and language- independent* network behavioral and spatial-temporal features for detecting organizational accounts.

3 Tweet Corpus

According to a 2016 census, Bangladesh's population is more than 168 million[3]. The number of Internet users in Bangladesh now stands at over 66.8 million, which means there is a 41% penetration. Facebook, with a usage of about 97.2%, is the most used social network while Twitter ranks second with 1.08% usage – approximately 1.7 million accounts. Our tweet corpus includes all tweets tweeted between June and October 2016 that were either geo-tagged as being from within Bangladesh or from users whose account location on their profile mention a city or a place in Bangladesh. The tweet corpus statistics are as shown in Table 1.

[3] Internet World Stats. Bangladesh Internet Usage and Telecommunications Reports. http://www.internetworldstats.com/asia/bd.htm.

Table 1. Tweets dataset

Feature	Value
Number of tweets	7,090,560
Number of users	150,000
Minimum timestamps	June 1, 2016
Maximum timestamps	October 31, 2016
Tweets location	Bangladesh
Languages used	English, Bangla

Table 2. Ground truth data

Account type	Count
Individual	30,957
NGOs	68
News	62
Political organizations	35
Celebrities	17

Table 3. User × User behavioral interactions network

Feature	Retweet network	Mention network	Followers network
Nodes	308,477	335,678	12,604,797
Edges	681,404	431,437	25,942,312
Connected components	4,305	11,755	1079
Average node degree	75	48	1078
GCC nodes	299,890	298,337	12,600,824
GCC edges	675,682	405,813	25,939,414

3.1 Ground Truth Labeling

The ground truth labeling followed a set of cold-start heuristics yielding 31,139 candidate accounts, which were further validated by a human labeler. Accounts were sorted in descending order according to their PageRank and degree centrality measures of retweet, follower and user-mentioned networks in an attempt to leverage the high importance of organizational presented in their accounts which are indicated by high centrality measures. We identified those accounts by their centralities as a preliminary step before identifying organizational accounts then a validating manual labeling step to follow on. The top 200 accounts on each list were retrieved and manually labeled. We also used heuristic matching rules by employing translations of the same keyword lists used by Wu et al. [14] to detect candidates to label, based on the following rules:

News: We created a list of the popular TV channels and newspapers in Bangladesh using Wikipedia. The Twitter handles for each channel/newspaper on this list were then matched within the corpus.

Celebrities: A list of Bangladeshi movie actors, politicians and public figures was collated from Wikipedia and matching handles were located and verified in the corpus. Some additional celebrities and other types of organizational accounts were located and labeled by taking a closer look at accounts with the largest number of followers.

Political Organizations: We made a list of political parties and groups with the help of three political scientists from Bangladesh, matching their names with

Twitter handles and profile information. Additionally, several Bengali keywords indicating political, social, religious groups and organizations were used to match and validate Twitter handles as belonging to them.
NGOs: A list of Bengali keywords indicating NGOs was created and matched, alongside local branches of globally active NGOs, for e.g., Red Cross, listed on Wikipedia.
Individuals: We created a list of regular expressions like "I am," "I'm," "I love," "I work at," "I like," etc. in Bengali to match and label roughly 30 K accounts as individuals. Table 2 shows the frequency distribution of each type of of labeled account as per the above heuristic approaches expectedly revealing imbalanced data and reflecting the reality of the actual population of accounts; hence imposing challenges in our training models - addressed in the experiment Sect. 5.

4 Features

In this section, we describe the network, behavioral and spatio-temporal features, listed in Table 4, that were used to differentiate between IvO accounts. These features are content- and language-independent, i.e., they rely only on the non-textual features of the Twitter accounts. Retweet, follow and user-mentioned networks were also created from the tweet corpus. The connectivity statistics of these networks are shown in Table 3.

4.1 Network Features

We created three directed weighted graphs based on retweet, follower and user-mentioned information. This means a directed edge is added from user A to B if user A retweeted user B, or if user A mentioned user B, or if user A follows user B in the corpus. The following subsections show the various network centrality measures we experimented with:

Table 4. Feature sets used in our learning model

Features				
Network			Account profile	Tweet location & Timestamp
Retweet	Followers	User-mentioned		
-Degree centrality	-Degree centrality	-Degree centrality	-Number of users	-Location entropy
-Pagerank centrality	-Pagerank centrality	-Pagerank centrality	-Number of favourites by users	-Location variance
-K-core centrality	-K-core centrality	-K-core centrality	-Friends to followers ratio	-Timestamp entropy
-Clustering coefficient	-Clustering coefficient	-Clustering coefficient	-Hashtag centralities and clustering coefficient	-Timestamp variance

Degree Centrality: This measure of a given node represents the fraction of nodes it is connected to Sabidussi [21]. The higher number a node has, the more nodes it connected. This is further divided into in-degree and out-degree, where values are normalized to keep them bound between [0,1]. Figure 1 (a, e and i) charts the log-log centrality distributions for IvO accounts. From this, we observe that news organizations and celebrities have high in-degree and low out-degree centralities, and organizational accounts are located towards the head of the corresponding power law distributions. For organizational accounts, user-mentioned centralities tend to be lower as they are mentioned more often than they mention others.

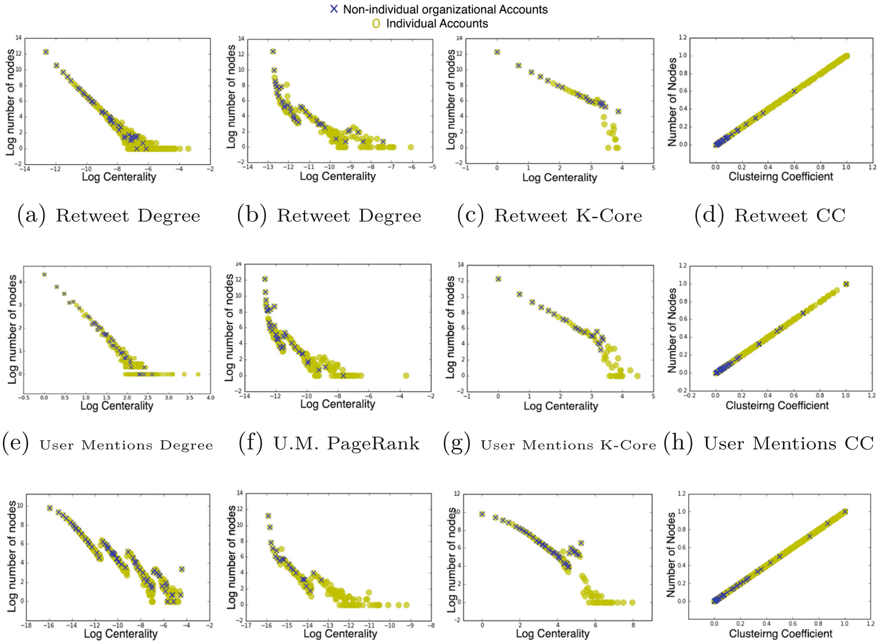
PageRank Centrality: This is an extension of the eigenvector centrality [2]. PageRank can be computed iteratively and the accounts' values explain their relative importance by using a damping factor until convergence. In Fig. 1 (f and j), we observe that organizational accounts tend to have lower PageRanks in follower and user-mentioned networks [13].

K-core Centrality: Another important centrality measure [23], the k-core decomposition process is initiated by removing all nodes with the degree $k = 1$. This causes new nodes with the degree $k \leq 1$ to appear. These are also removed, and the process is continued until the only nodes remaining are those of degree $k > 1$. The removed nodes and their associated links form the 1-shell. This pruning process is repeated for the nodes of degree $k = 2$ to extract the 2-shell, that is, in each stage the nodes with degree $k \leq 2$ are removed. The process is carried on until all higher-layer shells have been identified and all network nodes have been removed. In Fig. 1 (c, g and k), it can be observed that organizational accounts' k-core values are clustered towards the head of the distribution, separated from individual accounts' k-core measures, which are clustered towards the tail.

Clustering Coefficient: This coefficient is a measure of the degree of which nodes in a graph tend to cluster together [25]. For instance, friends of friends are likely to have a high clustering coefficient with the coefficient's value close to 1, while sparse and rarely connected graphs display a coefficient value closer to 0 [17]. We hypothesized that followers' connectivity levels would vary based on the type of account. For example, organizations would have a diverse set of followers who are not connected to each other. The graphs in Fig. 1 (d, h and l) show the clustering coefficients of IvO accounts where organizational accounts tend to cluster around lower coefficients with some outliers.

4.2 Tweet Location and Timestamp

Tweets originate from various locations and have different timestamps. Especially in the case of individuals tweeting from their mobile devices, users' spatial-temporal behaviors might help differentiate them from more stationary organizational accounts. Descriptions of the entropy- and variance-based temporal and spatial features that we experimented with are as follows:



(a) Retweet Degree (b) Retweet Degree (c) Retweet K-Core (d) Retweet CC

(e) User Mentions Degree (f) U.M. PageRank (g) User Mentions K-Core (h) User Mentions CC

(i) Followers Degree (j) Followers PageRank (k) Followers K-Core (l) Followers CC

Fig. 1. Degree Centrality, Page Rank, K-Core and Clustering Coefficients for three types of networks: Between, User-Mentioned, and Follower Behaviors. All points are not displayed, only 50 sampled points were taken for visualization, bearing in mind that some data points overlay others

Temporal Features: For every user, we have a distribution of timestamps marking times from when the user tweeted. Interestingly, entropy measures failed to capture a significant difference between IvO accounts, however, variance measures based on timing and locations of tweets showed a higher ability to display variations.

Spatial Features: We utilized geo-tagged tweets to compute spatial entropy and variance measures for accounts with spatial information, and found that organizational accounts tend to show lower variance in geolocation indicating their posting locations are more stationary. Entropy, nonetheless, did not display enough discriminative power, suggesting that variance can be better explained by geolocation deviation around the means per account.

4.3 User Profile Features

We gathered profile information for all the users in the database, such as user descriptions and their favorites count. We extracted the following features from user profiles:

Followers to Friends Ratio: This is one of the indicative features proposed by Can et al. [3] in their study, which showed a high correlation of user's posts being retweeted with the proposed ratio. We plotted the followers to friends ratio for all types of accounts and observed that organizational accounts have a lower friends-to-followers ratio.

Favorites Count: Twitter allows users to like any tweet and this information is captured as their favorites count. We observed that individual users have higher favorite counts as compared to organizational accounts [4].

List Count: Twitter lists allow users to create a curated list of Twitter accounts. On their timeline, users can then view a feed of tweets originating from the members of their lists. This functionality was introduced in 2013, and was first used as feature by Yasugi et al. [15]. Users can create their own lists or subscribe to preexisting ones. List counts represent the number of lists users are members of, and it was observed that organizational accounts tend to have lower list counts.

Username Frequency: Keywords in individuals' usernames tend to match many others, whereas there are keywords in organizational accounts' names that tend not to match frequently. When all keywords in a name are repeated multiple times among other Twitter handles, then the account is likely to belong to an individual. This variance motivated us to include it as feature in our learning model.

Hashtag Network: Hashtags are common terms or phrases that identify messages related to a specific topic or an event. Users who share common hashtags usually talk about similar topics. We made use of each accounts' hashtag usage to build a hashtag network as highlighted by Wagner et al. [7] study on distilling features from the social network. We began by collecting all the hashtags used by every account. Then for every common hashtag between accounts, we added a weighted edge, where accounts sharing multiple hashtags had higher weights. Next, we calculated centrality measures and clustering coefficients for each account in the hashtag network.

5 Experiments

This section highlights our experimental results and evaluations in detecting the rare organizational accounts in our tweet corpus – including social, political and religious groups, NGOs, celebrities, and public figures.

5.1 Normalizing the Measures

We normalized all feature values to make them scale invariant. For centrality measures and clustering coefficients, the values were already normalized to lie between 0 to 1, except for the k-core centrality. User profile-based features such as friends to followers ratio, favorites count and name frequency were normalized as well.

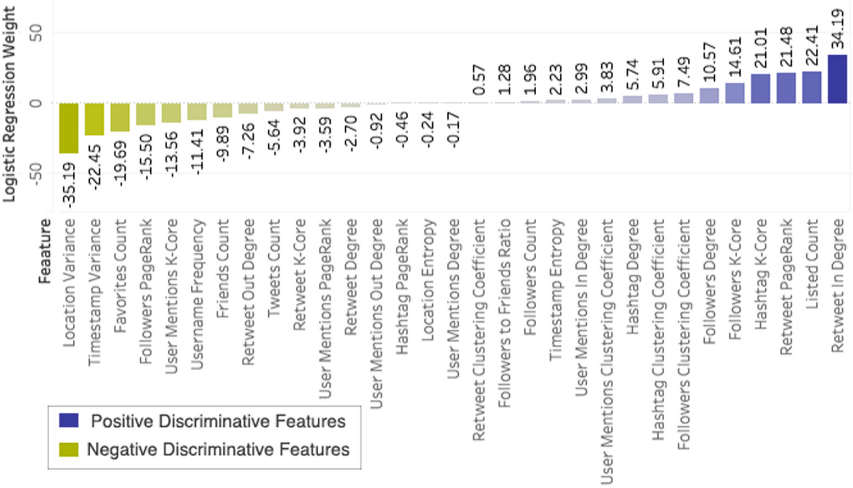


Fig. 2. Logistic regression model weights.

5.2 Detection of IvO Accounts

Besides showing labeled data distributions across various features for visual inspection in Sect. 4, Fig. 1, here we also conducted a statistical test to show that individual accounts behave differently from organizational ones. We employed a multivariate two-sample test between individual and organizational accounts, as proposed by Baringhaus and Franz [1]. We experimented with 1,000 permutation bootstrap-replications between two types of account distributions. The *critical-value* measure was obtained as 9362.8 based on 95%-value-of-confidence, and an *observed statistical value* of 38,9025 was calculated, which is significantly larger than the *critical-value*; thus confirming the alternative hypothesis that distributions corresponding to IvO accounts are significantly different from one another.

Next, we considered the labeled data to evaluate the performance of the classifiers. We reported precision, recall, and f-measure metrics for each category of accounts using 10-fold cross-validation on the resultant balanced data adapted by using the undersampling method [24] targeting individual dominant class. Table 5 shows that the logistic regression-based model outperformed other models such as Random Forrest, Multilayer Perceptron (MLB), AdaBoost and Gaussian Naive Bayes. Discriminating feature weights of the logistic regression model can be utilized to understand positively and negatively correlated features for individual and organizational account categories. In the logistic regression model, positively weighted features relate to organizational typed accounts, whereas negatively weighted features relate to the individual typed accounts. Figure 2 shows the corresponding weights of all features employed in the model.

5.3 Discriminating Features for Organizational Accounts

Organizational accounts tend to have higher valued:

- Retweet pagerank: Organizational accounts tend to be retweeted by other central users
- Hashtag k-core: Organizational accounts use central hashtags towards the core of the hashtag network
- Followers k-core: Organizational accounts are located towards the core of the followers network
- Followers clustering coefficient: Followers of organizational tend to follow each other as well
- Retweet in-degree centrality: Organizational accounts tend to be retweeted more often
- List count: Organizational accounts have higher list counts

Table 5. Classifier Performance for Individual and Organizational Account Types

Classifier	Precision	Recall	F1 Score
Logit Regression	**.682**	**.644**	**.662**
RandomForest	.563	.689	.62
MLP	.426	.611	.502
AdaBoost	.293	1.	.453
GaussianNB	.309	.622	.413

(a) Classification Performance the rare Organizational Accounts

Classifier	Precision	Recall	F1 Score
Logit Regression	**.989**	**.991**	**.99**
RandomForest	.991	.985	.988
MLP	.988	.976	.982
AdaBoost	1.	.93	.95
GaussianNB	.988	.96	.974

(b) Classification Performance for the Dominant Individual Accounts

5.4 Discriminating Features for Individual Accounts

Individual accounts tend to have higher valued:

- Spatial variance: Individuals are likely to Tweet from diverse locations
- Timestamp variance: Individuals have a random pattern of tweeting, where as organizational accounts are more time structured
- Favorites count: Individuals like/endorse others' Tweets more often compared to organizational accounts
- Followers pagerank: Organizational accounts tend not to follow many others
- User mentions k-core: Organizational accounts tend not to mention many others

6 Conclusions

In this paper, we proposed a classification model to detect organizational accounts on Twitter by employing a set of network-based, behavioral, temporal and spatial features independent of content and language. Community detection is a popular method employed to understand the network structure. Communities can further be described by locating their key influencers. Our method can be employed to automatically detect organizational accounts such as religious, political and educational groups, NGOs, news outlets, public figures and icons located in any community, thus helping name the key groups and actors driving these collectives.

Acknowledgements. This work was supported by the Office of Naval Research under Grant No.: N00014-16-1-2015, and N00014-15-1-2722.

References

1. Baringhaus, L., Franz, C.: On a new multivariate two-sample test. J. Multivar. Anal. **88**, 190–206 (2004)
2. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. J. Math. Sociol. **2**(1), 113–120 (1972)
3. Can, E.F., Oktay, H., Manmatha, R.: Predicting retweet count using visual cues. In: Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. ACM (2013)
4. Celli, F., Rossi, L.: The role of emotional stability in twitter conversations. In: Proceedings of WSASM. Association for Computational Linguistics (2012)
5. De Silva, L., Riloff, E.: User type classification of tweets with implications for event recognition. In: ACL 2014, p. 98 (2014)
6. McCorriston, J., et al.: Organizations are users too: characterizing and detecting the presence of organizations on twitter. In: ICWSM (2015)
7. Wagner, C., et al.: The wisdom in tweetonomies: acquiring latent conceptual structures from social awareness streams. In: Proceedings of the 3rd ISSW. ACM (2010)
8. Conover, M.D., et al.: Predicting the political alignment of twitter users. In: Privacy, Security, Risk and Trust (PASSAT), pp. 192–199. IEEE (2011)
9. Rao, D., et al.: Classifying latent user attributes in Twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents. ACM (2010)
10. Burger, J., et al.: Discriminating gender on Twitter. In: EMNLP 2011 (2011)
11. Kim, S.M., et al.: Distinguishing individuals from organisations on twitter. In: WWW 2017 (2017)
12. Varol, O., et al.: Online human-bot interactions: Detection, estimation, and characterization (2017) arXiv preprint arXiv:1703.03107
13. Page, L., et al.: The pagerank citation ranking: bringing order to the web. Technical report, Stanford InfoLab (1999)
14. Wu, S., et al.: Who says what to whom on twitter. In: Proceedings of the 20th International Conference on World Wide Web. ACM (2011)
15. Yasugi, N., et al.: Use of twitter as an instrument for disseminating public information in providing public goods and roles of e-government: evidence from Japanese prefectures. Int. J. Eng. Innovative Technol. **3**, 128–133 (2013)
16. Katz, E.: The two-step flow of communication: an up-to-date report on an hypothesis. Public Opin. Q. **21**, 61–78 (1957)
17. Klemm, K., Eguiluz, V.M.: Growing scale-free networks with small-world behavior. Phys. Rev. E **65**, 057102 (2002)

18. Mahmud, J., Nichols, J., Drews, C.: Home location identification of twitter users. ACM Trans. Intell. Syst. Technol. **5**, 47 (2014)
19. Nacos, B.: Politics and the twitter revolution: how tweets influence the relationship between political leaders and the public. Polit. Sci. Q. **128**, 178–179 (2013)
20. Nguyen, D.-P., Gravel, R., Trieschnigg, D., Meder, T.: "How Old Do You Think I Am?" a study of language and age in twitter (2013)
21. Sabidussi, G.: The centrality index of a graph. Psychometrika **31**, 581–606 (1966)
22. Savage, D., Zhang, X., Yu, X., Chou, P., Wang, Q.: Anomaly detection in online social networks. Soc. Netw. **39**, 62–70 (2014)
23. Seidman, S.: Network structure and minimum degree. Soc. Netw. **5**, 269–287 (1983)
24. Tomek, I.: Two modifications of CNN. IEEE Trans. Syst. Man Cybern. **6**, 769–772 (1976)
25. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**, 440–442 (1998)

A Study of How Opinion Sharing Affects Emergency Evacuation

Aravinda Ramakrishnan Srinivasan, Farshad Salimi Naneh Karan,
and Subhadeep Chakraborty(✉)

Department of Mechanical, Aerospace, and Biomedical Engineering,
University of Tennessee, Knoxville, TN, USA
schakrab@utk.edu

Abstract. Several factors like herding amongst evacuees, individual's impatience with crowded pathways, individual's familiarity with the building, and presence of credible leaders play a critical role in an emergency evacuation situation. A thorough understanding of how these factors interplay in a given building structure can lead to increased safety through appropriate evacuation planning. Past works have concentrated their efforts on developing an accurate movement model or an accurate decision theoretic model to study emergency evacuation. This work involves a unified modeling of individuals' movement along with a personalized decision mechanism. Further, to account for herding in the crowd explicitly, a spatially-bounded opinion sharing framework is incorporated. Utilizing this unified model, the interplay of several factors like knowledge about the available exits and presence of leaders with preferred route choice on the evacuation time were investigated in detail. For the given building geometry, we discovered that a crowd consisting of patient individuals with few appropriately informed leaders was able to evacuate the building quicker. Using this unified model, effects of these factors on other building structures can be studied, and it can help with improving the overall safety of the evacuees.

Keywords: Egress dynamics · Hybrid simulation model
Individual decision model–Markov decision process

1 Introduction

Since 1982, there have been at least 81 public mass shootings across the USA, with the killings occurring in 33 states from Massachusetts to Hawaii [1]. In response to this alarming trend, emergency evacuation of buildings have been identified as an important topic of research. Optimization of pedestrian flow can possibly decrease the time spent along non-optimal paths and hence reduce damage related to panic situations.

The existing literature has a rich body of work on modeling pedestrian movement and pedestrian destination choice. The current state of the art can be broadly divided into microscopic, macroscopic and experimental models. In

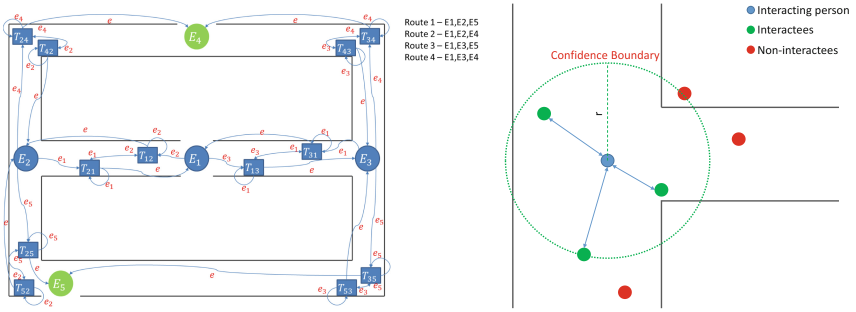
R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 176–182, 2018.
https://doi.org/10.1007/978-3-319-93372-6_19

microscopic modeling, the collective phenomena like bottlenecking, oscillations, etc. are observed from detailed modeling of the dynamics at the microscopic or node level. The microscopic category includes social force model [2], cellular automaton models of pedestrian movement [3], lattice gas method [4], and decision tree based modeling [5]. The macroscopic modeling technique involves describing the flow of pedestrian as analogous to fluid flow and deriving the flow equations necessary to understand and control the crowd movement [6].

In this work, the scope is to study the effect of opinion/information propagation [7] in a crowd of evacuating individuals. Our work, incorporates a sophisticated movement and decision model into a spatially-bounded opinion sharing model to study the effect of knowledge level and presence of leaders in the crowd. The next section provides details about our hybrid model.

2 Simulation Model

The building setup consists of two separate rooms that open up to a common hallway that wrap around to two different exits. The rooms were populated with people from different age and gender groups and were given walking speeds accordingly. The exits were placed such that the building has one shortest path (route 1 in Fig. 1(a)), a couple of paths of equal length (route 2 and 4 in Fig. 1(a)), and a longest path (route 3 in Fig. 1(a)).



(a) (b)

Fig. 1. (a) A finite state automata showing states and actions overlaid on the building layout, and (b) Illustration of interaction with spatially bounded confidence model (Color figure online)

2.1 Decision Model

The underlying decision logic for individuals is modeled as a Markov decision process. A Markov decision process is defined by $M = \{S, A, P, \gamma, R\}$ where:

S is the set of all possible decision states,
A is the set of all available decision/actions,

P is the transition probability $P(s, a, s')$. It gives the probability an individual assigns for successful physical transition to state s' from state s after deciding to take action a,

R is the set of rewards or payoffs assigned to the various decisions by an individual. The individual's overall route choice depends on the reward structure,

γ is the discount factor $\in [0, 1)$ - which make the computation of accumulated rewards mathematically tractable.

Each individual has exits 1, 2, 3, 4, and 5 marked as E_1, E_2, E_3, E_4, E_5, and the trails connecting the exits, marked as T_{ij} (see Fig. 1(a)) as available decision states. T_{ij} denotes the corridor connecting the i^{th} exit to the j^{th} exit. Every individual can decide to move towards one of the immediately available exit points and they will land in the state corresponding to their current position. The set of available actions consist of decisions to move towards exits and the action of exiting labeled as e_1, e_2, e_3, e_4, e_5, and e respectively in Fig. 1(a).

Initially, the transition probability $(P(s, a, s'))$ for all state and action pairs is set at 0.9, and $P(s, a, s) = 1 - P(s, a, s')$ to takes into account the environmental uncertainties. The transition probability for action e $(P(s, e, s'))$ is reduced as time progresses to account for impatience as expressed by, $P(s, e, s') = P(s, e, s') * exp(-\alpha \times t_{diff})$, where $t_{diff} =$ Time spent in state T_{ij} $-$ Estimated travel time to exit E_j. We have experimented with 3 different impatience growth rate, α to simulate different crowd behaviors.

The exits are given decreasing rewards from outward to inward ($E_4,\ E_5 > E_2,\ E_3 > E_1$). The trail state rewards are inversely proportional to the trail length and are upper bounded by the minimum reward for all the exits. Individuals will typically chose the shortest path. However, if the lanes are crowded, then they tend to move towards the next best available route to reach either exit 4 or 5 as quickly as possible. This reward structure enables the decision maker to seek the decision state that leads to the shortest path towards the exit, but the framework allows individuals to change their decision if the are unable to reach their desired exit within a reasonable time frame.

Individuals are assigned a decision timer (τ_i) from a normal random distribution. Each individual performs a planning routine whenever their decision timer expires. For planning their route, individuals compute the value of available states (exits and trails), compare the values, and decide to move along the trail with the highest value. The value of a state is the expected cumulative reward that can be obtained from that state. The discount factor is used in the summation to weigh the immediate reward more than the future rewards. Formally, a value iteration algorithm is used to find the value of states.

The value of states found with value iteration algorithm satisfies the Bellman optimality condition [8]. The Bellman optimality condition states that the action taken at a state has to result in landing at the best possible next state with respect to their calculated value. Thus each individual optimizes his/her route at every decision cycle.

2.2 Opinion Sharing Framework

Humans have a tendency to herd and it is captured in this paper with a spatially bounded confidence model. The bounded confidence model [9] is modified to suit the egress dynamics by using distance between individuals as the confidence boundary metric. Each individual after completing a value iteration cycle will interact with individuals within their herding range (r) and modify their perceived value of states according to $V_{self} = (1 - \mu) \times V_{self} + \mu \times average\ of\ V_{others\ within\ r}$, where μ is the herding level, which is how much weight individuals give to the herd's opinion. The value function is normalized for each individual to ensure that the herding effect is uniform.

An interaction process for an individual (blue) is depicted in Fig. 1(b). The boundary for the interaction/herding zone is shown with the green circle. Agents within the zone and not separated by walls are allowed to share opinion (green).

2.3 Movement Model

The position of the individuals are updated every one second. Each individual will attempt to move towards their respective exit choice. Every individual occupies a circle of one feet radius and additionally one feet radius is designated as personal space. Every individual attempts to move while respecting others' personal space and avoid collision with walls and people. With this hybrid model several scenarios were investigated. The results are presented and discussed in the following section.

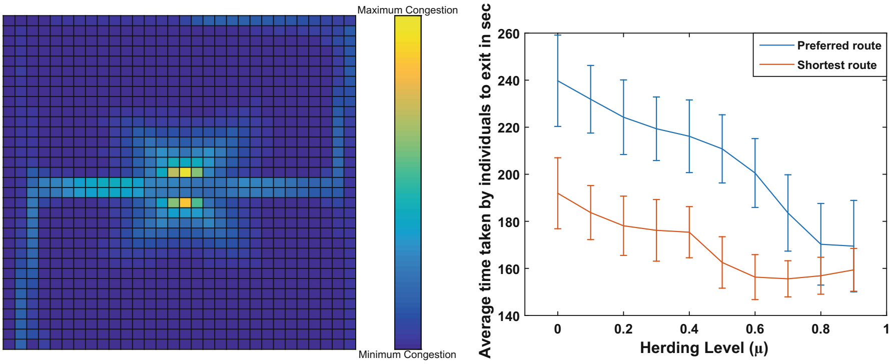
3 Results and Discussion

3.1 Shortest Path Decision Makers

For this set of simulations, each individual's decision model was assigned a reward/reinforcement function which preferred the shortest route. As evident from the congestion map (Fig. 2(a)), route 1 (left, then down) was the most utilized path and route 4 (right, then up) was the second most utilized path. Route 1 was the natural choice for the crowd since it is the shortest path to safety. As every individual tried to go through route 1 it became crowded, impatience grew resulting in part of the crowd starting to move along route 4. The highest congestion occurred at the room exits followed by the corridor just outside the rooms.

3.2 Familiar Path Decision Makers

For this set of simulation, the crowd was initialized with a familiar path reinforcement function. The crowd was randomly and evenly divided into four groups and each group was given a reward function that made one of the four available paths as the familiar route for the individuals in the group. At all herding levels, the shortest path crowd fared better than the familiar path crowd (Fig. 2(b)).



(a) (b)

Fig. 2. (a) Heat map indicating congestion along the routes and (b) Effect of different herding level (μ) on the average time taken by individuals to exit the building with shortest path and familiar path reward function (Common parameters: $N = 300$, $r = 10\ ft$, $\alpha = 0.05$, and $\tau = 4s$)

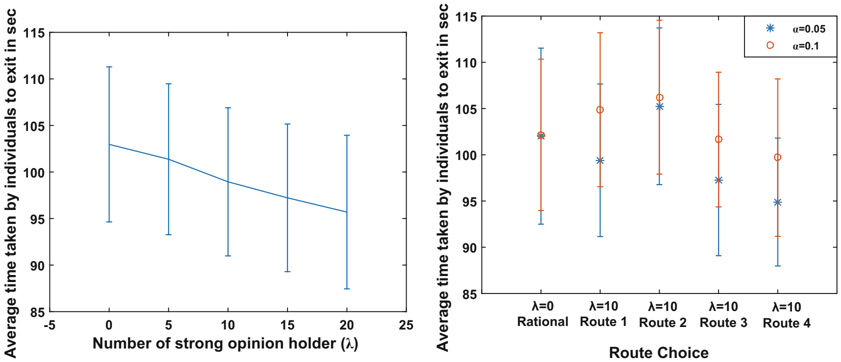
Quicker evacuation was observed when the crowd consisted of more receptive individuals. Cooperation was better when evacuees did not have complete unbiased knowledge of their environment.

3.3 Shortest Path Decision Maker with Familiar Path Leaders

The next set of simulations were conducted to study the effect of leaders with biased route choice on the crowd's egress dynamics. The leaders are characterized by a strong bias and stuck to their opinion (i.e.,) their exit choice is affected only by the environment and not by other individuals. The crowd is composed of a few leaders and many shortest path seeking individuals.

Effect of number of leaders (λ): The results with the specific simulation parameters are shown in Fig. 3(a). The average time to exit the building decreased with more leaders in the crowd. The crowd moved with the leaders and avoided congestion at route 1 and reached safety faster. Route 4 was chosen in particular because it was the second best choice among the available routes taking into account distance to travel and the potential congestion in the corridors.

Effect of route choice of leaders under different impatience levels: The final set of simulations were concerned about the route choice of the leaders. The simulations were conducted with fixed number of leaders ($\lambda = 10$) in a crowd of 110 people. The effect of leaders were diminished (Fig. 3(b)) when a crowd consisted of individuals with faster impatience growth ($\alpha = 0.1$). One logical explanation for this is the fact that in a crowd of highly impatient individuals, leaders ability to sway and hold opinion of other individuals for long time is diminished. With a lesser impatient crowd ($\alpha = 0.05$), except for route 2 which puts additional pressure on already crowded lane all other leaders route bias were helpful in getting the crowd to safety quicker.



(a) (b)

Fig. 3. (a) $\alpha = 0.05$, $N = 120$, and leader with route choice 4 - Effect of number of strong opinion holders on the average time to evacuate, and (b) Number of strong opinion holders, $\lambda = 10$ - Effect of different route choice of leaders on the average time taken by individuals to exit the building (Common parameters: $\tau = 4\,\text{s}$, $r = 10\,\text{ft}$, and $\mu = 0.4$)

4 Conclusion and Future Work

This work combines a naturalistic movement model and a decision making model with an explicit opinion sharing dynamics (the hybrid model) to study the effect of opinion sharing and several other factors on the crowd evacuation metrics for a given building structure. Factors such as how receptive the crowd is to opinion sharing, how fast the individuals tend to change their exit choice when confronted with crowded lanes/congestion, and the frequency of decision making affect the crowd's evacuation time. Ideally, a communicative crowd seeking shortest path with well informed leaders is well-suited for a quick evacuation of the given building. Herding is not detrimental for evacuation. However, over-herding can lead to under utilization of all the available routes leading to an increase in the evacuation time. People with strong opinions can contribute to faster egress out of the building, if their strong opinion aligns with the under-utilized route(s).

References

1. Us mass shootings, 1982–2017: Data from mother jones investigation, June 2016 http://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data
2. Helbing, D., Molnár, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282–4286 (1995)
3. Alizadeh, R.: A dynamic cellular automaton model for evacuation process with obstacles. Saf. Sci. **49**(2), 315–323 (2011)
4. Guo, R.Y., Huang, H.J.: A mobile lattice gas model for simulating pedestrian evacuation. Phys. A Stat. Mech. Appl. **387**(2–3), 580–586 (2008)

5. Pan, X., Han, C.S., Dauber, K., Law, K.H.: A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. Ai Soc. **22**(2), 113–132 (2007)
6. Hughes, R.L.: The flow of large crowds of pedestrians. Math. Comput. Simul. **53**(4–6), 367–370 (2000)
7. Karan, F.S.N., Chakraborty, S.: Dynamics of a repulsive voter model. IEEE Trans. Comput. Soc. Syst. **3**(1), 13–22 (2016)
8. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, Hoboken (2014)
9. Karan, F.S.N., Chakraborty, S.: Effect of zealots on the opinion dynamics of rational agents with bounded confidence. Acta Phys. Pol. B **49**(1), 73 (2018)

Fine-Scale Prediction of People's Home Location Using Social Media Footprints

Hamdi Kavak[1](✉), Daniele Vernon-Bido[2], and Jose J. Padilla[2]

[1] George Mason University, Fairfax, VA 22030, USA
hkavak@gmu.edu
[2] Virginia Modeling Analysis and Simulation Center,
Suffolk, VA 23435, USA
http://www.hamdikavak.com

Abstract. In this study, we develop a machine learning classifier that determines Twitter users' home location with 100 m resolution. Our results suggest up to 0.87 overall accuracy in predicting home location for the City of Chicago. We explore the influence of *time span of data collection* and *location-sharing habits of a user*. The classifier accuracy changes by data collection time but larger than one-month time spans do not significantly increase prediction accuracy. An individual's home location can be ascertained with as few as 0.6 to 1.4 tweets/day or 75 to 225 tweets with an accuracy of over 0.8. Our results shed light on how home location information can be predicted with high accuracy and how long data needs to be collected. On the flip side, our results imply potential privacy issues on publicly available social media data.

Keywords: Human mobility · Social media · Home location inference

1 Introduction

The availability of large-scale behavioral data allows researchers to scratch the surface of human behavior understanding and prediction [2]. The spatial movement of people provides a starting point as it has the potential to reveal the relationship between places, activities, and social interactions that make up one's daily life. In this respect, home is one of the most important hubs for people when transitioning from one daily activity to another [8]. However, predicting 'home' is challenging due to the sparsity of geo-tagged social media footprints.

There is a wide range of methods to infer people's home location from their social media content [4–7]. In this study, we develop a machine learning classifier that used geo-tagged tweets for home location prediction following and extending Hu et al. [4]'s work. We start building our classifier with five human mobility features identified to be important in [4]. We advance Hu et al. [4]' s work by (1) adding two mobility features (land use patterns and distance from most checked-in location), (2) constructing place visit history through clustering, and

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 183–189, 2018.
https://doi.org/10.1007/978-3-319-93372-6_20

(3) exploring the effect of data collection length, tweeting rate, and the number of tweets on classifier accuracy. Our study considers 100 m resolution and outperforms in applicable scope (100%) and accuracy (up to 0.87) of any previous studies. We also report data collection requirements necessary for home location prediction problem.

In Sect. 2, we describe the dataset used in this study and report the preparation process we follow. In Sect. 3, we present our home location prediction classifier and explain how we train and evaluate its performance. We then report our results in Sect. 4 and investigate several factors that are influential on prediction accuracy. Finally, we conclude the paper in Sect. 5.

2 Dataset Preparation

We perform three steps in preparing social media data for home location prediction. First, we collect data from Twitter and identify a subset suitable for the study. Next, we create a *ground-truth* dataset that contains tweets known to be sent from users' home. Finally, we clean the *ground-truth* dataset and cluster it to identify unique locations at the individual level.

We use Twitter's Streaming API to collect public tweets with exact Global Positioning System (GPS) locations. Tweet collection is performed between May 16, 2014 and April 27, 2015. We choose the city of Chicago, Illinois because it is one of the significant metropolitan areas in the United States. We focus on active users with at least five geo-tagged tweets [4]. The active users' dataset contains ≈7.78 million location footprints from 92,296 Twitter users.

We create a dataset with a portion of active users whose home locations are known with confidence. We follow a process similar to the one in [4] that relies on crowdsourcing to identify whether tweets that contain home-related keywords[1] are sent from home or not. We develop a web application for labeling these home-related tweets. The web application simply displays a tweet from the dataset and asks the user (crowdsource) to choose a label: *from home*, *not from home*, or *unsure*. Each question is displayed up to three times randomly. Precedence is given to the ones that already have an answer. We received 14,076 responses for 4,679 questions. We only focused on tweets with an agreement in all three that the user is sending the message from home. Approximately 38% of tweets (1,797) from 1,268 users satisfy this criterion.

Lastly, pre-processing consists of two steps: *cleaning* and *clustering*. Cleaning prevents biasing the dataset with a significant number of tweets from the same location in a short time interval. We clean tweets that are consecutively shared in less than sixty minutes and within 100 m distance. At the end of the cleaning, 62.2% of messages were removed. Clustering uses the cleaned tweets and identifies tweets sent from same places. This is important because GPS data usually has some inaccuracy even when shared from the same location. We cluster these points by giving them the same location ID label using the DBSCAN algorithm

[1] *home* and at least one of the following keywords: *shower*, *sofa*, *TV*, *sleep*, *nap*, *bed*, *alone*, *watch*, *night*, *sweet*, *stay*, *finally*, *tonight*, *arrived*.

[3] with 100-m as the maximum distance parameter and one as the minimum number of points in a cluster. Pre-processing generated a dataset containing 462,409 location footprints with the following properties: *anonymized user ID*, *local date-time*, *location (latitude-longitude)*, *cluster label ID*, and *is home*.

3 Home Location Prediction

We consider home location prediction a process that takes a set of location footprint history of a user and predicts the footprints that are shared from home. To this end, we create a method that receives a location footprint set of a user and generates a mobility feature set (X) that contains one record *per unique cluster label ID*. This mobility feature set has following features where the first six are identified in [4] and the last two are proposed by the authors (see online supplemental for detailed feature descriptions and value distributions).

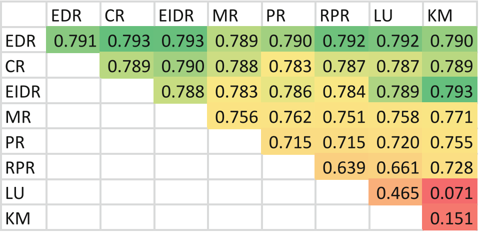
- Check-in Ratio (CR)
- Check-in Ratio during Midnight (MR)
- Check-in Ratio of Last Destination of a Day (EDR)
- Check-in Ratio of Last Destination of a Day with Inactive Midnight (EIDR)
- PageRank (PR)
- Reverse PageRank (RPR)
- Land Use Pattern (LU)
- Kilometer Distance from Most Checked-in Location (KM)

We use Support Vector Machines (SVM) with linear kernel as our classifier because it is a robust approach for binary classification problems and has been successfully implemented in the home location prediction problem [4]. SVM works by creating an optimally placed decision boundary (hyperplane) to separate elements of classes with maximum margin [1]. It is computationally costly to train an SVM classifier due to the involvement of numerical optimizations, but it is computationally efficient to use as a trained classifier. For instance, when a linear SVM classifier is trained, the classification problem turns into a simple calculation of $c = W * X + b$. When c is non-negative, it denotes one class and when it is negative, it denotes the other class given that W is the weight vector for the hyperplane, b is the intercept parameter, and X is the input whose class is investigated.

To train and test the classifier, we apply repeated 5-fold cross-validation. That is, we split the ground-truth users in five equal groups, train the classifier with four groups and test it with the remaining one, and repeat until all groups are used in training four times and in testing one time. We also repeat this procedure five times with randomly shuffling the user list in each time. In total, each evaluation takes 25 runs. In each fold, we predict home location at the user-level by calculating the SVM score for each unique location label ID and pick the one with the highest score. In the end, we capture average accuracy.

4 Results

We first report the accuracy of each mobility feature separately and with their combinations shown in Fig. 1. As a single feature, *End of Day Ratio* (EDR) has the highest accuracy with 0.791 while general *Check-in Ratio* (CR) and *End of Inactive Day Ratio* (EIDR) are marginally lower. *Midnight Ratio* (MR) is slightly lower with 0.756 followed by *PageRank* (PR) and *Reverse PageRank* (RPR) scores 0.715 and 0.639. Finally, *Land Use* (LU) feature has the accuracy score of 0.465 and *Kilometer Distance to Most Visited Location* (KM) feature performs the worst with the accuracy score of 0.151.



	EDR	CR	EIDR	MR	PR	RPR	LU	KM
EDR	0.791	0.793	0.793	0.789	0.790	0.792	0.792	0.790
CR		0.789	0.790	0.788	0.783	0.787	0.787	0.789
EIDR			0.788	0.783	0.786	0.784	0.789	0.793
MR				0.756	0.762	0.751	0.758	0.771
PR					0.715	0.715	0.720	0.755
RPR						0.639	0.661	0.728
LU							0.465	0.071
KM								0.151

Fig. 1. Accuracy scores of single features (diagonal) and all combinations of two features. The color intensity is given based on values. (Color figure online)

Pairing the features provides marginal improvements to the best performing single features; however, the lowest performing feature, KM, when combined with PR and RPR improved the accuracy by 8 to 14%. The combined best score, 0.793, outperforms the best-reported score in the literature [4] in that the accuracy is the same but the scope of applicability is greatly improved. Hu et al. [4] have an applicability of 30–40% while our classifier covers 100%. To check the robustness of this result, we examine the change of accuracy based on data collection length and number of footprints.

Figure 2 shows that accuracy scores are not linear with respect to data collection length. The top performing single and combined feature scores reach their maximum accuracy in 14 days (except for PR which peaks at 21 days). For single features, there is a slight difference in the rankings of the top three features although they are still very close. For combined features, their accuracy appears to be a bit better than the single features especially when data collection length is 30 days or lower.

In addition to the data collection length, we investigate the number of tweets per user and the user tweeting rate. The number of tweets per user (G_n) captures the direct relationship between footprint size and classifier accuracy. We define a measurement - *tweeting rate* - to standardize the unit over different tweeting habits of users (G_r). Tweeting rate (Eq. 1) is given as the total number of tweets of a user divided by the number of days between the first and last tweet.

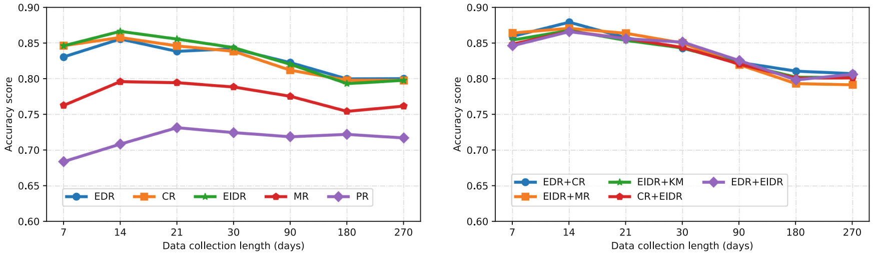


Fig. 2. Accuracy score based on data collection length for single best performing features (left) and combined best performing features (right).

$$G_r = \frac{number\,of\,tweets}{number\,of\,days\,between\,first\,and\,last\,tweet} \qquad (1)$$

We create four groups of users for each of the two measures keeping a similar number of users in each group (see supplemental). Figure 3 shows the average accuracy for the top five performing single and combined features. For users with lower tweeting rate, the classifier's accuracy averages 0.6 and 0.7 respectively for single features and combined features. Additionally, PR and RPR perform poorly on this group because of an insufficient number of check-ins.

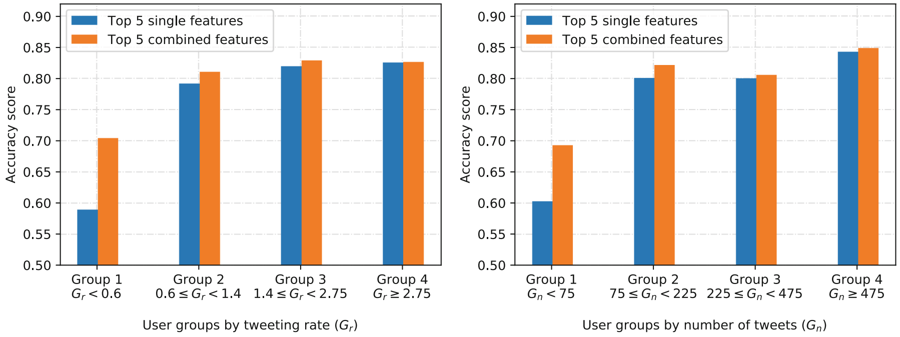


Fig. 3. The change of accuracy by four groups of users gathered based on tweeting rate (left) and number of tweets (right).

Groups 2 through 4 demonstrate no significant difference in accuracy. The higher tweet rates do not provide sufficient benefit to the classifier's accuracy rating to suggest that the increase is relevant. As such, we consider the Group 2 users to have the optimal tweet rate for predicting home location. We conclude that with our classifier, 0.6 to 1.4 daily geo-tagged tweet activity or 75 to 225 total tweets per individual allow for the predication of the user's home location with over 0.8 accuracy.

5 Conclusion

In this paper, we addressed the fine-scale (100-m) prediction problem of Twitter users' home locations. We developed an SVM classifier with several mobility features including check-in ratios at locations, graph-based features, and distance between locations. We then trained this classifier with geo-tagged Twitter data from the City of Chicago and explored the accuracy of home location prediction under different conditions. The best accuracy for the entire dataset was 0.795. When considering several subsets of the dataset, we gathered empirical insights that were not clearly present in the entire dataset. For instance, we found that a high number of tweets and high tweeting activity did not significantly increase prediction accuracy. In fact, 0.6 to 1.4 daily tweeting rate or 75 to 225 number of tweets is enough to perform over 0.8 accuracy. Lower numbers, on the other hand, reached 0.71 accuracy which is still very high given that our classifier covers all the instances in the applicable scope whereas the previous study by [4] only applies to 71–76% of instances for achieving a similar accuracy.

Notes: Additional information, code, and datasets of this manuscript are freely available at https://github.com/hamdikavak/home-location-prediction. We thank our colleagues at Old Dominion University who helped labeling our dataset.

Acknowledgments and Disclaimer. This material is based on research sponsored by the Office of the Assistant Secretary of Defense for Research and Engineering (OASD(R&E)) under agreement number FAB750-15-2-0120. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the OASD(R&E) or the U.S. Government.

References

1. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
2. Eagle, N., Pentland, A.S.: Reality mining: sensing complex social systems. Pers. Ubiquit. Comput. **10**(4), 255–268 (2006)
3. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd KDD. AAAI Press (1996)
4. Hu, T., Luo, J., Kautz, H., Sadilek, A.: Home location inference from sparse and noisy data: models and applications. In: Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, pp. 1382–1387 (2016)
5. Mahmud, J., Nichols, J., Drews, C.: Where is this tweet from? Inferring home locations of Twitter users. In: ICWSM, pp. 511–514 (2012)
6. Pontes, T., Magno, G., Vasconcelos, M., Gupta, A., Almeida, J., Kumaraguru, P., Almeida, V.: Beware of what you share: inferring home location in social networks. In: ICDMW 2012, pp. 571–578 (2012)

7. Ryoo, K., Moon, S.: Inferring Twitter user locations with 10 km accuracy. In: Proceedings of the 23rd International Conference on WWW (2014)
8. Schneider, C.M., Belik, V., Couronné, T., Smoreda, Z., González, M.C.: Unravelling daily human mobility motifs. J. R. Soc. Interface **10**(84), 20130246 (2013)

Formal Organizations, Informal Networks, and Work Flow: An Agent-Based Model

Thomas W. Briggs(✉)

George Mason University, Fairfax, VA 22030, USA
tbriggs@gmu.edu

Abstract. Few computational network models contrasting formal organization and informal networks have been published. A generalized organizational agent-based model (ABM) containing both formal organizational hierarchy and informal social networks was developed to simulate organizational processes that occur over both formal network ties and informal networks. Preliminary results from the current effort demonstrate "traffic jams" of work at the problematic middle manager level, which varies with the degree of micromanagement culture and supervisory span of control. Results also indicate that some informal network ties are used reciprocally while others are practically unidirectional.

Keywords: Organizations · Networks · ABM · Boundary spanning

1 Introduction

Organizational stakeholders often articulate the importance of informal networks: "it's not what you know, it's who you know" is a truism, and managers routinely use their own networks to accomplish goals and get work done [1]. Yet informal networks are seldom studied in organizations and are often erroneously presumed to be comprehensively known and understood by managers [2]. Companies rarely undertake network analysis prior to organizational actions, sometimes with disastrous consequences.

Network scientists who publish case studies in magazines and journals advocate a network approach to managerial decision making. For example, Cross et al. [3] tell a cautionary tale of an organization's office-space redesign gone wrong due to a failure to account for individuals' positions as important nodes in informal networks. Cross et al. [3] highlight successes at organizations such as the U.S. Defense Intelligence Agency (DIA) where organizational network analysis led to measurable, beneficial organizational outcomes. Computational modeling and simulation offer an abstracted and generalized methodology to study and quantify processes and outcomes that can occur as an organization's employees interact with each other to disseminate information, collaborate, or make individual decisions. Individual decisions have real organizational implications, like the choice to retire based on what others in one's network are doing [4]. The purpose of this effort is to develop an agent-based model (ABM) to simulate work and information flow over both formal, hierarchical networks and informal networks that cross formal organizational boundaries.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 190–196, 2018.
https://doi.org/10.1007/978-3-319-93372-6_21

Early network studies of organizations include the frequently-cited Allen and Cohen [5] study of information flow in research and development laboratories, which compared formal organizations and informal networks, and concluded that work-related technical communication resulted from both social relations (i.e., informal networks) and work structure (i.e., formal organization). The authors also observed differences in communication by status (i.e., individuals with PhDs and without PhDs) and differences related to individuals' position in the network as "sociometric stars," which afforded them the opportunity to serve as gatekeepers of information [5].

Katz and Tushman [6], also studying R&D laboratories, found differing patterns of information flow based on whether a project was focused on research, development, or service, with research projects generating significantly more intraproject and R&D laboratory communication, while service projects generated significantly more intraorganizational communication throughout the organization. As in [5], the authors also found a specialized role for certain individuals—boundary spanners—who served as informational interfaces between internal organizational stakeholders and external stakeholders such as customers/suppliers, other professionals, and consultants [6].

Social network field studies of organizations and the people in them present specific research methodology challenges [7] and executives may be reticent to release information on the internal workings of their organizations since doing so could potentially give competitors an upper hand or put the organization at legal risk. Perhaps the most well-known social network dataset in recent years that details information flow in an actual organization is the public release of the emails of 158 employees of the Enron Corporation in 2002 following the federal inquiries after Enron's demise. Diesner et al. [8], in a study using the Enron email data, praise the email corpus as being, "alluring and of particular interest with much academic value…a rare, authentic glimpse into the social network of an actual business organization" (p. 202). The authors enhanced the Enron email dataset by adding previously unknown names and producing much higher rates of email-to-individual mapping before they extracted social network data. In addition to enhancing the dataset, the authors found that the flow of information between employees diversified with respect to formal roles as the Enron crisis intensified, previously disconnected employees began communicating, and formal chains of communication were bypassed [8].

2 Computational Network Models of Organizations

A few scholars have modeled information flow in organizations using formal, mathematical models and agent-based models. Ben-Arieh and Pollatscheck [9] developed a model using dynamic programming linking hierarchical organizational productivity to information processing, finding that "information overload" caused declines not just in individual productivity, but also in overall organizational productivity. By running sensitivity analyses to optimize the parameters of information compression and information expansion across three hierarchical organization types—homogenous, semi-homogeneous, and non-homogeneous—they concluded that the higher the cost of information processing, the lower the amount of information should flow [9]. To validate the model, a pilot study was conducted in a real-world "high-technology communication company" which tracked information flow between levels, finding that information mostly flowed downward with only 28% of information flowing upward from middle management to the top. The authors noted interest in their model from the military intelligence community as an additional indication of validity, classifying that community solidly as an organization with homogeneous information flow [9].

Using agent-based modeling and dynamic network analysis, Lin and DeSouza [10] took a different approach to exploring information transfer in organizations; essentially, a "bottom-up" approach in which agents formed ties based on individual utility maximization, leading to the emergence of informal social networks. Highly knowledgeable individuals had a tendency to have fewer network connections, possibly due to the high cost of being a constant source of information to others [10]. High knowledge diversity in an organization led to good reachability in the informal networks that emerged, and when knowledge is diverse or becomes obsolete fast, interpersonal, relationship-based knowledge transfer is less effective at improving the average knowledge level in the organization [10].

Tsvetovat and Carley [11] constructed a multi-agent, network model of organizations —in this case, covert networks representing terrorist cells—on the premise that complex socio-technical systems like organizations can be modeled only by combining social networks and cognitively-plausible agents acting independently. The agents are bounded in their rationality and their information about the world is limited by their ability to perceive. More simply, an individual agent only knows what it knows from its own small corner of the world: it knows only the other agents in its ego network, and knows only about assigned tasks and resources, though it will attempt to form beliefs about other agents and what they know. Communication occurs as a function of social proximity, homophily, and need, and agents exchange knowledge and learn about other agents to execute complex tasks that require coordination and delegation between agents [11]. To prove the viability of the approach, a terrorist cell network was generated from known, empirical network statistics and a corresponding anti-terrorist team was tasked with discovering and then exploiting knowledge about the terrorist agent activities in order to successfully disrupt the network [11].

These network studies of organizations have demonstrated the important role of informal networks and the particular importance of boundary spanners who connect different teams or parts of the organization though they have no direct linkage in the formal, organizational hierarchy. Few computational network models contrasting formal organization and informal networks have been attempted, and those that have are often calibrated to a particular case study (e.g., Enron). A generalized computational organization model containing both formal organization hierarchy and informal social networks, including boundary spanning, is needed to simulate organizational processes that can happen over both formal network ties or informal networks, permitting the precise quantitative measure of when each (simulated) network tie is used.

3 Methods

Developed in NetLogo [12], a simulated organization is populated with employee agents at multiple levels (e.g., CEO, managers, workers) and at each time step agents carry out work tasks that require them to interact with other agents to receive and complete projects, similar in spirit to the manner in which Tsvetovat and Carley's [11] agents used their ego networks to gain access to needed knowledge.

The simulation of informal networks in the current study uses a simple Erdős–Rényi random network mechanism [13, 14], where each pair of nodes is connected with some probability P (typically set at 0.01). While alternative network mechanisms (e.g., Watts-Strogatz small world) may better represent the nonrandom nature of informal organizational networks, the goal of this simulation was to examine a network topology that contains both formal, hierarchical links, along with some number of informal network links such that agents are connected by more network ties than just those shown on the organization chart. With the existence of the informal network, employees can leverage informal ties to reach across organizational boundaries to complete their work.

Outcome measures of interest included task performance (i.e., completion, efficiency) and measures of information flow and work movement. The percentage of information flow occurring through the links that define the formal organization is also compared with the percentage flowing through informal network links. If work gets "stuck" in the hierarchy, for example, can agents work around the blockage?

Calibration data were sourced from [9] and Table 1 lists the model parameters and sample values used in this simulation. Empirical data like the [9] pilot study on the amount of information sent and received by organizational level permits one to adjust the "bottleneck" parameter B such that some agents' workloads—for example, those of middle managers—might become overloaded as the model runs, creating further delays in the transmission of work or information to subordinates and preventing work from flowing back up through those middle managers. Acknowledging the fact that a disproportionate amount of information flows downward and horizontally, as opposed to upward, the upward transmission constraint *U* determines the likelihood/possibility that an agent can go "over the head" of a middle manager to the next level up.

Table 1. Parameters for org networks agent-based model.

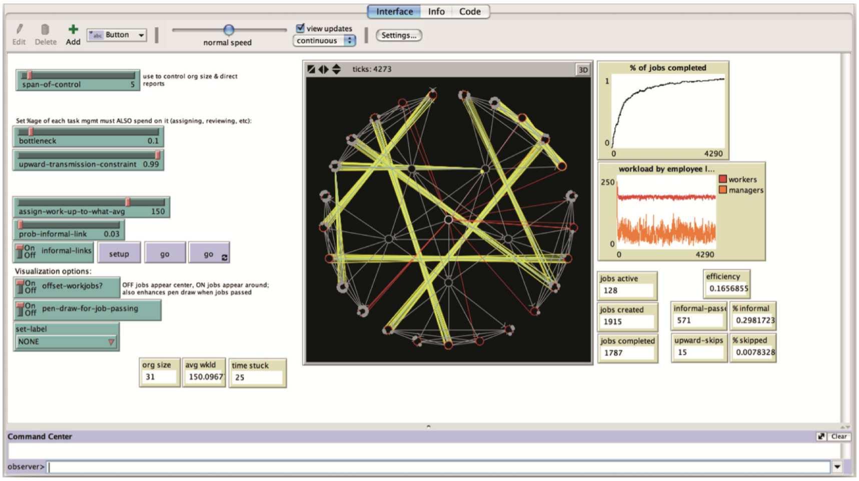
Parameter	Description	Sample values
S	Supervisory span of control	5, 10
P	Probability of rewiring (informal network)	0.01, 0.03
B	Bottleneck (i.e., micromanagement)	0.1, 0.5
U	Upward constraint on information transmission	0.90, 0.99

4 Results

Several hundred model runs were conducted to compare conditions – notably, the presence or absence of the informal network. The primary finding is the strong effect of the bottleneck parameter on organizational efficiency. The Bottleneck parameter (i.e., time managers spend on work as it transits down and up) interacts with span of control: with high bottleneck and high span of control, managers simply have too much work to do and tasks essentially get stuck at the manager, decreasing the efficiency with which employees receive new work. Work completed by employees is also approved more slowly. When middle managers are overwhelmed, employees more frequently utilize the option to "skip" their manager and go directly to the CEO, if allowed. However, when the bottleneck parameter is low and the organization is efficient – even with a moderately high span of control of 10 – employees rarely skipped their manager.

Efficiency, a metric calculated as the ratio of work units to units of time expended, was not improved by the addition of the informal network, though this is likely an artifact of the model's current implementation. Currently, the CEO has perfect information on the entire organization's workload and continues assigning work to maintain the workforce utilization level. In reality, the CEO's imperfect information would decrease efficiency, and this modification is planned for future model development.

Qualitative findings were explored using real-time model visualization. Figure 1 displays the GUI of the model showing a single model run after 11 (simulated) years, displaying use of informal network ties (yellow) to pass work. In Fig. 1, the worker near the 8 o'clock position has passed many jobs to a friend at the 10 o'clock position, but the cluster's triangular shape indicates that the 10 o'clock agent has not made the same use of the informal network tie. However, the 10 o'clock agent has an informal connection to the manager of the team in the northeast quadrant and has passed a substantial number of jobs to that manager to then be funneled to the CEO.



span-of-control 5
use to control org size & direct reports

Set %age of each task mgmt must ALSO spend on it (assigning, reviewing, etc):

bottleneck 0.1

upward-transmission-constraint 0.99

assign-work-up-to-what-avg 150

prob-informal-link 0.03

On Off informal-links

Visualization options:

On Off offset-workjobs?
OFF jobs appear center, ON jobs appear around; also enhances pen draw when jobs passed

On Off pen-draw-for-job-passing

set-label NONE

org size 31

avg wkld 150.0967

time stuck 25

jobs active 128

jobs created 1915

jobs completed 1787

efficiency 0.1656855

informal-passe 571

% informal 0.2981723

upward-skips 15

% skipped 0.0078328

Fig. 1. Org Network Model. S = 5, P = 0.03, B = 0.1, U = 0.99 after 11 simulated years

5 Conclusions and Discussion

The primary goal of the current effort was to generate an organizational network topology that considers both formal organization (hierarchy) and the informal networks that emerge in work organizations and to explore work flow over these co-occurring networks. A second goal was to lay the groundwork for future modelers to simulate "organizational life" on co-occurring networks. Such models permit studying how the process of information flow is affected by network characteristics as individuals participate in groups and in the larger organization.

Preliminary results from the current effort demonstrate "traffic jams" of work at the problematic middle manager level. When this occurs, employees end up with very little work to do because their manager is too busy processing work coming from both directions (up and down) to assign new tasks. This result depends greatly on the organization's "bottleneck" parameter. However, if employees are appropriately empowered, middle managers spend minimal time preprocessing or post-processing work. In a more risk-averse organization, middle managers spend substantial time micromanaging the work, decreasing efficiency and increasing employees' need to use their informal networks just to get their jobs done. Observing the visual path and frequency at which work projects travel over the informal network illustrates that network ties are not always utilized reciprocally – the network tie can appear to be almost unidirectional, though it's unlikely such a one-sided exchange would occur in real organizational life.

Limitations. Deliberate choices kept the model parsimonious. Agents did not differ in performance, which does not accurately represent the true distribution of performance. Introducing variability in performance may alter the dynamics of the entire organization; doing good work often begets more work, so if a given team happened to be "high performing," perhaps that team might serve as a conduit to pick up the backlog from lower-performing teams. The truly dynamic nature of organizational composition was not represented in the current model, but future work could simulate organizational growth, turnover, and personnel movement. An annual growth or contraction parameter, G, could control the addition of new agents/nodes or the replacement rate of agents who turn over. Networks should also be treated and implemented as dynamic [15].

Many questions remain and new questions have been raised in the current effort. What degree of connectivity, for example, is enough or is ideal to enable employees to work around middle-management bottlenecks or vacant managers, but not so much that the cost of maintaining network ties exceeds the value to be gained? What is an optimal network architecture to balance complexity with efficiency and performance [16]?

The current study demonstrates that a relatively simple ABM can be employed by researchers to simulate organizational life and to explore the use of various network connections that bridge or span network groups. The model can precisely quantify which work tasks were passed along which network ties at what time and under what conditions, providing rich and detailed data not easily gathered from real organizations. This initial examination of how informal networks are used when the formal organization hierarchy is unavailable demonstrates both the power and importance of informal networks. If the informal network did not exist, organizational efficiency would be negatively impacted and, at least in some places, work might nearly grind to a halt.

Acknowledgements. The author is grateful to Robert Axtell for many discussions of models of organizational life and to two anonymous peer reviewers for their helpful feedback and thoughtful suggestions on a draft of this paper.

References

1. Cross, R., Prusak, L.: The people who make organizations go–or stop. Harv. Bus. Rev. **80**, 104–112 (2002)
2. Krackhardt, D., Hanson, J.R.: Informal networks: the company behind the charts. Harv. Bus. Rev. **71**, 104–111 (1993)
3. Cross, R., Parise, S., Weiss, L.M.: The role of networks in organizational change. McKinsey Q. 3, 28–41 (2007)
4. Axtell, R.L., Epstein, J.M.: Coordination in transient social networks: an agent-based computational model of the timing of retirement. In: Epstein, J.M. (ed.) Generative Social Science: Studies in Agent-Based Computational Modeling. Princeton University Press (2006)
5. Allen, T.J., Cohen, S.I.: Information flow in research and development laboratories. Adm. Sci. Q. **14**, 12–19 (1969)
6. Katz, R., Tushman, M.: Communication patterns, project performance, and task characteristics: an empirical evaluation and integration in an R&D setting. Organ. Behav. Hum. Perform. **23**, 139–162 (1979)
7. Borgatti, S.P., Everett, M.G., Johnson, J.C.: Analyzing Social Networks. Sage, Los Angeles (2013)
8. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the Enron email corpus "it's always about the people. Enron is no different". Comput. Math. Organ. Theory **11**, 201–228 (2006)
9. Ben-Arieh, D., Pollatscheck, M.A.: Analysis of information flow in hierarchical organizations. Int. J. Prod. Res. **40**, 3561–3573 (2002)
10. Lin, Y., Desouza, K.C.: Co-evolution of organizational network and individual behavior: an agent-based model of interpersonal knowledge transfer. In: ICIS, p. 153 (2010)
11. Tsvetovat, M., Carley, K.M.: Modeling complex socio-technical systems using multi-agent simulation methods. Kuenstliche Intell. **2004**, 23–28 (2004)
12. Wilensky, U.: NetLogo. Northwestern University, Evanston, IL (1999)
13. Erdös, P., Rényi, A.: On random graphs. I. Publ. Math. Debr. **6**, 290–297 (1959)
14. Adamic, L.: Erdös-Renyi Degree Distribution NetLogo Model (2012)
15. Carley, K.M.: Dynamic network analysis. In: Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. National Academies Press, Washington, D.C. (2003)
16. Broniatowski, D.A., Moses, J.: Measuring flexibility, descriptive complexity, and rework potential in generic system architectures. Syst. Eng. **19**, 207–221 (2016)

Aspect Level Sentiment Classification with Attention-over-Attention Neural Networks

Binxuan Huang(✉), Yanglan Ou, and Kathleen M. Carley

Carnegie Mellon University, 5000 Forbe Avenue, Pittsburgh, USA
{binxuanh,kathleen.carley}@cs.cmu.edu, yanglano@andrew.cmu.edu

Abstract. Aspect-level sentiment classification aims to identify the sentiment expressed towards some aspects given context sentences. In this paper, we introduce an attention-over-attention (AOA) neural network for aspect level sentiment classification. Our approach models aspects and sentences in a joint way and explicitly captures the interaction between aspects and context sentences. With the AOA module, our model jointly learns the representations for aspects and sentences, and automatically focuses on the important parts in sentences. Our experiments on laptop and restaurant datasets demonstrate our approach outperforms previous LSTM-based architectures.

1 Introduction

Unlike document level sentiment classification task [4,16], aspect level sentiment classification is a more fine-grained classification task. It aims at identifying the sentiment polarity (e.g. positive, negative, neutral) of one specific aspect in its context sentence. For example, given a sentence "great food but the service was dreadful" the sentiment polarity for aspects "food" and "service" are positive and negative respectively.

Aspect sentiment classification overcomes one limitation of document level sentiment classification when multiple aspects appear in one sentence. In our previous example, there are two aspects and the general sentiment of the whole sentence is mixed with positive and negative polarity. If we ignore the aspect information, it is hard to determine the polarity for a specified target. Such error commonly exists in the general sentiment classification tasks. In one recent work, Jiang et al. manually evaluated a Twitter sentiment classifier and showed that 40% of sentiment classification errors are because of not considering targets [7].

Many methods have been proposed to deal with aspect level sentiment classification. The typical way is to build a machine learning classifier by supervised training. Among these machine learning-based approaches, there are mainly two different types. One is to build a classifier based on manually created features [7,27]. The other type is based on neural networks using end-to-end training without any prior knowledge [12,26,29]. Because of its capacity of learning representations from data without feature engineering, neural networks are becoming popular in this task.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 197–206, 2018.
https://doi.org/10.1007/978-3-319-93372-6_22

Because of advantages of neural networks, we approach this aspect level sentiment classification problem based on long short-term memory (LSTM) neural networks. Previous LSTM-based methods mainly focus on modeling texts separately [24,29], while our approach models aspects and texts simultaneously using LSTMs. Furthermore, the target representation and text representation generated from LSTMs interact with each other by an attention-over-attention (AOA) module [2]. AOA automatically generates mutual attentions not only from aspect-to-text but also text-to-aspect. This is inspired by the observation that only few words in a sentence contribute to the sentiment towards an aspect. Many times, those sentiment bearing words are highly correlated with the aspects. In our previous example, there are two aspects "appetizers" and "service" in the sentence "the appetizers are ok, but the service is slow." Based on our language experience, we know that the negative word "slow" is more likely to describe "service" but not the "appetizers". Similarly, for an aspect phrase, we also need to focus on the most important part. That is why we choose AOA to attend to the most important parts in both aspect and sentence. Compared to previous methods, our model performs better on the laptop and restaurant datasets from SemEval 2014 [18]

2 Related Work

Sentiment Classification

Sentiment classification aims at detecting the sentiment polarity for text. There are various approaches proposed for this research question [13]. Most existing works use machine learning algorithms to classify texts in a supervision fashion. Algorithms like Naive Bayes and Support Vector Machine (SVM) are widely used in this problem [11,16,28]. The majority of these approaches either rely on n-gram features or manually designed features. Multiple sentiment lexicons are built for this purpose [15,19,23].

In the recent years, sentiment classification has been advanced by neural networks significantly. Neural network based approaches automatically learn feature representations and do not require intensive feature engineering. Researchers proposed a variety of neural network architectures. Classical methods include Convolutional Neural Networks [6,8], Recurrent Neural Networks [10,25], Recursive Neural Networks [20,30]. These approaches have achieved promising results on sentiment analysis.

Aspect Level Sentiment Classification

Aspect level sentiment classification is a branch of sentiment classification, the goal of which is to identify the sentiment polarity of one specific aspect in a sentence. Some early works designed several rule based models for aspect level sentiment classification, such as [3,14]. Nasukawa et al. first perform dependency parsing on sentences, then they use predefined rules to determine the sentiment about aspects [14]. Jiang et al. improve the target-dependent sentiment classification by creating several target-dependent features based on the sentences' grammar structures [7]. These target-dependent features are further fed into an SVM classifier along with other content features.

Later, kinds of neural network based methods were introduced to solve this aspect level sentiment classification problem. Typical methods are based on LSTM neural networks. TD-LSTM approaches this problem by developing two LSTM networks to model the left and right contexts for an aspect target [24]. This method uses the last hidden states of these two LSTMs for predicting the sentiment. In order to better capture the important part in a sentence, Wang et al. use an aspect term embedding to generate an attention vector to concentrate on different parts of a sentence [29]. Along these lines, Ma et al. use two LSTM networks to model sentences and aspects separately [12]. They further use the hidden states generated from sentences to calculate attentions to aspect targets by a pooling operation, and vice versa. Hence their IAN model can attend to both the important parts in sentences and targets. Their method is similar to ours. However, the pooling operation will ignore the interaction among word-pairs between sentences and targets, and experiments show our method is superior to their model.

3 Method

Problem Definition

In this aspect level sentiment classification problem, we are given a sentence $s = [w_1, w_2, \dots, w_i, .., w_j, \dots, w_n]$ and an aspect target $t = [w_i, w_{i+1}, \dots, w_{i+m-1}]$. The aspect target could be a single word or a long phrase. The goal is to classify the sentiment polarity of the aspect target in the sentence.

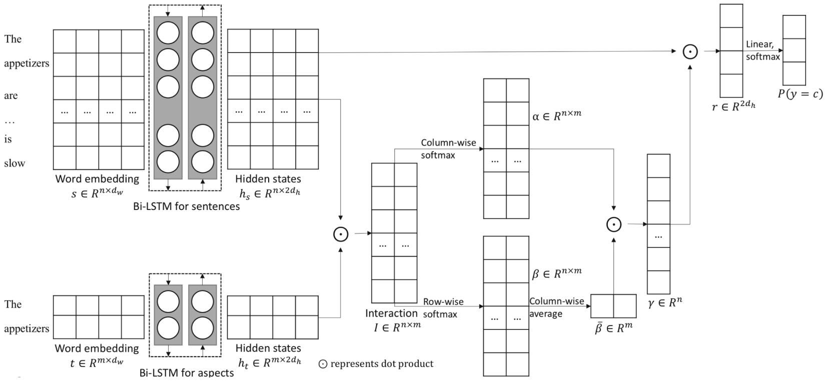


Fig. 1. The overall architecture of our aspect level sentiment classification model.

The overall architecture of our neural model is shown in Fig. 1. It is mainly composed of four components: word embedding, Bidirectional-Long short-term memory (Bi-LSTM), Attention-over-Attention module and the final prediction.

Word Embedding

Given a sentence $s = [w_1, w_2, \ldots, w_i, .., w_j, \ldots, w_n]$ with length n and a target $t = [w_i, w_{i+1}, \ldots, w_{i+m-1}]$ with length m, we first map each word into a low-dimensional real-value vector, called word embedding [1]. For each word w_i, we can get a vector $v_i \in R^{d_w}$ from $M^{V \times d_w}$, where V is the vocabulary size and d_w is the embedding dimension. After an embedding look up operation, we get two sets of word vectors $[v_1; v_2; \ldots; v_n] \in R^{n \times d_w}$ and $[v_i; v_{i+1}; \ldots; v_{i+m-1}] \in R^{m \times d_w}$ for the sentence and aspect phrase respectively.

Bi-LSTM

After getting the word vectors, we feed these two sets of word vectors into two Bidirectional-LSTM networks respectively. We use these two Bi-LSTM networks to learn the hidden semantics of words in the sentence and the target. Each Bi-LSTM is obtained by stacking two LSTM networks. The advantage of using LSTM is that it can avoid the gradient vanishing or exploding problem and is good at learning long-term dependency [5].

With an input $s = [v_1; v_2; \ldots; v_n]$ and a forward LSTM network, we generate a sequence of hidden states $\overrightarrow{h_s} \in R^{n \times d_h}$, where d_h is the dimension of hidden states. We generate another state sequence $\overleftarrow{h_s}$ by feeding s into another backward LSTM. In the Bi-LSTM network, the final output hidden states $h_s \in R^{n \times 2d_h}$ are generated by concatenating $\overrightarrow{h_s}$ and $\overleftarrow{h}_s$. We compute the hidden semantic states h_t for the aspect target t in the same way.

$$\overrightarrow{h_s} = \overrightarrow{LSTM}([v_1; v_2; \ldots; v_n]) \tag{1}$$

$$\overleftarrow{h_s} = \overleftarrow{LSTM}([v_1; v_2; \ldots; v_n]) \tag{2}$$

$$h_s = [\overrightarrow{h_s}, \overleftarrow{h_s}] \tag{3}$$

Attention-over-Attention

Given the hidden semantic representations of the text and the aspect target generated by Bi-LSTMs, we calculate the attention weights for the text by an AOA module. This is inspired by the use of AOA in question answering [2]. Given the target representation $h_t \in R^{m \times 2d_h}$ and sentence representation $h_s \in R^{n \times 2d_h}$, we first calculate a pair-wise interaction matrix $I = h_s \cdot h_t^T$, where the value of each entry represents the correlation of a word pair among sentence and target. With a column-wise softmax and row-wise softmax, we get target-to-sentence attention α and sentence-to-target attention β. After column-wise averaging β, we get a target-level attention $\overline{\beta} \in R^m$, which indicating the important parts in an aspect target. The final sentence-level attention $\gamma \in R^n$ is calculated by a weighted sum of each individual target-to-sentence attention α, given by Eq. (7). By considering the contribution of each aspect word explicitly, we learn the important weights for each word in the sentence.

$$\alpha_{ij} = \frac{exp(I_{ij})}{\sum_i exp(I_{ij})} \tag{4}$$

$$\beta_{ij} = \frac{exp(I_{ij})}{\sum_j exp(I_{ij})} \tag{5}$$

$$\bar{\beta}_j = \frac{1}{n}\sum_i \beta_{ij} \tag{6}$$

$$\gamma = \alpha \cdot \bar{\beta}^T \tag{7}$$

Final Classification

The final sentence representation is a weighted sum of sentence hidden semantic states using the sentence attention from AOA module.

$$r = h_s^T \cdot \gamma \tag{8}$$

We regard this sentence representation as the final classification feature and feed it into a linear layer to project r into the space of targeted C classes.

$$x = W_l \cdot r + b_l \tag{9}$$

where W_l and b_l are the weight matrix and bias respectively. Following the linear layer, we use a softmax layer to compute the probability of the sentence s with sentiment polarity $c \in C$ towards an aspect a as:

$$P(y = c) = \frac{exp(x_c)}{\sum_{i \in C} exp(x_i)} \tag{10}$$

The final predicted sentiment polarity of an aspect target is just the label with the highest probability. We train our model to minimize the cross-entropy loss with L_2 regularization

$$loss = -\sum_i \sum_{c \in C} I(y_i = c) \cdot log(P(y_i = c)) + \lambda ||\theta||^2 \tag{11}$$

where $I(\cdot)$ is an indicator function. λ is the L_2 regularization parameter and θ is a set of weight matrices in LSTM networks and linear layer. We further apply dropout to avoid overfitting, where we randomly drop part of inputs of LSTM cells.

We use mini-batch stochastic gradient descent with Adam [9] update rule to minimize the loss function with respect to the weight matrices and bias terms in our model.

4 Experiments

Dataset

We experiment on two domain-specific datasets for laptop and restaurant from SemEval 2014 Task 4 [27]. Experienced annotators tagged the aspect terms of

Table 1. Distribution by sentiment polarity category of the datasets from SemEval 2014 Task 4. Numbers in table represent numbers of sentence-aspect pairs.

Dataset	Positive	Neutral	Negative
Laptop-train	994	464	870
Laptop-test	341	169	128
Restaurant-train	2164	637	807
Restaurant-test	728	196	196

the sentences and their polarities. Distribution by sentiment polarity category are given in Table 1.

Hyperparameters Setting

In experiments, we first randomly select 20% of training data as validation set to tune the hyperparameters. All weight matrices are randomly initialized from uniform distribution $U(-10^{-4}, 10^{-4})$ and all bias terms are set to zero. The L_2 regularization coefficient is set to 10^{-4} and the dropout keep rate is set to 0.2 [21]. The word embeddings are initialized with 300-dimensional Glove vectors [17] and are fixed during training. For the out of vocabulary words we initialize them randomly from uniform distribution $U(-0.01, 0.01)$. The dimension of LSTM hidden states is set to 150. The initial learning rate is 0.01 for the Adam optimizer. If the training loss does not drop after every three epochs, we decrease the learning rate by half. The batch size is set as 25.

Model Comparisons

We train and evaluate our model on these two SemEval datasets separately. In order to further validate the performance of our model, we compare it with several baseline methods. We use accuracy metric to measure the performance.

Table 2. Comparison results. For our method, we run it 10 times and show "best (mean ± std)". Performance of these baselines are cited from their original papers.

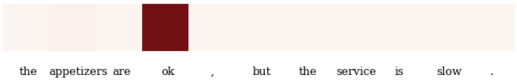
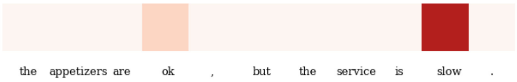
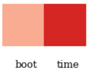
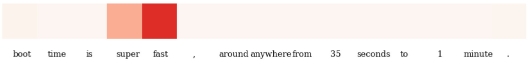
Methods	Restaurant	Laptop
TD-LSTM [24]	0.756	0.681
AT-LSTM [29]	0.762	0.689
ATAE-LSTM [29]	0.772	0.687
IAN [12]	0.786	0.721
AOA-LSTM	**0.812** (0.797 ± 0.008)	**0.745** (0.726 ± 0.008)

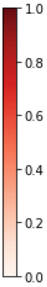
In our implementation, we found that the performance fluctuates with different random initialization, which is a well-known issue in training neural networks [22]. Hence, we ran our training algorithms 10 times, and report the average accuracy as well as the best one we got in Table 2. All the baseline methods only reported a single best number in their papers. On average, our algorithm is better than these baseline methods and our best trained model outperforms them in a large margin.

Case Study

In Table 3, we use some typical examples in test set to show the effectiveness of our model when learning the sentiment polarities of different aspects in sentences qualitatively. To analyze which word contributes the most to the aspect sentiment polarity, we visualize the final sentence attention vectors γ. In the first two examples, there are two aspects "appetizers" and "service" in the sentence. We can observe that when there are two aspects in the sentence, our model can automatically point to the right sentiment indicating words for each aspect. In the last example, the aspect is a phrase "boot time." From the sentence content, this model can learn "time" is the most important word in the aspect, which further helps it find out the sentiment indicating part "super fast."

Table 3. Examples of final attention weights for sentences. The color depth denotes the importance degree of the weight in attention vector γ.

Aspect	Sentence	Ans./Pred.
appetizers	 	0/0
service	 	-1/-1
 	 	+1/+1



Error Analysis

The first type of major errors comes from non-compositional sentiment expression which also appears in previous works [26]. For example, in the sentence "it took about 2 1/2 h to be served our 2 courses," there is no direct sentiment expressed towards the aspect "served." Second type of errors is caused by idioms used in the sentences. Examples include "the service was on point - what else you would expect from a ritz?" where "service" is the aspect word. In this case, our model cannot understand the sentiment expressed by idiom "on point." The third factor is complex sentiment expression like "i have never had a bad meal (or bad service) @ pigalle." Our model still misunderstands the meaning this complex expressions, even though it can handle simple negation like "definitely not edible" in sentence "when the dish arrived it was blazing with green chillis, definitely not edible by a human".

5 Conclusion

In this paper, we propose a neural network model for aspect level sentiment classification. Our model utilizes an Attention-over-Attention module to learn the important parts in the aspect and sentence, which generates the final representation of the sentence. Experiments on SemEval 2014 datasets show superior performance of our model when compared to those baseline methods. Our case study also shows that our model learns the important parts in the sentence as well as in the target effectively.

In our error analysis, there are cases that our model cannot handle efficiently. One is the complex sentiment expression. One possible solution is to incorporate sentences' grammar structures into the classification model. Another type of error comes from uncommon idioms. In future work, we would like to explore how to combine prior language knowledge into such neural network models.

Acknowledgments. This work is supported by Minerva - State Stability N00014-13-1-0835/N00014-16-1-2324 and Minerva - Dynamic Statistical Network Informatics - SCM N00014-15-1-2797. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of Minerva.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**(Feb), 1137–1155 (2003)
2. Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., Hu, G.: Attention-over-attention neural networks for reading comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp. 593–602 (2017)
3. Ding, X., Liu, B.: The utility of linguistic rules in opinion mining. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 811–812. ACM (2007)
4. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the 28th International Conference on Machine Learning (ICML-2011), pp. 513–520 (2011)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Huang, B., Carley, K.M.: On predicting geolocation of tweets using convolutional neural networks. In: Lee, D., Lin, Y.-R., Osgood, N., Thomson, R. (eds.) SBP-BRiMS 2017. LNCS, vol. 10354, pp. 281–291. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60240-0_34
7. Jiang, L., Yu, M., Zhou, M., Liu, X., Zhao, T.: Target-dependent twitter sentiment classification. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol, 1, pp. 151–160. Association for Computational Linguistics (2011)
8. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. Association for Computational Linguistics (2014)

9. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR) (2015)
10. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, vol. 333, pp. 2267–2273 (2015)
11. Liu, B., Blasch, E., Chen, Y., Shen, D., Chen, G.: Scalable sentiment classification for big data analysis using Naive Bayes classifier. In: 2013 IEEE International Conference on Big Data, pp. 99–104. IEEE (2013)
12. Ma, D., Li, S., Zhang, X., Wang, H.: Interactive attention networks for aspect-level sentiment classification. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-2017, pp. 4068–4074 (2017)
13. Medhat, W., Hassan, A., Korashy, H.: Sentiment analysis algorithms and applications: a survey. Ain Shams Eng. J. **5**(4), 1093–1113 (2014)
14. Nasukawa, T., Yi, J.: Sentiment analysis: capturing favorability using natural language processing. In: Proceedings of the 2nd International Conference on Knowledge Capture, pp. 70–77. ACM (2003)
15. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Sentiful: generating a reliable lexicon for sentiment analysis. In: 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009, pp. 1–6. IEEE (2009)
16. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, pp. 79–86. Association for Computational Linguistics (2002)
17. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
18. Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., et al.: Semeval-2016 task 5: aspect based sentiment analysis. In: ProWorkshop on Semantic Evaluation (SemEval-2016), pp. 19–30. Association for Computational Linguistics (2016)
19. Qiu, G., Liu, B., Bu, J., Chen, C.: Expanding domain sentiment lexicon through double propagation. In: Proceedings of the 21st International Jont Conference on Artifical Intelligence, vol. 9, pp. 1199–1204 (2009)
20. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
22. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: International Conference on Machine Learning, pp. 1139–1147 (2013)
23. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. Comput. Linguist. **37**(2), 267–307 (2011)
24. Tang, D., Qin, B., Feng, X., Liu, T.: Effective LSTMs for target-dependent sentiment classification. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 3298–3307 (2016)

25. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422–1432 (2015)
26. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 214–224 (2016)
27. Wagner, J., Arora, P., Cortes, S., Barman, U., Bogdanova, D., Foster, J., Tounsi, L.: DCU: aspect-based polarity classification for semeval task 4 (2014)
28. Wang, S., Manning, C.D.: Baselines and bigrams: simple, good sentiment and topic classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2. pp. 90–94. Association for Computational Linguistics (2012)
29. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: EMNLP, pp. 606–615 (2016)
30. Zhu, X., Sobihani, P., Guo, H.: Long short-term memory over recursive structures. In: International Conference on Machine Learning, pp. 1604–1612 (2015)

Analyzing Social Bots and Their Coordination During Natural Disasters

Tuja Khaund(✉), Samer Al-Khateeb(✉), Serpil Tokdemir(✉), and Nitin Agarwal(✉)

University of Arkansas at Little Rock, Little Rock, AR 72204, USA
{txkhaund, sxalkhateeb, sxtokdemir, nxagarwal}@ualr.edu

Abstract. Social bots help automate many sociotechnical behaviors such as tweeting/retweeting a message, 'liking' a tweet, following users, and coordinate or even compete with other bots. Social bots exist as advertising bots, entertainment bots, spam bots, and influence bots. In this research, we focus on influence bots, i.e., automated Twitter accounts that attempt to affect or influence behaviors of others. Some of these bots operate independently and autonomously for years without getting noticed or suspended. Furthermore, some of the more advanced influence social bots exhibit highly sophisticated coordination and communication patterns with complex organizational structures. This study aims to explore the role of Twitter social bots during the 2017 natural disasters and evaluate their coordination strategies for disseminating information. We collected data from Twitter during Hurricane Harvey, Hurricane Irma, Hurricane Maria, and Mexico Earthquake that occurred in 2017. This resulted in a total of over 1.2 million tweets generated by nearly 800,000 Twitter accounts. Social bots were detected in the data. Social networks of top bot and top non-bot accounts were compared to examine characteristic differences in their networks. Bot networks were further examined to identify coordination patterns. Hashtag analysis of the tweets shared by bots further helped in identifying hoaxes (such as, 'shark swimming on freeway') and non-relevant narratives (black lives matter, DACA, anti-Semitic narratives, Kim Jong-Un, nuclear test, etc.) that were disseminated by bots in several languages, such as French, Spanish, Arabic, Japanese, Korean, etc., besides English.

Keywords: Social bots · Twitter · Natural disasters · Crisis events
Severe weather · Hurricane · Earthquake · Disinformation · Coordination
Hoaxes · Alternate narratives

1 Introduction

A bot is a computer application that is designed to perform automated tasks over the Internet. The main idea behind creating a bot is to run simple tasks that are also structurally repetitive, at a rate much higher than humans [1]. A botnet refers to a collection of computer agents or bots that are programmed to act in a coordinated manner. Bots that mimic social behaviors of humans are referred to as social bots. Social bots could be of different types, viz., advertising bots, entertainment bots, spam bots, and influence bots [2]. In this research, we focus on influence bots, i.e., automated

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 207–212, 2018.
https://doi.org/10.1007/978-3-319-93372-6_23

or programmed accounts that attempt to affect or influence behaviors of the users with whom they interact. Social bots have various capabilities, e.g., they can affect users' perceived influence and learn the social graph to analyze people's posts and decide what to say and to whom. With all these capabilities, social bots have inarguably played an active role in affecting public discourse in online spaces (e.g., social media and chat forums) [3]. In this research, we investigate the role of social bots during four natural disaster events, namely *Hurricane Harvey*, *Hurricane Irma*, *Hurricane Maria*, and *Mexico Earthquake* that occurred in 2017. We seek answers to the following questions: (1) Are there characteristic differences between bot networks and human networks?; (2) What are these differences?; (3) Are there hoaxes or alternate narratives being disseminated during these events?; (4) What are these hoaxes and alternate narratives?; and (5) Do bots play a role in disseminating these narratives? To answer these questions, we examine social networks of bots and humans, coordination strategies used by bots, and analyze tweets shared by bots.

2 Literature Review

There is a growing body of research on detecting social bots. Journalists, analysts, and researchers have increasingly reported examples of the potential dangers ushered in by social bots [4]. Widespread diffusion of information by social bots may have unwarranted consequences on society. Social bots reportedly mislead, exploit, and manipulate social media discourse with rumors, spam, malware, misinformation, or simply noise. Sophisticated social bots can generate credible personas, and thus are more difficult for both people and filtering algorithms to detect [4]. In 2010, Chu et al. [5] proposed a classification system to determine whether a tweet belongs to a human, bot, or cyborg [5]. Over 500,000 accounts were studied to find their differences in tweeting behavior and content [5]. Wang et al. [6] have explored the possibility of crowdsourcing bot detection, i.e., using legions of human workers to detect bots. The authors assumed that bot detection is a simple task for humans because humans have a natural ability to evaluate conversational nuances (e.g., sarcasm or persuasive language) and to observe emerging patterns or anomalies. The authors observed the detection rate for hired workers drops off over time, although it remains good enough to be used in a majority voting protocol [6]. Abokhodair et al. [7] studied the "Syrian Social Bots" (SSB) which was used during the Syrian crisis in 2012. The authors focused on one botnet that was active for over eight months before Twitter detected and suspended it. The study analyzed the life and the activities of the botnet where it focuses on the content of the tweets. They found out that bots tend to share more news articles, fewer opinion tweets, no testimonial tweets, and fewer conversational tweets than any other legitimate Arabic or English Twitter user.

3 Methodology

We collected four datasets of Twitter users, using Google TAGS [8], including who tweeted and retweeted about four natural disasters events namely: *Hurricane Harvey, Hurricane Irma, Hurricane Maria*, and *Mexico Earthquake*. The data was collected from 08/30/2017 to 09/28/2017. This resulted in a total of 1,219,454 tweets that were generated by 776,702 Twitter accounts. To analyze persistent bot activity, we filtered down to those twitter accounts that engaged with all the four events. This resulted in 633,903 accounts that either tweeted or retweeted during the four events.

We calculated bot likelihood scores for the accounts using BotOrNot API [9], which ranges between 0 and 100, 100 being the highest likelihood for an account being a bot. For a robust and reliable analysis, we considered top 100 bot accounts (with BotOrNot score ranging between 90%–100%) and top 100 non-bot/human accounts (with BotOrNot score ranging between 0%–3%). We collected social network (i.e., friends and followers) of the top bot accounts and non-bot/human accounts. Twitter had already suspended some of the bot accounts and some human accounts were set to private, so we were able to collect social network information for 72 bot accounts (Fig. 1) and 82 human accounts (Fig. 2). We also analyzed the tweets shared by the top bot and non-bot accounts that resulted in 76,928 unique hashtags for the four events.

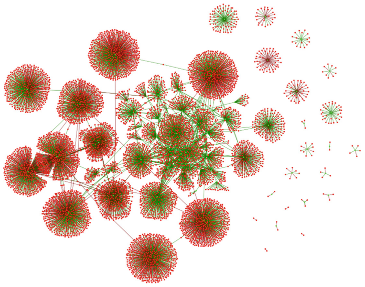


Fig. 1. Bot network, green nodes indicate bots and red nodes indicate their friends and followers. (Color figure online)

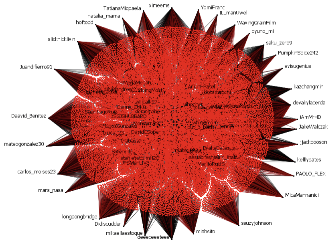
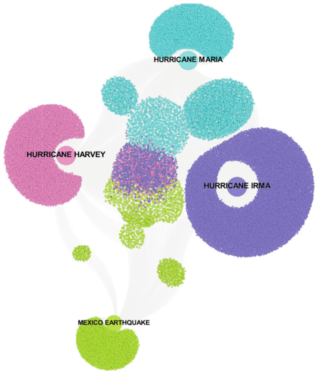


Fig. 2. Human network, black nodes indicate humans, and red nodes indicate their friends and followers. (Color figure online)

Community detection algorithm proposed by Blondel et al. [10] which optimize modularity score was run on the bot and human networks. We observed that human networks have more number of communities as compared to bot networks. Furthermore, human communities are smaller in size and denser as compared to bots. In other words, while humans have more tightly knit and focused communities, bots tend to make connections with rather weaker sense of belongingness to a community. Hence, bot communities tend to be bigger and less tightly knit (or less focused connections). Moreover, community detection also revealed strikingly different structural patterns between bots and humans. Bots' communities are more hierarchical in structure, i.e., there is a central core of members who connect more strongly among themselves as opposed to the peripheral members, who are weakly connected with the core as well as among themselves.

In addition to analyzing structural differences between human and bot networks, we examined their content, especially the hashtags. By constructing a hashtag co-occurrence network we identify hashtags that are (1) specific to the event and (2) common across multiple events. Further applying clustering to the hashtag co-occurrence network [10], we detect four main groups of hashtags (Fig. 3, left). Colors depict different clusters – one for each disaster event: pink for Hurricane Harvey hashtags, purple for Hurricane Irma, blue for Hurricane Maria, and green for Mexico earthquake. Hashtags common to multiple events were also identified in the hashtag co-occurrence network. Although, majority of the common hashtags refer to support, relief operations, prayers, and solidarity for the victims during these disaster events, we did observe a substantial presence of non-related hashtags that were common to multiple events. More troubling was the strategy bots used in pushing these non-related hashtags, i.e., non-related hashtags were latched on to the event-related hashtags. These non-related hashtags included hoaxes (such as 'shark swimming on freeway') and alternate narratives ('black lives matter', deferred action for childhood arrivals - commonly known as DACA, anti-Semitic narratives, Kim Jong-Un, nuclear test, etc.) that were disseminated by bots in several languages, such as French, Spanish, Arabic, Japanese, Korean, etc., besides English (Fig. 3, right).



Language	Hashtag (Translation)
English	#DACA, #BlackLivesMatter
Spanish	#VenezuelaDemocraciaYDiálogo (Venezuela Democracy and Dialogue), #Cáncer (Cancer)
Arabic	(The demise Of Israel) #زوال_إسرائيل, (The Jews) #اليهود
French	#Nucléaire (Nuclear), #GendarmerieEnOpération (Gendarmerie Special Operations)
Japanese	金正恩 (Kim Jong-un), 核试验 (Nuclear Test)

Fig. 3. Hashtag co-occurrence network (left) for the four natural disaster events in 2017 and non-relevant hashtags and their stated language (right). (Color figure online)

On digging deeper, we found 765 tweets that discussed the *Shark* hoax. We ran a modularity community detection algorithm on the tweet-retweet network which resulted in 69 clusters as shown in Fig. 4 (left) - colors depict different clusters - with the biggest cluster in blue. The blue cluster refer to a tweet "*A shark photographed on I-75 just outside of Naples, FL This is insane. #HurricaneIrma*" that was retweeted 420 times. We found that few accounts that disseminated the tweet had greater than 80% bot scores while a majority of the accounts had less than or equal to 50% bot scores (see, Fig. 4, right). Although the majority of the accounts had 50% or lower bot scores many of these accounts exhibit bot-like behaviors. One possible explanation is that these accounts may have helped dissemination of the hoax without proper

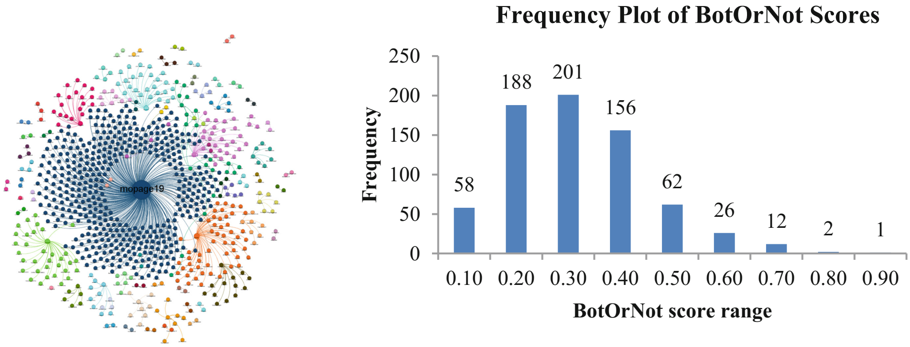


Fig. 4. Retweet network (left) of users who shared the 'shark' hoax during the four natural disaster events in 2017 and their bot scores (right). (Color figure online)

fact-checking. Alternatively, these accounts could be sophisticated bots that mimic human behavior to remain undetected by Twitter bot detection algorithms. Further research is needed to clarify the nature of these accounts.

4 Conclusion and Future Work

Social bots can disrupt discourse in online spaces. Social bots evolve constantly and become more sophisticated over time. We compared the social networks of bot and non-bot accounts that were identified during the 2017 natural disaster events. We observed, while humans have more tightly knit and focused communities, bots tend to make connections with rather weaker sense of belongingness to a community. Analysis of their content revealed that the discourse was not just limited to the disaster events. Non-relevant hashtags including hoaxes and alternate narratives were latched on to the event-specific hashtags and were disseminated in Spanish, Arabic, French, Japanese, among other languages.

The overarching research agenda is to investigate the different strategies that social bots use to coordinate disinformation campaigns and successfully manipulate online discourse. For future work, we will investigate the accounts that exhibit bot-like behavior despite having a low bot score. We will compare the communication network of bot and human accounts to identify other information maneuver tactics. We plan to expand our analysis to include entertainment and sport events to study the role of social bots in disseminating hoaxes, alternate narratives, uncertain, or ambiguous information.

Acknowledgments. This research is funded in part by the U.S. National Science Foundation (IIS-1636933, ACI-1429160, and IIS-1110868), U.S. Office of Naval Research (N00014-10-1-0091, N00014-14-1-0489, N00014-15-P-1187, N00014-16-1-2016, N00014-16-1-2412, N00014-17-1-2605, N00014-17-1-2675), U.S. Air Force Research Lab, U.S. Army Research Office (W911NF-16-1-0189), U.S. Defense Advanced Research Projects Agency (W31P4Q-17-C-0059) and the Jerry L. Maulden/Entergy Fund at the University of Arkansas at Little Rock. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding organizations. The researchers gratefully acknowledge the support.

References

1. Gayer, O.: What is an Internet Bot | How Bots Can Hurt Your Business. https://www.incapsula.com/blog/understanding-bots-and-your-business.html
2. Types of Bots: An Overview of Chatbot Diversity | botnerds.com, http://botnerds.com/types-of-bots/
3. @DFRLab: Le Pen's (Small) Online Army (2017). https://medium.com/dfrlab/le-pens-small-online-army-c754058630f0
4. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. Commun. ACM **59**, 96–104 (2016)
5. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on Twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 21–30. ACM (2010)
6. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., Zhao, B.Y.: Social turing tests: crowdsourcing sybil detection. ArXiv Prepr. arXiv:1205.3856 (2012)
7. Abokhodair, N., Yoo, D., McDonald, D.W.: Dissecting a social botnet: growth, content and influence in Twitter. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 839–851. ACM (2015)
8. Hawksey, M.: TAGS v6.0 ns. https://docs.google.com/spreadsheets/d/1EqFm184RiXsAA0TQkOyWQDsr4eZ0XRuSFryIDun_AA4/edit?pli=1&usp=embed_facebook&usp=embed_facebook
9. Botometer by OSoMe. https://botometer.iuni.iu.edu
10. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **2008**, P10008 (2008)

Sentiment Dynamics of *The Chronicles of Narnia* and Their Ranking

Kaiyun Dai[1,2](✉), Menglan Ma[1,3](✉), and Jianbo Gao[1,2,3](✉)

[1] Institute of Complexity Science and Big Data Technology, Guangxi University,
100 Daxue Road, Nanning 530004, Guangxi, People's Republic of China
1085917180@qq.com, lily.mml@qq.com, jbgao.pmb@gmail.com
[2] School of Computer, Electronics and Information, Guangxi University, 100 Daxue Road,
Nanning 530004, Guangxi, People's Republic of China
[3] School of Journalism and Communication, Guangxi University, 100 Daxue Road,
Nanning 530004, Guangxi, China

Abstract. Everyone in a civilized society grows up by reading stories. Fictions, including those for children, are an important type of stories, as they reflect social and cultural reality to some degree. The plot, the figures, and the environment of a fiction are the three main elements of a fiction. In particular, the development of the plot is pivotal for a fiction to be successful. It is now generally thought that sentiment dynamics of the fiction can well reflect the plot development. With the availability of a number of algorithms to automatically obtain the sentiment dynamics of a fiction, it has become increasingly desirable to fully understand its sentiment dynamics. This motivates us to use random fractal theory to study a set of popular children's fictions, *The Chronicles of Narnia*, written by the famed author, C.S. Lewis. We find the sentiment dynamics of each novel of the series possesses persistent long-range correlations, characterized by a Hurst parameter larger than 1/2. This has offered a mechanism to understand why many sentiment dynamics occurring naturally in a society or imagined by an author of a fiction can arouse strong emotions in humans. Interestingly, the value of the Hurst parameter for the series is strongly positively correlated with the score of the novels from Goodreads, suggesting that the scaling law governing sentiment dynamics can be used to objectively appraise the optimality of a fiction.

Keywords: *The Chronicles of Narnia* · Sentiment dynamics · Fractal analysis

1 Introduction

Stories are indispensable for civilized society development. We are surrounded with various stories during the process of growth. It seems that the stories have a special attraction to everyone. Stories are concentrated, summarized, and refined to the social reality, even myth or fairy tale is also based on reality rather than a figment of people's imagination. Fictions, including those for children, are an important type of stories, as they reflect social and cultural reality to some degree. Therefore, fiction is closely related to social reality.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 213–219, 2018.
https://doi.org/10.1007/978-3-319-93372-6_24

Aristotle once said: “We are unable to influence others through intelligence, emotions can do this.” Actually, Emotions play an important role in communications among human beings, as well as in rational learning. Every decision we made is more or less affected by emotions. In recent years, researchers are dedicated to sentiment analysis, with applications ranging from automated analysis of reviews and social media for purposes of marketing and customer service, to the monitoring of political issues, among many others [1]. While significant efforts have been made to detect sentiment [1–3], the analysis carried out thus far has largely been confined to detecting a polarity, or a mood, according to a limited set of emotions. That is far away from enough, compared with the works on static sentiments, sentiment dynamics is more worth studying. Better understanding the sentiment dynamics occurred naturally in society or in books means a lot to us.

Recent work in literary text analysis has suggested that, shifts in sentiment can serve as a useful proxy for plot development [4, 5]. After filtering, the spikes, troughs, and zeros of the smooth trend signals of sentiment are different and meaningful, because the stories are ingeniously conceived and these points represent the turning points in the story. This can explain the reason why the sentiment dynamics can serve as a useful proxy for plot development. Thus, in order to find which kind of sentiment dynamics can easily resonate with readers and the explanation of such resonation, we do sentiment analysis on the *The Chronicles of Narnia*, which is widely recognized as a successful series of fictions.

Random fractal theory is one of the most important theories developed for data analysis, the other one is Chaos. Many physiological and behavioral processes exhibit fractal dynamics. This means the measured patterns of change over time—the behavioral time series—exhibit certain properties, including self-similarity and scaling [6]. Adaptive fractal analysis (AFA) is a relatively new fractal analysis method that may hold promise in dealing with many types of real-world data. In this paper, we apply this method to do analysis.

The paper is organized as following: In Sect. 2, we give the brief description of AFA and research object. In Sect. 3, we perform this method on the sentiment dynamics of children literature *The Chronicles of Narnia* and show the experiment and results. Conclusions are in Sect. 4.

2 Method and Data

2.1 Method

Many fractal analyses concentrate explicitly on how to measure the variability scales with the size of a time window over which the measure is calculated [7]. Gao et al. provided a succinct and comprehensive treatment of various fractal analysis methods [8].

A parameter called the Hurst exponent, H, provides a way to quantify the “memory” or serial correlation in a time series [7]. Different H values have different meanings. In fact, $H = 0.5$ indicates the process is random. A finding of $0.5 < H < 1$ indicates the process will develop as the current trend, that is to say, the process has long-range correlations. In contrast, $0 < H < 0.5$ indicates an anti-persistent process, which means motion is likely to move in the opposite direction to the current trend.

AFA is based on a nonlinear adaptive multiscale decomposition algorithm [9]. The first step involves partitioning an arbitrary time series under study into overlapping segments of length $w = 2n + 1$, where neighboring segments overlap by $n + 1$ points. In each segment, the time series is fitted with the best polynomial of order M, obtained by using the standard least-squares regression; the fitted polynomials in overlapped regions are then combined to yield a single global smooth trend [10]. Denoting the fitted polynomials for the ith and (i + 1)th segments by $y^{i}\left(l_1\right)$ and $y^{(i+1)}(l_2)$, respectively, where $l_1, l_2 = 1, \ldots, 2n + 1$, we define the fitting for the overlapped region as

$$y^{(c)}(l) = w_1 y^{(i)}(l + n) + w_2 y^{(i+1)}(l),\ l = 1, 2, \ldots, n + 1 \tag{1}$$

where $w_1 = \left(1 - \frac{l-1}{n}\right)$ and $w_2 = \frac{l-1}{n}$ can be written as $(1 - dj/n)$ for $j = 1, 2$, and where dj denotes the distances between the point and the centers of $y^{(i)}$ and $y^{(i+1)}$, respectively. Note that the weights decrease linearly with the distance between the point and the center of the segment. Such a weighting is used to ensure symmetry and effectively eliminate any jumps or discontinuities around the boundaries of neighboring segments. As a result, the global trend is smooth at the non-boundary points, and has the right and left derivatives at the boundary [7]. The global trend thus determined can be used to maximally suppress the effect of complex nonlinear trends on the scaling analysis. The parameters of each local fit are determined by maximizing the goodness of fit in each segment. The different polynomials in overlapped part of each segment are combined using Eq. (1) so that the global fit will be the best(smoothest) fit of the overall time series. Note that, even if $M = 1$ is selected, i.e., the local fits are linear, the global trend signal will still be nonlinear. With the above procedure, AFA can be readily described. For an arbitrary window size w, we determine, for the random walk process u(i), a global trend v(i), $i = 1, 2, \ldots, N$, where N is the length of the walk. The residual of the fit, $u(i) - v(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to the following scaling equation:

$$F(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v(i))^2\right]^{1/2} \sim W^{H}. \tag{2}$$

Thus, by computing the global fits, the residual, and the variance between original random walk process and the fitted trend for each window size w, we can plot $\log_2 F(w)$ as a function of $\log_2 w$. The presence of fractal scaling amounts to a linear relation in the plot, with the slope of the relation providing an estimate of H [10]. The above are the basic steps of applying AFA.

2.2 Data

The Chronicles of Narnia is a series of seven fantasy novels by C. S. Lewis. It is considered a classic of children's literature and is the author's best-known work, having sold over 100 million copies in 47 languages. *The Chronicles of Narnia* has been adapted several times, complete or in part, for radio, television, the stage, and film. Lewis was awarded the 1956 Carnegie Medal for The Last Battle.

Set in the fictional realm of Narnia, a fantasy world of magic, mythical beasts, and talking animals, the series narrates the adventures of various children who play central roles in the unfolding history of that world. Except in The Horse and His Boy, the protagonists are all children from the real world, magically transported to Narnia, where they are called upon by the lion Aslan to protect Narnia from evil and restore the throne to its rightful line. The books span the entire history of Narnia, from its creation in The Magician's Nephew to its eventual destruction in The Last Battle.

The Lion, the Witch and the Wardrobe is the first published and best known of seven novels in *The Chronicles of Narnia*. *TIME* magazine included the novel in its "All-TIME 100 Novels" (best English-language novels from 1923 to 2005) [11]. In 2003, the novel was listed at number 9 on the *BBC*'s survey The Big Read [12].

Lewis began *The Magician's Nephew* soon after completing The Lion, the Witch and the Wardrobe, but he needed more than five years to complete it. The story includes several autobiographical elements and explores a number of themes with general moral and Christian implications, including atonement, original sin, temptation and the order of nature.

3 Experiment and Result

3.1 Long Range Correlations in Sentiment Dynamics

We take two fictions of the *The Chronicles of Narnia*, *The Lion, the Witch and the Wardrobe* and *The Magician's Nephew* as examples to show how we analyze the sentiment time series of the fiction and they are shown in Figs. 1 and 2.

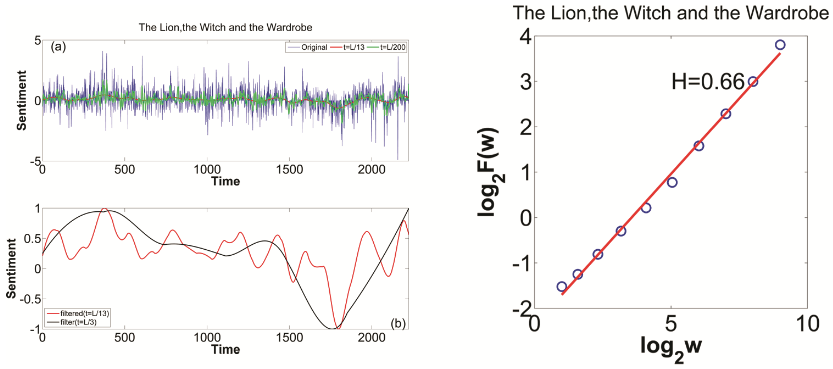


Fig. 1. Sentiment time series of *The lion, the witch and the wardrobe*. The blue, red, and green are for raw, smoothed data with w = 223, and smoothed data with w = 11, respectively. The blue circles denote the results from the sentiment time series, the red line is the best linear least squares fitting (Color figure online)

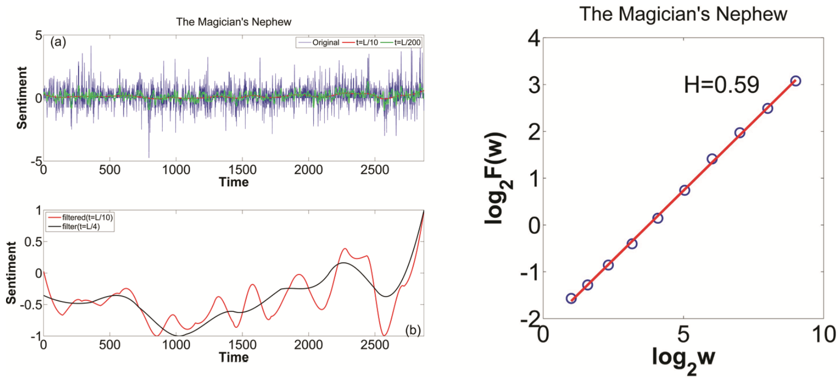


Fig. 2. Sentiment time series of *The Magician's Nephew*. The blue, red, and green are for raw, smoothed data with w = 287, and smoothed data with w = 14, respectively. The blue circles denote the results from the sentiment time series, the red line is the best linear least squares fitting (Color figure online)

To better know how the sentiment change, we rescale the value to the range of -1 to 1, and as showed in part (b) of Fig. 1. Since the red curves is still not smooth enough to learn about the mainly plot development, we choose a larger window to repeat above steps and get the black curve. Except the filters of raw sentiment, we use Hurst parameter to examine whether the raw sentiment has long-range persistence. After calculation, the Hurst parameter of the sentiment is 0.66 which means that the sentiment time series has long-range persistence.

Figure 2 shows result of the sentiment analysis on the book called The Magician's Nephew. The Magician's Nephew is the prequel to these series and mainly show "how all the comings and goings between our own world and the land of Narnia first began".

As we can find in part (b) of Fig. 2, the black curve dip to a low point near the 2500th which corresponds to the plot that the leading characters, Digory and his partners, got the apple which could protect the land of Narnia from the Witch through a hard adventure. Then the black curve rises to the highest point corresponding to the happy ending of this story that Digory successfully find the way to protect Narnia and cure his diseased mother. The Hurst value of this book is 0.59, which means the sentiment dynamics in this book is not in a mess and it has long-range correlations.

Each book has a H. After calculation, we get the Hurst value of the series book, the score of each book is 0.59, 0.66, 0.63, 0.59, 0.64, 0.59, 0.65 respectively (according to the sequence of story development). The Hurst value of the two most influential books is up to 0.66 and 0.65.

3.2 Comparison of Hurst Parameter and Scores from *Goodreads*

In order to explore the correlation between the score that we obtained from the goodreads (https://www.goodreads.com/) and the Hurst value we calculated, we draw the following image.

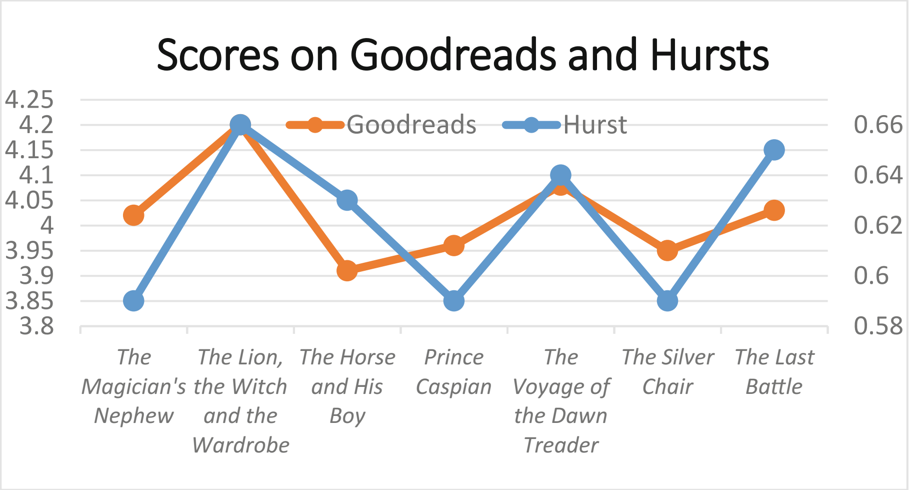


Fig. 3. Correlation between the Hurst parameter and scores on Goodreads website for the seven books.

Goodreads is the world's largest site for readers and book recommendations, which is a "social cataloging" website that allows individuals to freely search its database of books, annotations, and reviews. Readers can rate books from 1 to 5 stars. The score of each book is the average of each rating. The readers can rate the book following his mind. Figure 3 shows that the trend of Hurst and reader scores are basically the same. After calculation, the correlation coefficient between the two is 0.64, so the two have significant positive correlations.

4 Conclusion

No matter in stories or in daily lives, emotions count. In order to explore what kind of emotion changes can affect other people effectively, we use AFA method to study the sentiment dynamics in *The Chronicles of Narnia*. We calculate the Hurst of each book, the results show the Hurst of each book is larger than 1/2, we find that the sentiment dynamics of each novel of the series possesses persistent long-range correlations, which means sentiment dynamics is not cluttered and abrupt. Such sentiment dynamics are more receptive and in accordance with sentiment logic that can arouse strong feelings in readers which can explain the popularity among readers.

People's judgments on books are based on their own understanding. However, each person's cognitive level and reading experience are different. This makes people's evaluation of books highly subjective. This article starts from the dynamic point of view. Scientifically gives an objective method of measuring novels, which is unprecedented. Besides, value of the Hurst parameter for the series is positively correlated with the score of the fiction from Goodreads, suggesting that the scaling law governing sentiment dynamics can be used to appraise the optimality of a fiction and providing a reference for their ranking.

References

1. Cambria, E.: Affective computing and sentiment analysis. IEEE Intell. Syst. **31**(2), 102–107 (2016)
2. Cambria, E., Schuller, B., Xia, Y., et al.: New avenues in opinion mining and sentiment analysis. IEEE Intell. Syst. **28**(2), 15–21 (2013)
3. Frankenstein, W., Joseph, K., Carley, K.M.: Contextual sentiment analysis. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, pp. 291–300. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39931-7_28
4. Archer, J., Jockers, M.: The Bestseller Code. St. Martins Press, New York (2016)
5. Jockers, M.: Syuzhet: extracts sentiment and sentiment-derived plot arcs from text (2015)
6. Lebovitch, L.S., Shehadeh, L.A.: Introduction to fractals. In: Rileyand, M.A., Van Orden, G.C. (eds.) Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences, pp. 178–266 (2005). http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp. Accessed 28 June 2012
7. Riley, M.A., Bonnette, S., Kuznetsov, N., et al.: A tutorial introduction to adaptive fractal analysis. Front. Physiol. **3**, 371 (2012)
8. Gao, J.B., Cao, Y.H., Tung, W.W., Hu, J.: Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond. Wiley, Hoboken (2007)
9. Gao, J., Hu, J., Tung, W.: Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. PLoS ONE **6**(9), e24331 (2011)
10. Gao, J., Fang, P., Liu, F.: Empirical scaling law connecting persistence and severity of global terrorism. Phys. A Stat. Mech. Appl. **482**, 74–86 (2017)
11. Grossman, L.: All-TIME 100 Novels: The Lion, The Witch and the Wardrobe. Time, 16 October 2005. Accessed 25 May 2010
12. BBC - The Big Read. BBC, April 2003. Accessed 19 Oct 2012

Sign Prediction in Signed Social Networks Using Inverse Squared Metric

Mahboubeh Ahmadalinezhad(✉) and Masoud Makrehchi

University of Ontario Institute of Technology,
Oshawa, ON L1H 7K4, Canada
{mahboubeh.ahmadalinezhad,masoud.makrehchi}@uoit.ca

Abstract. This paper investigates the edge sign prediction problem in signed social networks, in which edges have either positive and negative sign. The main goal of this research is to effectively predict the sign of the edges using a newly proposed metric and with an emphasis on reducing the computational cost. In this study a new metric is introduced based on both neighbourhood and distance based link prediction measures. The sign of a connection between two users is subjected to the context of their relations. In the absence of these information, we can utilize topological and structural information of the network. The proposed metric has two components: (i) the importance of a node which is measured by node degree and (ii) the distance between two nodes which is penalizing the first component can be estimated by any shortest path algorithms. The new metric outperforms other sign prediction methods and also is computationally more affordable.

Keywords: Signed social networks · Sign prediction · Graph theory
Inverse squared metric

1 Introduction

This research focuses on signed social networks which are not explored to a good extend in the literature. Both positive and negative links provide important information about the social behaviour. In real-world problems, negative interactions can be important as positive connections. For example in social media analysis in which users express their trust or distrust corresponding to the other user's opinion although they can express approval or disapproval of a nomination.

However, few number of studies have focused on both positive and negative interactions in social networks [5,10,12]. For instance, in Epinions which is a product rating website, a link means trust or distrust of reviewers. In technology news website Slashdot, a link means either "friend" and "foe". However the interesting question is that how we can predict the sign of a given link between two nodes. The second question is what are the best predictive features and metrics for sign prediction task in absence of contextual information exchanged between the nodes by depending only on topological information from the network.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 220–227, 2018.
https://doi.org/10.1007/978-3-319-93372-6_25

The sign prediction problem is defined as predicting a missing sign based on the information from the rest of the network can include link, sign and/or direction of the links [4,9]. In this paper, a new measure called "Inverse Square Metric" is proposed to model the interactions between nodes in a social network.

The structure of this paper is organized as follows. In Sect. 2, related works are brought. In Sect. 3, the sign prediction problem is defined and the proposed method is mainly introduced. In Sect. 4, datasets for experimental purposes are introduced and in Sect. 5, implementation results are demonstrated and discussed. Finally, Sect. 6 concludes with future directions.

2 Related Works

In this section, a survey on some works related to sign prediction is conducted. Propagation of trust and distrust in the signed network of Epinions is introduced by Guha et al. [5], in which the adjacent matrix is used as the features of the model. Kunegis et al. [10] have investigated the features of the nodes and links in the friend/foe network of Slashdot Zoo. In [19], the effect of users' behaviour and their social interactions are analyzed, while in [1] an edge sign prediction problem is defined as a matrix factorization problem.

The applicability of balance and status theories from social psychology in sign prediction problem are demonstrated in [6,12], which provide valuable insight into the networks. Each link in the network is represented by a high-dimensional feature space, which is categorized into two classes. The first set of the features are related to node properties, and the second set deals with the number of cliques and their type. Zhang et al. [21] have extended the clique to rectangle pattern and used it together with the modified version of PageRank measure as the features of an edge in signed social networks.

Different ranking algorithms are applied to sign prediction problem. Shariari and Jalili [16] proposed the reputation and optimism as features which are obtained from ranking algorithms. In [15], frequent subgraph patterns are proposed. By modeling social neighbourhood of two users u and v as a synthesis of frequent subgraphs, the existence of certain graph patterns are shown, which help to estimate the link sign. Community-based algorithms, which group similar nodes, are used in sign prediction problem [17]. Transfer learning approach [20] is proposed to predict the signs for a newly formed signed social network based on the information of the existing and mature signed network. Trust and distrust prediction based on a low-rank matrix factorization method is proposed in [7].

3 Problem Statement

Let's $G = (V, E, S)$ be a directed and signed graph where $V = \{v_1, v_2, \ldots v_n\}$ is a subset of nodes, $E \subset V \times V$ is a subset of links between the nodes, and S is the sign of the edges ($|E| = |S|$). In signed social networks, each edge has a positive or negative label.

Due to sparsity of social networks, we have $|E| \ll |V \times V|$. A social network can be represented by an adjacency matrix $A = (a_{ij})_{n\times n}$ such that $a_{ij} \in \{-1, 0, 1\}$. A is a sparse matrix.

$$a_{ij} = \begin{cases} -1 < v_i, v_j > \in E, sgn(v_i, v_j) = -1 \\ 1 \quad < v_i, v_j > \in E, sgn(v_i, v_j) = 1 \\ 0 \qquad\qquad Otherwise \end{cases} \tag{1}$$

where $a_{ij} \neq a_{ji}$.

3.1 Inverse Square Metric

Barabasi and Albert [2] proposed a model of a growing network based on preferential attachment. The main idea of preferential attachment is the anecdote saying the rich get richer. In the proposed model, when a new node is added to the network, it prefers to get connected to highly connected nodes. It means the likelihood of the links between new node and other existing nodes is not uniform.

Moreover, Zhoa et al. [22] analyzed the network evolution on directed social networks and investigated some properties including node distribution and degree correlation. The results of the study by [22] on directed social networks show that the attributes of in-degree, out-degree and total degree follow the preferential attachment pattern. In this study, a new metric is introduced based on the preferential attachment model on signed social networks. The proposed metric is formulated as follows:

$$IVS(u, v) = \frac{deg(u) \cdot deg(v)}{|path|^2} \tag{2}$$

The proposed metric has two main components: (i) The importance and intensity of both nodes in a dyadic relation estimated by multiplying their node degrees, (ii) The distance between the two nodes which penalizes the first component is measured by the shortest path.

Since the input is a signed and directed social network, 16 different types of this formula is considered. For example, for one of the interactions, both with positive in-degree, the formula will be as follows:

$$IVS_{PIPI}(u, v) = \frac{Indeg_P(u) \cdot Indeg_P(v)}{|path|^2} \tag{3}$$

The proposed formula is similar to inverse-square law, where the distance between two point charges has negative effect on the amount of the electrostatic force.

$$F = \frac{q_1 \cdot q_2}{r^2} \tag{4}$$

In inverse-square law, point charges, q_1 and q_2 located at a distance r from each other, apply an electrostatic force F represented by Eq. 4 on each other. The charges with same sign repulse each other and vice versa. The links between the

Table 1. Statistics of the datasets used in the paper

	Epinions	Slashdot	Wikipedia
The number of nodes:	119,217	82,144	7,118
The number of edges:	841,000	594,202	103,747
Positive edges:	85.0%	77.4%	78.7%
Negative edges:	15.0%	22.6%	21.2%

users in a social network is in terms of the positive and negative links. A positive node means the node has more positive links than negative ones and acts similar to a positive charge. However, in social networks, nodes with same sign attract each other and reverse. In this study, the sign of a node is determined by the nodes in-degree and out-degree as discussed before. In the future works, the aim is to investigate other measures to determine the sign or intensity of a node in the network.

4 Dataset

In this paper, three large online social networks are considered in which each link is explicitly labeled as positive or negative: Epinions, Slashdot and Wikipedia. The statistics of these datasets are shown in Table 1. These online social networks are available through https://snap.stanford.edu/data/.

1. **Epinions:** is an online product rating site that users can express their opinions about products. The network consists of individual users connected by directed trust and distrust links [11].
2. **Slashdot:** is a signed social network of technology news site users (slashdot.org), connected by directed “friend” and “foe” relations [3].
3. **Wikipedia:** is an election network for choosing one user for administrator role. In this networks, users are able to vote a nominator as administrator in public dialogues and talks [4].
   In all networks approximately 80% of the edges have positive label.

5 Results

The proposed method is implemented on both imbalanced and balanced networks. In the imbalanced networks, about 80% of the edges are positive. A balanced network is created based on the framework of [5] and Leskovec et al. [12], in which the number of positive and negative edges are equal. For every negative edge, a random positive edge is chosen to ensure the balance of the network.

The results are based on 10-fold cross-validation and the average prediction F-scores is reported for all experiments. Different methods are used as baseline methods and defined as:

Table 2. Results of different methods for sign prediction problem on balanced datasets. F-score is reported in each cell. Higher F-score is desirable.

Method	Epinions	Slashdot	Wikipedia
Local features	0.878	0.862	0745
Status theory	0.848	0.734	0.714
Local and status theory	0.896	0.873	0.758
Link prediction features	0.880	0.866	0.746
Topological features	0.761	0.648	0.614
IVS features	**0.925**	**0.909**	**0.812**

Status Theory [13]**:** This theory is based on the extracted principles from social psychology, in which we can figure out the type of relationship between two nodes by taking advantage of the information from a third party. Status theory works for directed and signed social networks.

Local Features: This method considers local connections of a node to the rest of the network. Each edge is represented as a 7-dimensional vector of degree features. The edge from node u to node v is represented by in-degree of v, positive in-degree of node v, negative in-degree of node v, out-degree of node u, positive out-degree of node u, negative out-degree of node u, and the number of common neighbors.

Status Theory and Local Features [12]**:** This method uses 23-dimensional vectors for representation of each edge. 7 local features and 16 type of triadic status are considered as features.

Link Prediction Features: This method extracts 15 link prediction features for representation of each link in the networks: node degree, in-degree, out-degree, common neighbour, Jaccard [14], Adamic/Adar [14], preferential attachment [14], structural similarity [18], the length of shortest path, and Katz [8]. This group of features is selected, which provide maximum information at the lowest computational cost.

Topological Features: Information about in-degree and out-degree of the nodes in any network brings a large amount of data into the sign prediction analysis. Therefore, covers the deficiencies of the sign prediction method. In order to conduct a better evaluation of the sign prediction method, the authors believe that a more accurate comparison can be drawn by removing this information. As a result, only the topological features (node degree of two nodes (u, v), common neighbor, structural similarity, preferential attachment, Jaccard, Adamic, Katz, the length of shortest path) are used in this section of evaluation.

IVS Features: Inversed Squared Metric (IVS) attempts to combine the neighbourhood and distance-based information within a network and provide insight about the signs of interactions between the users.

Table 3. Supervised classification using different models by 16 IVS features for sign prediction problem. F-score are reported in each cell. Higher F-score is desirable

Classifier	Epinions	Slashdot	Wikipedia
Linear Regression	0.871	0.820	0.746
Ridge Regression	0.917	0.896	0.811
Lasso	0.917	0.899	0.812
SVM	0.680	0.760	0.599

Table 2 illustrates the results of all methods on balanced datasets. It should be noted that in balanced datasets the number of positive and negative edges in the network are equal. Table 2 shows that the F-score of IVS method improves significantly for all networks. Another strong point of the proposed method, IVS, is its lower computational cost.

The full datasets are most probably imbalanced. For example, Epinions has 80% positive links. Therefore, it is more probable that the training model predicts a positive link. In this way, it can be concluded that using balanced dataset makes the prediction task more difficult. In addition, if a method gives a good result on a balanced dataset, it is more likely that it can do even better on an imbalanced dataset [12,16].

Furthermore, different supervised learning models are evaluated by using 16 IVS features. Four classifier models (SVM, Linear Regression, Ridge Regression, and Lasso) are used and the results are compared. The main goal is to evaluate the performance of various models to predict the sign of an edge. The reason for selecting these four classifiers is to be in accordance with most of the previous works to solve sign prediction problems so the results can be compared [9,12]. Table 3 shows F-score of the four classifiers using IVS features. Lasso is observed to the best model across all datasets.

6 Conclusion

Negative links in social networks convey valuable information about users; however, they are neglected in many studies so far. Analyzing users' behaviour based on both positive and negative relations in the networks and finding a proper pattern are the challenging problems in social media. In this study, a new metric is introduced which has two components: (i) the importance of the two nodes are measured by node degree, and (ii) the distance between the two nodes is measured by the shortest path which is considered as a penalty of the first component. The improvement of this method in comparison to the baseline methodologies is significant for all three datasets.

One of the main findings of this research was proposing a new representation with acceptable complexity. The main computation cost of the new metric is that the distance-based algorithms which had efficient complexity and desirable F-score in comparison to the previous works.

References

1. Agrawal, P., Garg, V.K., Narayanam, R.: Link label prediction in signed social networks. In: IJCAI (2013)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
3. Chiang, K.Y., Hsieh, C.J., Natarajan, N., Dhillon, I.S., Tewari, A.: Prediction and clustering in signed networks: a local to global perspective. J. Mach. Learn. Res. **15**(1), 1177–1213 (2014)
4. Chiang, K.Y., Natarajan, N., Tewari, A., Dhillon, I.S.: Exploiting longer cycles for link prediction in signed networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 1157–1162. ACM (2011)
5. Guha, R., Kumar, R., Raghavan, P., Tomkins, A.: Propagation of trust and distrust. In: Proceedings of the 13th International Conference on World Wide Web, pp. 403–412. ACM (2004)
6. Heider, F.: Attitudes and cognitive organization. J. Psychol. **21**(1), 107–112 (1946)
7. Hsieh, C.J., Chiang, K.Y., Dhillon, I.S.: Low rank modeling of signed networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 507–515. ACM (2012)
8. Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953)
9. Kumar, S., Spezzano, F., Subrahmanian, V., Faloutsos, C.: Edge weight prediction in weighted signed networks. In: 2016 IEEE 16th International Conference on Data Mining (ICDM), pp. 221–230. IEEE (2016)
10. Kunegis, J., Lommatzsch, A., Bauckhage, C.: The slashdot zoo: mining a social network with negative edges. In: Proceedings of the 18th International Conference on World Wide Web, pp. 741–750. ACM (2009)
11. Kunegis, J., Preusse, J., Schwagereit, F.: What is the added value of negative links in online social networks? In: Proceedings of the 22nd International Conference on World Wide Web, pp. 727–736. ACM (2013)
12. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web, pp. 641–650. ACM (2010)
13. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1361–1370. ACM (2010)
14. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. J. Am. Soc. Inf. Sci. Technol. **58**(7), 1019–1031 (2007)
15. Papaoikonomou, A., Kardara, M., Tserpes, K., Varvarigou, T.A.: Predicting edge signs in social networks using frequent subgraph discovery. IEEE Internet Comput. **18**(5), 36–43 (2014)
16. Shahriari, M., Jalili, M.: Ranking nodes in signed social networks. Soc. Netw. Anal. Min. **4**(1), 172 (2014)
17. Shahriary, S.R., Shahriari, M., Noor, R.: A community-based approach for link prediction in signed social networks. Sci. Program. **2015**, 5 (2015)
18. Xu, X., Yuruk, N., Feng, Z., Schweiger, T.A.: Scan: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 824–833. ACM (2007)

19. Yang, S.H., Smola, A.J., Long, B., Zha, H., Chang, Y.: Friend or frenemy?: predicting signed ties in social networks. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 555–564. ACM (2012)
20. Ye, J., Cheng, H., Zhu, Z., Chen, M.: Predicting positive and negative links in signed social networks by transfer learning. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1477–1488. ACM (2013)
21. Zhang, T., Jiang, H., Bao, Z., Zhang, Y.: Characterization and edge sign prediction in signed networks. J. Ind. Intell. Inf. **1**(1) (2013)
22. Zhao, J., Lui, J.C., Towsley, D., Guan, X., Zhou, Y.: Empirical analysis of the evolution of follower network: a case study on Douban. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 924–929. IEEE (2011)

Detecting and Characterizing Bot-Like Behavior on Twitter

SiHua Qi(✉), Lulwah AlKulaib, and David A. Broniatowski

George Washington University, 2121 I St NW, Washington, DC 20052, USA
{qisihu,lalkulaib,broniatowski}@gwu.edu

Abstract. Social media is becoming a platform of choice for people to voice their opinion on topics of discussion. To evaluate these opinions, it is important to have an accurate assessment of who is saying what. Unfortunately, social media are also the home of bots which makes the assessment difficult. Bots are computer programs designed to mimic human behavior online in social networks. They are used to pursue a variety of goals, including, but not limited to, spreading information, and influencing targets.

In this paper, we describe a machine learning framework that uses content-based features extracted from Twitter to detect bot-like behavior on the platform. Unlike other machine-learning approaches to bot detection, we seek to generate explanations of why specific accounts are categorized as bots; thus allow us to modify these criteria as bots' behaviors evolve. We have therefore developed the criteria mentioned in an article published in Medium [1] to detect bot-like behavior in our dataset then evaluate the results. We then explain the different types of bots that used as our datasets and compare the significant features for each type of bots in a logistic regression method.

Keywords: Bots · Bot-like behavior · Logistic regression

1 Introduction

Online social network sites are becoming more popular each day. According to a report by Dream Grow, Twitter is considered among the top 15 most popular social networking sites. It has 330 M active users monthly, which puts it in 4th place on that list. [2] Unfortunately, the number of bot accounts is surprisingly large. A new paper from University of Southern California and Indiana University suggests that up to 15% of twitter accounts are in fact bots rather than people [3].

Bots on Twitter are accounts controlled by a software, automatically producing content, and interacting with other users. Some of these bots use Twitter as a tool to announce news headlines, others utilize the platform for marketing, such bots are considered useful bots. However, there is a growing record of misuse of bot accounts. These accounts would be designed to mimic human behavior, then sold to users aiming to boost their popularity with fake followers, [4] used to promote terrorist propaganda, [5] or used by some organizations to influence public opinion [6].

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 228–232, 2018.
https://doi.org/10.1007/978-3-319-93372-6_26

Twitter allows bots on the platform that adhere to their rules, automated likes are not allowed, and automated retweets are only allowed for entertainment, informational, or novelty purposes. [twitter-automation] These rules, and other issues have been addressed after the DARPA challenge which was a Twitter bot detection challenge to study malicious activities carried by bot accounts. [7] While working on this project we noticed that published papers talk about bot detection. Meanwhile, users that try to hide that their accounts constantly violate Twitter rules, are not always fully automated. Bots can be turned on and off as needed. When the program is not running the account, a human would be posting, which makes these accounts harder to detect.

In this paper, we defined our criteria that determines what a bot account acts like as bot-like behavior. We used criteria from an article published by Nimmo [1], and some others that we added as the study evolved. Unlike other approaches that try to predict whether an account is a bot or not based on holdout data [3], we use a statistical approach that aims to provide explanatory insight into why our assignment is made. The rest of this paper is organized as followed. In Sect. 2, related work is discussed. In Sect. 3, We propose criteria used for bot-like-behavior detection. We use the criteria to train our model which detects accounts that satisfy any of the criteria and generates a report for each dataset with the results. In Sect. 4, we explain the results generated by the study.

2 Related Work

Twitter has been widely used since 2006, and the open structure of twitter lead people to question who is tweeting early on. Chu et al. [8], classified twitter users into human, bot, and cyborg accounts using 4 components each of which checks a specific criterion, then compute a score that enables classification. Davis et al. [9], created a service that evaluates the extent to which a Twitter account exhibits the similarity to the known characteristics of social bots. Their platform fetches a given account's recent activity, then computes and returns a bot-likelihood score. The DARPA Twitter bot challenge [7] also addressed four different features they assigned to different teams to work on in the bot detection challenge. The detection systems created in the challenge were all semi-supervised and all teams used human judgement to augment automated bot identification processes.

3 Bot-Like Behavior Detection

3.1 Data Collection

Twitter has a set of API functions [10] that supports user information collection. Our data was collected using the Twitter API, where we crawled the most recent 200 posts by users from a known bot list [11]. We used a dataset consisting of 4 types of manually verified twitter bots: Fake Followers, Traditional Spam Bots, Social Spam Bots and Content Polluters. We also pulled a list of verified legitimate users from the same source.

3.2 Methods

We designed a study to describe a list of user behaviors for each twitter account. Using the article by Nimmo [1], we created a program that would detect the features that indicate bot-like behavior. After data collection, we ran the script that generated results for each user against our criteria. Using those results, we applied a stepwise logistic regression model based on Akaike Information Criteria (AIC) values to determine which of the 19 features were relevant when detecting bot-like behavior. Features are explained in Table 1.

Table 1. Features used for bot-like-behavior detection.

Feature name	Explanation
digit_screen_name	screen_name consists of digits only
scramble_name	screen_name consists of alpha numeric scrambles
default_profile_image	using default profile image
default_background_image	using default background image
url_shortner	using url shorteners in tweet content
low_post_high_result	retweet count or like count is more than number of followers for given account
multi_language	more than 2 languages appeared in tweets crawled
tweet_frequency	average daily tweet number
time_range	average days between two consecutive tweets
rt_number	number of retweets/total tweets crawled
#of_mentions	average number of mentions in original tweets crawled for this account
#of_hyperlinks	average number of hyperlinks in original tweets crawled for this account
#of_friends	number of friends
#of_followers	number of followers
status_num	number of tweets
#of_favorites	number of favorited tweets
most_recent_time	most recent tweet timestamp
tweet_avg_word_number	average number of words in each original tweet
tweet_lexical_diversity	number of unique words used in all crawled original tweets

4 Analysis

We tested our script on all four bot categories data that we collected. Examining the results, we used the cut off value $|z| = 2$ as a threshold to extract features which are more relevant to the model. The $|z| = 2$ cut off value corresponds to two-sided hypothesis test with a significance level of = 0.05. A big magnitude of z-score indicates that the corresponding true regression coefficient is not 0 and that the variable matters. Based on the features we got, we described a series of bot-like behaviors.

For all four types of bots, there were 2 features in common: “most_recent_time” with a negative z-score and “status_num” with a positive z value which indicated if a user is more active recently, the user is most-likely not a bot and the user with a high number of tweets are more likely be a bot.

Based on the reports generated (Table 2), we noticed that each bot type has features more significant to that type. Fake follower bots are “simple accounts that inflate the number of followers of another account” [9]. Our results showed that fake follower accounts do not tweet frequently, but they have a significant number of friends consistent with their purpose: making other accounts popular. On the other hand, content polluters bots are designed to generate spam while masquerading as humans [12]. According to our analysis, content polluters have a high average number of tweets per day, and a significant number of friends. This corresponds to the idea of spam accounts in general, where accounts are trying to increase their outreach. In contrast, traditional spam bots are “a group of automated accounts spamming job offers” [9], which are easily identifiable as automated. In our dataset, the average time between two posts by traditional spam bots is short. Traditional spambots also rarely post retweeted content. Such behavior is consistent with accounts that are designed to posting job advertisements. Finally, social spam bots are “spammers of products on sale at Amazon.com” or “spammers of paid apps for mobile devices” [9]. The report shows that these accounts post several of hyperlinks, and do not engage in conversations (twitter mentions). Social spam bot behavior that we found here is consistent with their content suggested by the source.

Table 2. Features relevant to bo types.

Feature name	Fake followers		Content polluters		Traditional spam		Social spam	
	z	P > |z|	z	P > |z|	z	P > |z|	z	P > |z|
status_num	4.64	0	13.555	0	3.367	0.001	2.71	0.007
tweet_frequency	4.541	0	13.598	0	−4.378	0	−4.986	0
#of_friends	3.409	0.001	5.242	0	4.239	0	–	–
avg_word_number	−2.239	0.025	–	–	–	–	–	–
multi_language	−2.732	0.006	–	–	–	–	–	–
most_recent_time	−3.177	0.001	−5.005	0	−5.316	0	−8.441	0
scramble_name	–	–	−2.994	0.003	4.03	0	−3.55	0
rt_number	–	–	−4.51	0	−3.74	0	–	–
#of_favorites	–	–	3.648	0	–	–	–	–
url_shortner	–	–	–	–	−2.379	0.017	–	–
avg_time_btw_status	–	–	–	–	−3.263	0.001	–	–
#of_hyperlinks	–	–	–	–	−4.098	0	2.93	0.003
#of_followers	–	–	–	–	–	–	4.383	0
default_background_ image	–	–	–	–	–	–	−2.147	0.032
low_post_high_result	–	–	–	–	–	–	−2.579	0.01
#of_mentions	–	–	–	–	–	–	−2.797	0.005

5 Conclusion

Our results demonstrate that bot-like behavior differs significantly with bot design. Specifically, one may be able to infer the functional purpose for which the bot was created by exploring the specific features along which that particular bot type differs from human users. Future work in the area of bot detection could benefit from combining explanatory approaches grounded in traditional statistical analyses, in addition to the machine-learning approaches that are already in widespread usage.

References

1. Nimmo, B.: #BotSpot: Twelve ways to spot a bot, 28 August 2017. https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c
2. Kallas, P.: Top 15 most popular social networking sites and apps, February 2018. https://www.dreamgrow.com/top-15-most-popular-social-networking-sites/
3. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization (2017). http://arxiv.org/abs/1703.03107
4. Confessore, N., Dance, G., Harris, R., Hansen, M.: The follower factory. New York Times, 27 January 2018. https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html
5. The ISIS Twitter census: defining and describing the population of ISIS supporters on Twitter. States News Service, 13 March 2015
6. Ferrara, E., Wang, W., Varol, O., Flammini, A., Galstyan, A.: Predicting online extremism, content adopters, and interaction reciprocity (2016). https://doi.org/10.1007/978-3-319-47874-6_3. http://arxiv.org/abs/1605.00659
7. Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F.: The DARPA Twitter bot challenge. Computer, **49**(6), 38–46 (2016). https://doi.org/10.1109/mc.2016.183. http://ieeexplore.ieee.org/document/7490315
8. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on Twitter. Paper presented at the 21–30, 6 December 2010. https://doi.org/10.1145/1920261.1920265. http://dl.acm.org/citation.cfm?id=1920265
9. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: BotOrNot: a system to evaluate social bots (2016). https://doi.org/10.1145/2872518.2889302. http://arxiv.org/abs/1602.00975
10. Rules and policies. https://help.twitter.com/en/rules-and-policies/twitter-automation
11. Bot repository (2017). https://botometer.iuni.iu.edu/bot-repository/datasets.html
12. Lee, K., Eoff, B., Caverlee, J.: Seven months with the devils: a long-term study of content polluters on Twitter

Initializing Agent-Based Models with Clustering Archetypes

Samaneh Saadat, Chathika Gunaratne, Nisha Baral, Gita Sukthankar(✉), and Ivan Garibay

University of Central Florida, Orlando, FL, USA
gitars@eecs.ucf.edu

Abstract. Agent-based models are a powerful tool for predicting population level behaviors; however their performance can be sensitive to the initial simulation conditions. This paper introduces a procedure for leveraging large datasets to initialize agent-based simulations in which the population is abstracted into a set of archetypes. We show that these archetypes can be discovered using clustering and evaluate the benefits of selecting clusters based on their stability over time. Our experiments on the GitHub dataset demonstrate that simulation runs performed with the clustering archetypes are more successful at predicting large-scale activity patterns.

Keywords: Agent-based models · GitHub archetypes
Unsupervised learning · Stable clustering

1 Introduction

The aim of our research is to create a versatile agent-based model for simulating large-scale usage trends of the GitHub collaborative development tool. GitHub repositories typically have several developers working on the project as a virtual team. However open source projects hosted on GitHub can be downloaded and copied thousands of times, spawning an ecosystem of related repositories. Although it is possible to predict usage trends on GitHub using purely machine learning approaches [1], we believe that a hybrid approach of agent-based modeling and data mining is more promising, enabling us to explore a richer range of community interactions.

One challenge of modeling developer behavior is that GitHub has become popular as a general purpose hosting and communication tool for myriad types of efforts, ranging from personal software archives to large open source projects with millions of users. Many repositories are not directly related to software development but are instead used to curate document collections [2]. Previous studies of computer-supported cooperative software development have attempted to survey the developers to understand the differences between individual contributors

This research was supported by DARPA program HR001117S0018.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 233–239, 2018.
https://doi.org/10.1007/978-3-319-93372-6_27

vs. rockstar programmers and popular curators [3]. There is a large amount of variability in the usage rates of GitHub, with some developers submitting hundreds of changes in a month, but with most users remaining completely dormant or passively observing.

Expressing the diversity of the user population within a single agent-based framework is demanding. Rather than relying on existing taxonomies of user behavior created from survey data, we extract the archetypes from the user's contribution history from the most stable clusters found by k-means clustering. This approach also has the advantage of simultaneously producing the relative distribution of each archetype across the developer population, along with the monthly activity. Our results show that our archetype extraction and simulation initialization procedure produces more accurate predictions of population behavior, as measured by the Gini coefficient of contributor activities.

2 Approach

Our dataset consists of all GitHub users and repositories created before March 2017 along with the activity data from January 2015 to February 2017. We divided 26 months of data into 20 months for training the clustering and 6 months for testing the simulation. There are approximately seven million users with at least one activity during the training period, but we restrict our analysis to the three million users with greater than ten total activities.

The GitHub activity dataset consists of 14 event types: *CommitComment*, *Create*, *Delete*, *Fork*, *Gollum*, *IssueComment*, *Issue*, *Member*, *Public*, *Pullrequest*, *PullrequestReviewComment*, *Push*, *Release*, and *Watch*. These can be grouped into three general categories: contributions, watches, and forks (copies). We created activity profiles of GitHub users using the average monthly activity per event type to be used as clustering features. Since GitHub users have a wide range of activity levels, first we partitioned users based on their average monthly activity and then clustered each partition separately. Table 1 shows the number of users in each partition.

Since the range of values for different event types varies widely, we normalized features by scaling them to lie between zero to one. We clustered each partition separately using k-means but restricted our analysis to partitions with greater than one hundred users.

Table 1. GitHub user partitions

Partition	Average monthly activity	Number of users
1	(0,10]	1.4 M
2	(10, 100]	1.5 M
3	(100, 1 K]	44 K
4	(1 K, 10 K]	741
5	(10 K, inf]	69

2.1 Cluster Stability

One question is whether clustering the data from different time periods yields the same archetypes. Are the data-driven archetypes more sensitive to monthly activity fluctuations than archetypes described in survey studies? To examine this question, we performed a cluster stability analysis to measure whether similar clusters are observed from month to month. Similarity is computed between clusterings of consecutive months, and the stability score is the average similarity score of all consecutive months [4]. The Adjusted Rand Index (ARI) is used to measure similarity between clusterings. ARI is 1.0 when clusters are identical and close to 0.0 for random labeling [5]. The following procedure is used to calculate stability score for each k value:

Given a set $M = m_1, m_2, ..., m_n$ of monthly user activity profiles in the training months, the k-means algorithm takes the number of clusters, k as input:

```
1. For $k = 2, ..., k_{max}$
  1.1. For $i = 1, ..., n$
         Cluster data of $m_i$ into $k$ clusters to obtain model $CM_i$ and clusters $C_i$
  1.2. For $i = 2, ..., n$
         Cluster data of $m_i$ using $CM_{i-1}$ to obtain clusters $C_i'$
  1.3. Compute stability as the mean similarity between clustering $C_i$ and $C_i'$
```

$$Stability(k) = \frac{1}{(n-1)^2} \sum_{i=2}^{n} Similarity(C_i, C_i') \tag{1}$$

```
2. Choose the parameter $k$ that gives the highest stability:
```

$$K = \arg\max_k Stability(k) \tag{2}$$

2.2 Archetype Model

The clustering results were then used to initialize the archetypes included in our agent-based model of GitHub repository contribution, developed on NetLogo 6.0.2 [6]. *General User* archetypes were created using the best and second best clusters from partitions 1 to 4 in Table 1. *Hyperactive Users* were defined by aggregating the event activity profiles of the 69 users in partition 5 into mean frequency per event type. Accordingly, the cluster size for *Hyperactive Users* was set equal to the count of users in partition 5.

Using the above archetypes, a scaled-down agent-based model was constructed. Two agent breeds were modeled: (1) *Users* and (2) *Repositories. Repositories* were considered a non-active breed of agents that kept track of contributions made by *User* agents. *User* agents were allowed to perform one out of a set of actions, U, that reflected actual GitHub events plus the event *Idle* for the case that a User did not perform an event during that time step. Since activity was partitioned on monthly basis, the per minute frequency of a GitHub event

($U_j^{GH} : U^{GH} = U - Idle$) being triggered by a *General User* ($a_i : i <= 16$) was referred to as $F_{U_j^{GH},a_i}$ and calculated as:

$$F_{U_j^{GH},a_i} = \frac{A_{U_j^{GH},a_i}}{43200} \tag{3}$$

For *General Users*, $F_{U_j^{GH},a_i} < 1$ and was modeled as the probability that a user of archetype a_i would trigger a GitHub event U_j^{GH} in a discrete simulation time step. Accordingly, each discrete time step was used to represent a minute. *Hyperactive Users* were modeled as triggering $F_{U_j^{GH},a_17}$ where $F_{U_j^{GH},a_17} > 1$ such that many events of type U_j^{GH} were generated per simulation time step. This difference in event triggering could have been accommodated by modeling simulation time steps as milliseconds but was performed to reduce the runtime of the simulations to a computationally feasible limit. The model was scaled down by 1000th of the population size of GitHub in the training data and cluster sizes.

In addition to the rate at which *Users* performed events of different types which was informed by the clustering results, there was the question of modeling the target repository for each event. To handle this, we obtained the mean number of repositories a user would interact with during a month from the data (mean = 3.6, st.dev = 37.046). These values were used to calculate the maximum number of repositories a user would work with in the simulated month, μ, through a gamma distribution ($\alpha = \frac{3.6^2}{37.046^2}$, $\lambda = \frac{3.6}{37.046^2}$).

Users maintained a list of familiar repositories. Each simulation time step, a user agent selected a behavior to perform based on the event frequency defined by its archetype. If this event was a contribution event, one of the repositories in its contribution list would be selected as the target of this event. If, in a simulation time step, a user decided to perform a *Watch* or *Fork* event, the user chose σ repositories at random from the repository population and selected the repository with the highest sum of *Fork*s and *Watch*es from this subset as the target for this action. This repository would then be added to its list of familiar repositories, displacing a repository already in this list at random if the list was already at capacity μ.

Some key ABM parameters were directly inferred from statistics of the training data. New *Users* were injected into the simulation at a probability of 0.008 per time step. *Create* events generated new repositories at a probability of 0.481955 as actual create events can result in repositories, branches and tags.

3 Results

Our experiments examine three questions:

1. do stable clusters exist across consecutive months in the partitioned GitHub data?
2. does the ABM configured with the extracted archetypes outperform the simple mean model?

3. are archetypes extracted from more stable clusters better than those from less stable clusters?

First, we examine the stability of the clustering across consecutive months. Table 2 shows the stability score for k-means clustering with $k = 3, ..., 9$ in all 4 partitions. k values of 4, 3, 4, and 3 generate the most stable clusters for partition 1, 2, 3, and 4 respectively. Partitions 1, 2, and 3 definitely exhibit stable clusters as their best stability scores exceed 0.9.

Table 2. Stability score for different partitions

# Clusters	0–10	10–100	100–1 K	1 K–10 K
3	0.931	**0.996**	0.920	**0.685**
4	**0.949**	0.988	**0.928**	0.671
5	0.912	0.919	0.841	0.557
6	0.825	0.989	0.867	0.604
7	0.795	0.934	0.751	0.594
8	0.791	0.974	0.796	0.577
9	0.769	0.973	0.617	0.571

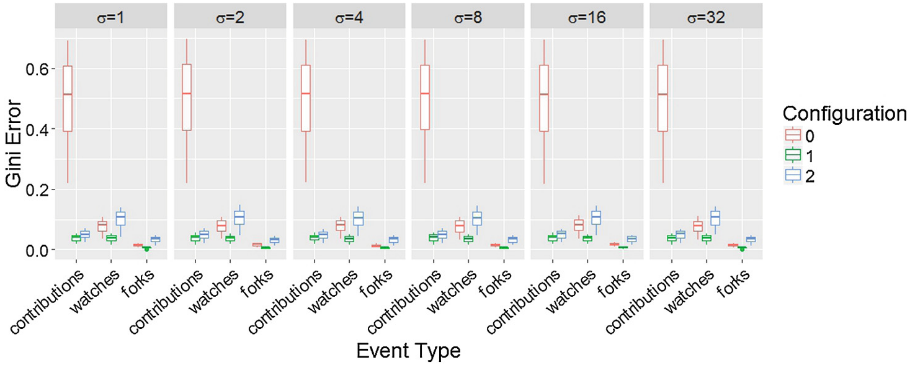


Fig. 1. Error of Gini coefficient for users. Configuration 0 (without cluster information) performs badly at predicting the dispersion of contributions across users. Configuration 1 (most stable cluster) is the best performer yielding a small improvement vs. using the second most stable cluster to initialize the archetypes.

To evaluate the simulation performance of our extracted archetypes, we ran several configurations of the agent-based model. The baseline (Configuration 0) models the entire population using the mean event frequencies. Configuration 1 uses the archetypes from the most stable clustering result on each partition, yielding 15 archetypes (14 *General Users* and 1 *Hyperactive User*). Configuration

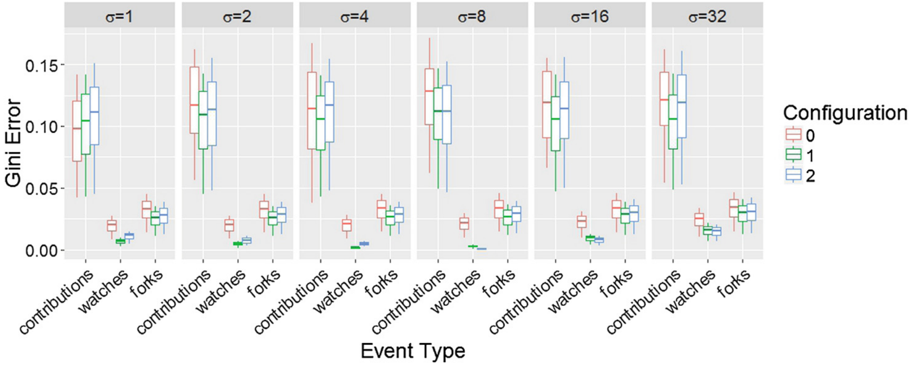


Fig. 2. Error of Gini coefficient for repositories. Configurations 1 and 2 (cluster based archetypes) yield slightly better performance. However, the ad hoc heuristics used by the simulation for repository assignment do not perform as well at allocating events across repos.

2 used the second best clustering results yielding 17 archetypes. The simulations were run for 43200 time steps, simulating a month of GitHub activity with $\sigma \in 1, 2, 4, 8, 16, 32$. Each configuration was repeated ten times to obtain aggregate simulation results.

Although our ABM is designed to answer questions about many types of GitHub trends, we are particularly interested in accurately modeling the relative activity levels of users and repositories since these are core aspects of the ABM that affect many population-level trends. Our experiments measure the absolute error of different initial archetype populations at predicting the Gini coefficient over one month of test data. Rather than looking at the errors of specific event types, we group the events into meaningful action categories: (1) contributions, (2) watches, and (3) forks. Figure 1 shows the performance of the cluster-based archetypes at predicting the Gini coefficient over user contributions for one month of test data. We also study the performance of our repository allocation heuristics at predicting the Gini coefficient over repository activity (Fig. 2).

4 Conclusion and Future Work

The stable cluster-based user archetypes outperform the baseline and the less stable clusters at predicting the dispersion of activity across users. These archetypes offer slight improvements in calculating the dispersion across repos, however the heuristics for repo assignment do not perform as well. In future work, we plan to extend our clustering approach to discover repo archetypes that can be used to make a more informed decision about the assignment of events to repos.

References

1. Borges, H., Hora, A., Valente, M.T.: Predicting the popularity of GitHub repositories. In: Proceedings of the International Conference on Predictive Models and Data Analytics in Software Engineering (2016)
2. Wu, Y., Kropczynski, J., Prates, R., Carroll, J.M.: Rise of curation in GithHub. In: AAAI Conference on Human Computation and Crowdsourcing (2015)
3. Blincoe, K., Sheoran, J., Goggins, S., Petakovic, E., Damian, D.: Understanding the popular users: following, affiliation influence and leadership on GitHub. Inf. Softw. Technol. **70**, 30–39 (2016)
4. Von Luxburg, U., et al.: Clustering stability: an overview. Found. Trends Mach. Learn. **2**(3), 235–274 (2010)
5. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
6. Wilensky, U.: Netlogo. Technical report, Center for Connected Learning and Computer-based Modeling, Northwestern University, Evanston, IL (1999)

Applications for Health and Well-Being

Predicting Alcoholism Recovery from Twitter

Jennifer Golbeck(✉)

University of Maryland, College Park, USA
jgolbeck@umd.edu

Abstract. We show that social media data, in the form of Twitter profiles, can be used to automatically and accurately predict whether or not an alcoholic entering treatment will achieve and maintain sobriety. This analysis is based on a dataset of 270 Twitter (225 in the training set and 45 in the test set) users who announced that they were attending their first Alcoholics Anonymous meeting and, subsequently, their sobriety or return to drinking. Our model uses a tree-based machine learning approach to make predictions over a feature set developed from automated text analysis, social network analysis, and a quantified estimation of relevant factors identified by the addiction research community. The model correctly predicts recovery status after 90 days with 80% accuracy and ROC AUC of 0.815. We describe how this data works together to produce a model, and discuss the opportunities and challenges resulting from the ability to make these types of predictions.

1 Introduction

Through social media, the web has transformed from something reserved for the technologically savvy into something where more than a billion people share the details of their daily lives. This growing repository of personal information has triggered a new field of research focused on inferring insights about users. Predicting product preferences [3], personal attributes [26], behaviors, and even futures [8] are just a few examples.

Our research investigates the ability to predict a person's potential for future success from their past actions. We accurately predict whether alcoholics will achieve and maintain sobriety by analyzing their Twitter behavior *before* they begin treatment. First, we found several hundred alcoholics on Twitter who announced that they were attending their first Alcoholics Anonymous (AA) meeting. From their subsequent tweets, we selected those individuals who explicitly stated whether they had maintained sobriety or returned to drinking at the 90 day mark.

Then, using data from before their first AA meeting, we developed a feature set that leverages insights from the addiction community. The features include data about subjects' social connections, language, and psychological attributes. From this, we can predict whether each individual will maintain sobriety at 90 days with 80% accuracy.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 243–252, 2018.
https://doi.org/10.1007/978-3-319-93372-6_28

2 Related Work

Over the past five years, a large body of work has been produced that shows the power to infer information about individuals and to predict their future behavior.

Personality traits have been a popular target for inference algorithms, and they leverage a variety of data as inputs. These include social network connections [38], social media profiles and interactions [1,12,19,20,28], and language analysis [11,33,35,38]. Other work inferred personal value systems [4,22], political preferences [18], intelligence [26], sexual orientation [25], and demographic attributes like age, gender, religion, and race [26].

Similarly, models can infer insights about relationships online, including tie strength [15], trust between two people [2,16,21,27,39], and even the likelihood a romantic relationship will last [3].

Inferring or predicting future psychological and medical conditions has also seen attention and success in these models, including work on depression [8,36], post-partum depression [7], heart disease risk [10], and post-traumatic stress disorder (PTSD) [9].

3 Methods

We have lightly paraphrased all the tweets to follow to prevent deanonymization of subjects through a Twitter search.

3.1 Subject Selection

We originally began by searching Twitter for any mention of “first AA meeting” between January 1, 2013 and November 1, 2015. We selected users who were announcing they themselves were attending their first meeting for themselves, eliminating jokes, people attending to support an alcoholic friend or family member, and people who were attending for class assignments (a common practice in nursing and counseling programs). For the remaining users (N = 526), we used the Twitter API to collect their most recent 3,200 tweets, the maximum allowable under Twitter’s Terms of Service. These were separated into tweets “before” and “after” the first AA meeting.

The addiction community marks 90 days of sobriety as the first step in short term recovery [37]. To establish if a subject was recovered or not at the 90 day mark, we looked at all tweets made after attending the first AA meeting through the end of the data that was available to us. If anyone posted about drinking before the 90 day mark, we established that they were not recovered (e.g. *I am hungover as balls lol* or *Taking 5 shots of vodka after I left work tonight was not a good idea).* If the subjects tweeted about their sobriety (e.g. *I’ve officially been sober for 4 months*), we knew they were recovered. For each subject, we required clear, unambiguous evidence that explicitly discussed drinking or recovery. We excluded all subjects from our data who did not have such statements. This eliminated 301 of 526 initial users, leaving 225 in our sample.

Note that some alcoholics may be considered recovered if they are drinking with no symptoms of abuse or dependence over a 1-year period [6]. Since we are looking at subjects who are struggling with alcohol and thus attending AA, we did not consider that any drinking within a 90-day period of a first AA meeting could be considered "symptom free"', and thus only considered abstainers as recovered.

We acknowledge that this is self-reported information about recovery. However, it is derived from spontaneous statements made publicly on social media. Thus, while it is possible that someone is maintaining a false facade of recovery or remission, we believe the information is, overall, an accurate reflection of subjects' drinking status.

For each of these users, we confirmed that we had access to 1,000 words in their "before" tweets which would support our ability to do a robust automated text analysis. The result was a data set of 225 Twitter users for whom we had recovery/relapse data at 90 days. Among our analyzed population, 22.8% were sober after 90 days. Alcoholism recovery rates are hard to pin down. Short term (1-year) recovery rates range from 20–50% depending on the severity of the addiction and the criteria for recovery [31]. NIH reports a long term 18.2% recovery rate [6]. We are looking at initial recovery in the 90-day period, but the recovery rate among our subjects is generally in line with these figures.

In the first phase of this work, we experimented only with these users. However, because we engineered and selected down our feature set (described below), we were at risk of overfitting to this dataset. Thus, we designated this original group of subjects as a training set and collected a second test set of data. We used the same process but selected users who posted about attending their first AA meeting between November 1, 2015 and September 1, 2016. After the same filtering process, the test set had 45 users.

3.2 Feature Set Development

We developed our feature set by analyzing the training set data. First, we collected raw features. We used the Receptiviti API[1] to do basic psycholinguistic text analysis on our subjects' tweets, which includes analysis from the 2015 Linguistic Inquiry and Word Count (LIWC) [34] plus composite features. This generated 141 different features. We also gathered basic Twitter behavioral information, including frequency of tweeting, retweeting, mentions, etc.

We also analyzed tweets to determine whether subjects were over 21 (the legal drinking age in the US). Since users do not provide their age in their profiles, we made a best guess by reading the tweets. To prevent bias, we reviewed only the tweets from before the subject's first AA meeting and did not know if they were recovered or not when determining age. Some accounts tweeted explicitly about being under 21 (e.g. *Only 37 months until I turn 21!*). The most common way

[1] http://receptiviti.com.

of identifying those who were under 21 was through posts that made it clear the author was still in high school. These included posts about class periods, teachers, and events like prom (e.g. *I'm not excited about prom at all so I probably won't even go this year*). Signs of someone over 21 included discussions of their children going to school, getting divorced, or posting their age on their birthdays. For nearly every author, there were fairly clear indications of where the person fell on the age spectrum. Still, this is a human-generated heuristic since existing age-inference algorithms (e.g. [32]) were unavailable to us.

To account for social factors, we collected a list of each subject's bi-directional friends on Twitter (i.e. people who followed the subject and whom the subject followed back). For each of these friends, we collected their 200 most recent tweets as a snapshot of their tweeting behavior. We then analyzed the user's—and their friends'—tweets for mentions of alcohol, using an alcohol word dictionary we developed[2]. For both subjects and friends, we counted the frequency with which they used our set of identified alcohol words, and the percentage of friends who tweeted about alcohol. We also engineered a feature measuring the frequency of alcohol-related tweets by the subjects' friends, weighted according to the frequency of their interactions.

Coping styles represent psychological and behavioral strategies people use to deal with stressful situations, and they are known to be closely tied to addiction [5]. They may be *adaptive*, helping to reduce stressors; or they may be *maladaptive*, which tend to alleviate symptoms without addressing the underlying problem. Adaptive (positive) coping encompasses an analytic approach to problem solving and use of healthy relationships for support [30]. Maladaptive (negative) coping styles are often linked to conditions that grow from an inability to deal with stress. For example, alcoholism is often tied to and even predicted by an avoidance-based coping style [13,31]. Similarly, PTSD is linked to a dissociative coping style [14].

We used an existing automatic classifier developed by [17] to determine if each subject had a "good, adaptive" coping style, or a "poor, maladaptive" one.

Finally, we performed a basic bag-of-words document classification using the Naïve Bayes Multinomial algorithm in the open source machine learning toolkit, Weka [23]. While this did not achieve high accuracy when trained on the training set and evaluated on the test set (55% accuracy (ROC AUC = 0.53), it did serve as a useful indicator in our larger feature set.

4 Classification Results

From these selected and engineered features, we used a tree-based machine learning algorithm - REPTree as implemented in the Weka toolkit [23] - to classify users as "recovered" or "not recovered" at the 90-day mark. The classifier was built on the training set and we evaluated its accuracy on the test set. Because the classes are unbalanced, with many more subjects relapsing, we used the Cost Sensitive meta classifier in Weka to do class balancing.

[2] Link to public repository removed for anonymous submission.

When the classifier was tested against the training set, we achieved 78.6% accuracy with an area under the receiver-operating characteristic curve (AUC) of 0.915. However, as mentioned above, this has a risk of overfitting. We evaluated the model using the test set. Our algorithm correctly predicted if people would maintain their sobriety after 90 days with 80% accuracy. Table 1 shows detailed accuracy statistics by class and overall.

The overall AUC is 0.747 This indicates that, given one member of each class, the algorithm will correctly classify them 74.7% of the time. This shows the algorithm performs far better than random guessing (Table 2).

Table 1. Detailed accuracy measures for both classes and overall at 90 days.

	N	Precision	Recall	F-Measure	AUC
90 days - Not Recovered	28	0.7	0.824	0.757	0.747
90 days - Recovered	17	0.88	0.786	0.83	0.747
90 days - Weighted Avg.	45	0.812	0.8	0.802	0.747

Table 2. Confusion mmatrix for Recovered (R) and Not Recovered (NR) at 90 days.

Prediction		
NR	R	True Class
14	3	NR
6	22	R

We separated the training and test sets because we developed the model from the training set. We worried about overfitting if we did n-fold validation over the data that included the training set. However, for completeness of the analysis we also the classifier with all the data merged into a single set. We used a 10-fold validation and achieved better results that those reported on the test set. Accuracy was 85%, ROC AUC = 0.815, and the confusion matrix is shown in Table 3.

Table 3. Confusion mmatrix for Recovered (R) and Not Recovered (NR) at 90 days.

Prediction		
NR	R	True Class
37	28	NR
12	192	R

4.1 Linking Features to Addiction Research

Machine learning is often criticized for being a black box. One can build a large feature set, throw it at an algorithm, and often produce good results—but without any real understanding of how the model was built, or why certain features are predictive.

Sometimes this is because correlations emerge that are purely statistical in origin—especially in big data. Other times, correlations emerge that really do represent, or map to, meaningful real-world connections. In either case, it is both satisfying and validating to build models that leverage features that reinforce known relationships in existing literature.

In this work, we have a mix of both cases. One of our major features comes from performing document classification with Weka. It performs classification based solely on a vector of words used by each subject. This approach is common and often successful, but it yields few insights into the data. We also selected and engineered features which have ties to addiction literature. In a review of the alcoholism literature, Moos and Moos [31] found that recovery was linked to adaptive coping styles and having supportive, sober friends. Thus, we explicitly created features to represent that.

We infer coping style from a linguistic analysis tweets using a technique presented in [17]. We have several features that reflect a supportive, sober social network, as well. These include the percentage of friends who tweet about alcohol, and a weighted measure of friend alcohol influence that leverages data about user interactions on Twitter.

Among the psycholinguistic features we included were words related to *affect*, *insecurity*, and *drives* (including *goal orientation*). Being goal-driven is a known predictor of alcoholism recovery, while insecurity and affective disorders are known predictors of relapse [29]. The same study also found language skills that were tied to success, which we measured with the grammatical features in our set.

We also included a categorical feature indicating whether the subject was over or under 21. Prior research shows that when people begin drinking under 21, especially early in adolescence, they are more likely to experience addiction, and their addictions are more severe [24].

Taken together, these linked features mean we can estimate known predictors of alcoholism recovery by automatically processing social media accounts, and then use that to build highly accurate predictors. If this technique generalizes, it means that many tools could be created to predict people's future behaviors.

5 Discussion

We show that recovery from alcoholism can be accurately predicted by analyzing people's social media activities before starting the recovery process. These results demonstrate our algorithmic abilities in this particular context—but they also represent a step forward in predicting people's future behavior from their past behavior. Both aspects hold promise and cause for concern.

There is potential for using this algorithm in treatments and interventions. Recall that all of our subjects publicly announced the first step in treating their addictions. Many expressed enthusiasm for changing their lives, yet failed to remain sober over even the short term. A tool that can predict recovery outcomes could be used in a supportive role for people likely to recover, and in a cautionary role (i.e. a signal that deeper changes may be necessary) for people unlikely to recover.

We envision potential for an app where people who are considering entering treatment for alcoholism can analyze their own profiles. Because our algorithm is based on features that relate to known factors in addiction recovery, this app could analyze a user's Twitter profile and highlight the positive and negative factors in a their environment to suggest helpful action. For example, it could suggest that the user has a good network of Twitter friends who are not focused on alcohol use, and echo that this is important in offline friends as well. Someone with a maladaptive coping style could be guided to treatments, like cognitive-behavioral therapy, that can help them adjust their stress reactions and improve their chances of recovery. This app need not show a prediction of if the user will actually succeed in recovery (which could be discouraging). The success of these features in creating a predictor illustrate how an algorithm could be combined with advice from addiction specialists to offer initial guidance to alcoholics.

As a broader example of predicting future outcomes, this demonstrates the potential for work that leverages existing literature to predict success in overcoming other addictions, making lifestyle changes related to health (such as weight loss), or even achieving success in relationships.

At the same time, serious concerns follow in terms of both application and privacy.

On the application side, this algorithm could be used in way that may negatively or unfairly affect people's lives. Consider that programs like AA are often offered as alternatives to jail time for DUI offenses. One could imagine this algorithm being used to determine whether or not to offer a person alternative treatment; that means an algorithmic prediction of non-recovery could lead to greater jail time. This is a concerning application of the technology.

On the privacy side, these sort of behavioral prediction can be made without users' consent and can reveal information they prefer to keep private. Furthermore, because the features used in these predictions are not easily controllable (that is, they rely on linguistic characteristics or friends' behavior), users have little means of preventing the inferences from happening. One need only look to concerns about applications to see why people may feel violated by this technology.

Where do we go from here? There is a necessary ethical, legal, and policy debate to come: how should these technologies be used, and what control should users should have over that data (explicit or inferred)? This discussion needs to happen sooner rather than later, lest users face the choice of either forced privacy invasion or withdrawal from online life.

References

1. Adalı, S., Golbeck, J.: Predicting personality with social behavior: a comparative study. Soc. Netw. Anal. Min. **4**(1), 1–20 (2014)
2. Avesani, P., Massa, P., Tiella, R.: A trust-enhanced recommender system application: Moleskiing. In: Proceedings of the 2005 ACM symposium on Applied computing, pp. 1589–1593. ACM (2005)
3. Backstrom, L., Kleinberg, J.: Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 831–841. ACM (2014)
4. Chen, J., Hsieh, G., Mahmud, J.U., Nichols, J.: Understanding individuals' personal values from social media word use. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2014, New York, NY, USA, pp. 405–414. ACM (2014)
5. Cooper, M.L., Russell, M., George, W.H.: Coping, expectancies, and alcohol abuse: a test of social learning formulations. J. Abnorm. Psychol. **97**(2), 218 (1988)
6. Dawson, D.A., Grant, B.F., Stinson, F.S., Chou, P.S., Huang, B., Ruan, W.: Recovery from dsm-iv alcohol dependence: United States, 2001–2002. Addiction **100**(3), 281–292 (2005)
7. De Choudhury, M., Counts, S., Horvitz, E.: Predicting postpartum changes in emotion and behavior via social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3267–3276. ACM (2013)
8. De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: ICWSM (2013)
9. Coppersmith, G., Harman, C., Dredze, M.: Measuring post traumatic stress disorder in Twitter (2014)
10. Eichstaedt, J.C., Schwartz, H.A., Kern, M.L., Park, G., Labarthe, D.R., Merchant, R.M., Jha, S., Agrawal, M., Dziurzynski, L.A., Sap, M., et al.: Psychological language on twitter predicts county-level heart disease mortality. Psychol. Sci. **26**(2), 159–169 (2015)
11. Farnadi, G., Zoghbi, S., Moens, M.-F., De Cock, M.: Recognising personality traits using facebook status updates. In: Proceedings of WCPR, pp. 14–18 (2013)
12. Ferwerda, B., Schedl, M., Tkalcic, M.: Predicting personality traits with instagram pictures. In: Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015, EMPIRE 2015, New York, NY, USA, pp. 7–10. ACM (2015)
13. Fromme, K., Rivet, K.: Young adults' coping style as a predictor of their alcohol use and response to daily events. J. Youth Adolesc. **23**(1), 85–97 (1994)
14. Gil, S., Caspi, Y.: Personality traits, coping style, and perceived threat as predictors of posttraumatic stress disorder after exposure to a terrorist attack: a prospective study. Psychosom. Med. **68**(6), 904–909 (2006)
15. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 211–220. ACM (2009)
16. Golbeck, J.: Trust on the world wide web: a survey. Found. Trends Web Sci. **1**(2), 131–197 (2006)
17. Golbeck, J.: Predicting coping style from twitter. In: Proceedings of the International Conference on Social Informatics (SocInfo16) (2016)

18. Golbeck, J., Hansen, D.: A method for computing political preference among Twitter followers. Soc. Netw. **36**, 177–184 (2014)
19. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), pp. 149–156. IEEE (2011)
20. Golbeck, J., Robles, C., Turner, K.: Predicting personality with social media. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 253–262. ACM (2011)
21. Golbeck, J.A.: Computing and applying trust in web-based social networks (2005)
22. Gou, L., Zhou, M.X., Yang, H.: Knowme and shareme: understanding automatically discovered personality traits from social media and user sharing preferences. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, CHI 2014, pp. 955–964. ACM, New York (2014)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
24. Hingson, R.W., Heeren, T., Winter, M.R.: Age at drinking onset and alcohol dependence: age at onset, duration, and severity. Arch. Pediatr. Adolesc. Med. **160**(7), 739–746 (2006)
25. Jernigan, C., Mistree, B.F.T.: Gaydar: facebook friendships expose sexual orientation. First Monday **14**(10) (2009)
26. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proc. Natl. Acad. Sci. **110**(15), 5802–5805 (2013)
27. Levin, R., Aiken, A.: Attack resistant trust metrics for public key certification. In: 7th USENIX Security Symposium, January 1998
28. Markovikj, D., Gievska, S., Kosinski, M., Stillwell, D.: Mining facebook data for predictive personality modeling. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM 2013), Boston, MA, USA (2013)
29. Miller, L.: Predicting relapse and recovery in alcoholism and addiction: neuropsychology, personality, and cognitive style. J. Subst. Abuse Treat. **8**(4), 277–291 (1991)
30. Moos, R.: Coping With Life Crises: An Integrated Approach. Springer, Heidelberg (1976)
31. Moos, R.H., Moos, B.S.: Rates and predictors of relapse after natural and treated remission from alcohol use disorders. Addiction **101**(2), 212–222 (2006)
32. Nguyen, D.-P., Gravel, R., Trieschnigg, R.B., Meder, T.: How old do you think I am? A study of language and age in Twitter (2013)
33. Park, G., Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Kosinski, M., Stillwell, D.J., Ungar, L.H., Seligman, M.E.P.: Automatic personality assessment through social media language. J. Personal. Soc. Psychol. **108**(6), 934 (2015)
34. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Assoc. **71**, 2001 (2001)
35. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: Predicting personality with Twitter. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), pp. 180–185. IEEE (2011)
36. Reece, A.G., Danforth, C.M.: Instagram photos reveal predictive markers of depression. arXiv preprint arXiv:1608.03282 (2016)

37. Searles, J.S., Helzer, J.E., Rose, G.L., Badger, G.J.: Concurrent and retrospective reports of alcohol consumption across 30, 90 and 366 days: interactive voice response compared with the timeline follow back. J. Stud. Alcohol **63**(3), 352–362 (2002)
38. Staiano, J., Lepri, B., Aharony, N., Pianesi, F., Sebe, N., Pentland, A.: Friends don't lie: inferring personality traits from social network structure. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, UbiComp 2012, pp. 321–330. ACM, New York (2012)
39. Ziegler, C.-N., Lausen, G.: Spreading activation models for trust propagation. In: Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Service, Taipei, Taiwan. IEEE Computer Society Press, March 2004

The Portrayal of Quit Emotions: Content-Sensitive Analysis of Peer Interactions in an Online Community for Smoking Cessation

Vishnupriya Sridharan[1], Trevor Cohen[1], Nathan Cobb[2],
and Sahiti Myneni[1](✉)

[1] The University of Texas School of Biomedical Informatics, Houston, TX, USA
{Vishnupriya.Sridharan, Sahiti.Myneni}@uth.tmc.edu
[2] Georgetown University Medical Center, Washington, DC, USA

Abstract. Tobacco use causes serious emotional harm among smokers and it manifests in the form of mood disorders such as depression and anxiety. The effects of smoking cessation on quality of life are well documented. However, our understanding of emotional well-being of an individual in the window of quit and relapse period to provide just in time support is quite limited. In this study, we focus on social engagement, communication attributes, and emotional landscape of successful quitters as manifested in peer interactions of an online health community for smoking cessation. Further, we employed Word Embedding techniques to analyze the content-specific communication attributes in a given quit episode at scale. Results indicate users were highly engaged after a quit. The emotional index of successful quitters highlighted the fragile and complex nature of sentiments associated with a quit episode. The behavior change techniques popular before quit were 'goals and planning' and 'self-belief' and after quit were 'feedback and monitoring' and 'goals and planning'. Communication genres popular before quit were 'family and friends' and 'quit readiness', whereas focus on 'traditions', 'quit progress' and 'quit obstacles' was high after quit. Implications for development of real-time interventions that are mindful of emotional and informational support are discussed.

Keywords: Behavior change · Smoking cessation · Text analysis

1 Introduction

Tobacco use causes harm in nearly every organ of the body and is still the leading cause of preventable death [1]. Tobacco use also affects the mood of smokers and research shows that depression is common among smokers than among non-smokers [2]. Smoking takes a toll on both aspects of a smoker's life - physical and emotional well-being [1]. As far as emotional well-being is concerned, its relationship with smoking is not unidirectional. Emotional factors such as stress and depression can also trigger smoking habits irrespective of age, gender and behavioral status (such as first-time smokers or relapsers) [3, 4]. The process of quitting can also add to emotional

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 253–262, 2018.
https://doi.org/10.1007/978-3-319-93372-6_29

stress, however, time-varying abstinence has an inverse relation to concurrent levels of depressive symptoms [5]. Considering the complex nature of the quit process, researchers suggest that a smoker's quit journey needs to be mindful of physical and emotional factors [2]. Previous research efforts in this area have focused on studying quality of life when smoking and after cessation [6], factors affecting difference in quality of life between ex-smokers and current smokers [7] and changes in life satisfaction after a prolonged period of quit [8]. Other studies compared successful quitters and relapsers on factors such as self-control [9] and medications [10] leading to their successful or failed attempts. However, previous research has not focused on studying a successful quit attempt from the viewpoint of emotional well-being at various windows of time before and after a quit event. Recent advances in unstructured data analytics can provide insights into this issue in the context of social interactions in online platforms [11]. Online health communities are gaining popularity as behavior change venues. These platforms can provide valuable insight into the emotional aspects of users undergoing the process of behavior change, as mood manifestations are ecological and organic among peer interactions [12]. From such peer interactions, which are unprompted and longitudinal in online communities, we gain a spontaneous snapshot of the users' journey across smoking statuses. Specifically, it allows researchers to follow users' progress through time to understand their behavior change efforts which eventually lead to a successful or failed attempt. Interventions addressing just-in-time informational, behavioral and emotional needs of users can be developed by understanding these underlying behavioral and emotional trends. The overarching aim of this study is to follow users of an online health community for smoking cessation to analyze their successful quit attempts and underlying emotional attributes embedded in their communication content. Our specific objectives are to characterize user behavior from the following perspectives: (a) user engagement: determined based on post frequency, (b) content-sensitive characterization of quit episodes, and (c) emotional features embedded in messages exchanged between these users. Studying these three factors offers insights into the dynamics of emotional and communication attributes underlying abstinence from tobacco use. Patterns, if any, observed in terms of user engagement and content-specific emotional attributes will help inform researchers of the possible interventional solutions that can amplify the efficacy of traditional cessation infrastructure.

2 Materials and Methods

QuitNet is one of the first online health communities for behavior change and has been in continuous existence for the past 20 years [13]. Each QuitNet forum message has a message id, a thread id (the thread in which the message was exchanged), a sender id and recipient id. Each user also has a record of their quit events outlining abstinence status and time stamp associated with a quit episode. From this, each user was categorized into one of the following behavioral categories during the study period.

1. Smoker – A user whose abstinence status remained '0'.
2. Ex-smoker – A user whose abstinence status remained '1'.

3. Successful quitter – A user whose abstinence status changed from ‘0’ to ‘1’.
4. Relapser – A user whose abstinence status changed from ‘1’ to ‘0’.
5. Flipflopper – A user whose abstinence status changed multiple times.

For the purpose of this study, we considered the successful quitters and their quit journey from the year 2014. The total number of forum messages available during the study period are 65, 910 and the number of unique users are 2,354.

2.1 QuitNet User Engagement

The user engagement was estimated from forum post frequency of the successful quitters during the study period. The temporal base point was the quit event. The number of messages exchanged by this group of users were estimated at different time points ranging from two days to ninety days before and after a quit event.

2.2 Content-Specific Characterization of QuitNet User Interactions

We conducted qualitative coding to describe the manifestation of behavior change techniques and communication themes embedded in the messages exchanged by QuitNet users. The first step of this coding process was guided by the Taxonomy of behavior change techniques [14]. 2000 messages were randomly selected using a random number generator. Each of these messages were manually coded to the 16 techniques of the behavior change taxonomy by two independent coders. The techniques, their definition and sub-categories are available in [14]. The inter-rater reliability was estimated to ensure objectivity of the coding process. In this step, we disregarded eight of the 16 techniques of the taxonomy due to insufficient positive examples (less than 8% positive examples). The themes that were considered for further analysis were Goals and Planning, Feedback and Monitoring, Social Support, Natural Consequences, Comparison of Behavior, Comparison of Outcomes, Rewards and Threat and Self-belief. It is essential to note that a single message on QuitNet may be assigned to multiple techniques of the taxonomy. Further, we mapped these techniques to communication themes (e.g. Cravings, Traditions) derived using grounded-theory based analysis of QuitNet messages. The themes capture behavioral, interpersonal and individualistic concepts in QuitNet communication [15, 16].

Scaling up to QuitNet Dataset

Vector Generation. In order to scale our methods to large datasets, word representation techniques were used in conjunction with machine learning algorithms. The word representation techniques were used to derive implicit meaning in communication. Specifically, a method called the neural word embedding was implemented using the Skipgram with Negative Sampling Algorithm [17]. This process was implemented using the open source package called Semantic Vectors [18]. This algorithm helps identify meaning based on the context in which it is present. Wikipedia corpus containing 1.9 billion words and 4.4 million articles was used as a background corpus to provide sufficient context to the QuitNet messages. A stopword list [19] was used to eliminate commonly used words in the English language which do not offer semantic information. The neural Word Embedding algorithm was applied to the Wikipedia corpus to create input and output weights. The input and output weights were created for each word of the corpus by using a sliding window to create a context of terms. These vectors were then trained on the QuitNet corpus to obtain input and output weights for each term of the corpus and each message of QuitNet. This resulted in "QuitNet message vectors" which were obtained by using a message vector to predict surrounding messages. The vectors thus generated were of dimension 500.

Machine Learning Classification. The resulting vectors were then used as attributes for the machine learning classifier. An open source tool called Weka [20] was used for this purpose. For each of the vector, the 500 dimensions were considered as individual attributes for the machine learning algorithm. A binary classifier was implemented using each of the technique of the taxonomy as a target for classification. The specific classifier used for this purpose was the Random Forest classifier [21]. The classifier was trained using 2000 manually coded messages. Further, a ten-fold cross validation was conducted to ensure accuracy. The trained model was then used to classify the entire dataset of 65,910 messages used in the study.

Emotional Index. In order to calculate the emotional index of the successful quitters from their messages, a guide called LabMT word list [22] was used. The LabMT word list was created by a team of researchers by drawing on the most frequent words from sources such as Twitter and Google Books [23]. The researchers obtained emotion ranking of these words using Amazon's Mechanical Turk [24]. The LabMT words' emotion ratings range from 1.3 to 8.5 (least happy to most happy). Sample words from the list and their corresponding emotion score is listed in Table 1.

Table 1. Sample words from the LabMT word list and the corresponding emotional score

Sample LabMT word	Emotional score
Laughter	8.5
Smoked	4.86
Depressed	2.18
Suicide	1.3

For the purpose of this study, in order to observe emotional index of the users, the LabMT terms were grouped together into seven groups of emotional index based on their emotional score. In order to match each of the QuitNet messages to a specific emotional index group on a semantic level, word representation techniques were used. In a similar process outlined in the previous section, vectors of dimension 500 were generated for each of the emotional index groups, which we call "QuitNet emotions vectors". We then calculated semantic similarity score (cosine product) between QuitNet message vectors and QuitNet emotions vectors. The emotional index with the highest similarity score was assigned to each of the QuitNet messages.

3 Results and Discussion

The percentage of successful quitters during the study period was 30% and contributed 19% of total messages of successful quitters. 70% of the successful quitters were female users with an average age of 42.2. The average age of the male quitters was 44.6.

3.1 User Engagement

Figure 1A shows the total number of messages exchanged at various points of time, 2 days to three months before and after a quit event. Increasing the number of observation days before a quit event amounted to only a small change in proportion (less than 1.2%). However, after a quit, increasing the number of days amounted to increases as high as 27%. At even the lowest number of days of observation (2 days), the increase in proportion before and after a quit event was as high as 3%. These results suggest that successful quitters are highly dependent on the community to sustain their quit. Increase in engagement after a quit suggests that the community offers the necessary informational and emotional support, in addition to motivation to stay quit. Figure 1B shows the proportion of messages exchanged by successful quitters within the points of observation as a percentage of the total messages populated by this group of users during the study period. Within 14 days after a quit, the successful quitters shared 10% of the total messages they would populate through the entire study period. Within three months of their quit, the successful quitters shared 31.7% of the total messages they would populate through the year. Because the successful quitters were highly active in the community after a quit event, the proportion of messages is higher after the quit rather than before.

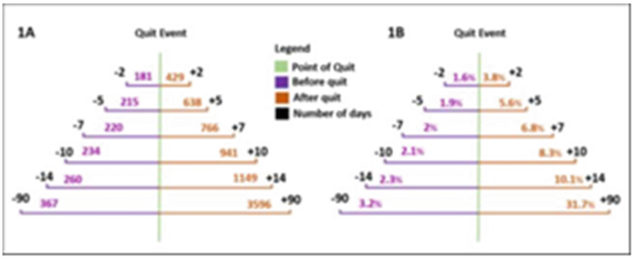


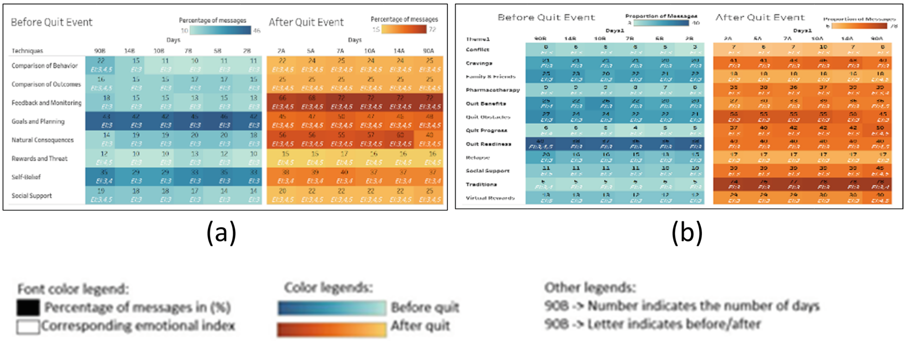
Fig. 1. User engagement at various points in time before and after a quit event

In summary, in terms of user engagement, a characteristic trait of successful quitters is that they stay active in the community, especially within a three month window after a quit event.

3.2 Content-Specific Characterization of QuitNet User Interactions

Assignment of Taxonomy Classes. The inter-rater reliability estimated for the two independent manual coders was 0.76. The precision, recall and f-measures of the automated classification were 0.81, 0.8 and 0.8 respectively.

Mapping to Taxonomy of Behavior Change Techniques. Figure 2a shows the distribution of messages in each technique of the behavior change taxonomy at various time points before and after a quit event. As can be seen from the figure, the technique with the highest presence before quit were the techniques ‘Goals and Planning’ and ‘Self-Belief’ with a proportion of 42% and 33% at the three month time window. In their messages, QuitNet users set goals to quit and made specific plans to quit. The themes with the least presence before quit were ‘Rewards and Threat’ and ‘Comparison of Behavior’ with a proportion of 10% and 11% respectively at the three month time window. This can be attributed to the fact that users at this point were focused on themselves, their past and future quit journey.



(a)

(b)

Fig. 2. Distribution of messages in the taxonomy of behavior change techniques and grounded theory based communication themes before and after a quit

After a quit event, the technique most popular among users were ‘Feedback and Monitoring’ and ‘Goals and Planning’ with a proportion of 72% and 48% at the three month time window. Users after a quit typically ended their messages with their quit statistics such as the number of days it has been since their quit, number of unsmoked cigarettes which contributed to the technique ‘Feedback and Monitoring’. Users also participated in traditions such as ‘taking pledges to not smoke for the day’, which contributed to the technique ‘Goals and Planning’. As seen from the figure, the technique ‘Natural Consequences’ had a steep fall (20%) after the 14 day window. This can be attributed to the fact that users’ health consequences alleviate with time.

Mapping to Grounded Theory Themes. Figure 2b shows the average distribution of messages in each theme before the quit event at various time windows. The highly popular themes in the messages exchanged among successful quitters were ‘Quit Readiness’ followed by ‘Quit Obstacles’, ‘Family and Friends’ and ‘Quit Benefits’ with a distribution of 40%, 27%, 25% and 25% respectively at the three month window before a quit. Users discussed their past quit attempts, the reason for their previous relapses, and how it had affected their family and friends. They also reassured themselves that this quit event would be successful especially because of their learnings from their past attempts. The themes that were least popular before a quit were 'Traditions' and 'Quit Progress' with a distribution of 6% and 6% at the three month window. Since users have not participated in the quit at that point, they had not participated in any of the traditions or 'Quit Progress' related discussions. The proportion of messages in the theme 'Conflict' was also low (8% at three month window) before quit since users were relatively new to the community and hence were not socially situated to have conflicts with other users.

Figure 2b also shows the average distribution of messages in each theme after the quit event at various time windows. After a quit, the messages with the highest proportion were 'Traditions' followed by 'Quit Progress' with an average proportion of 78% and 50% respectively at the three month window. The theme with lowest proportion after quit was 'Conflict' followed by 'Relapse' with a distribution of the 13% and 17% respectively at the three month window. Users did express concerns about possible failure and relapse thus contributing to 17% of the messages. The theme with the highest increase in proportion after quit event was the theme 'Traditions' with an increase of 69% followed by 'Quit Obstacles' with an increase of 35%. The emotional indices were linked to the techniques and themes and as can be seen from Fig. 3. Detailed discussion of emotions embedded in users' messages can be found below in Sect. 3.3.

3.3 Emotional Index

Figure 3 describes the emotional state of QuitNet's successful quitters at different time points during the study periods. Well before a quit (three months before) majority of users messages reflected a low emotional score. The low scores before a quit may have stemmed from the discussions surrounding previous failed quit attempts and how their smoking/failed quit attempts had affected their families. However, users are also indulged in positive, motivated self-talk which is reflected in a portion of messages that belonged to a higher emotional score.

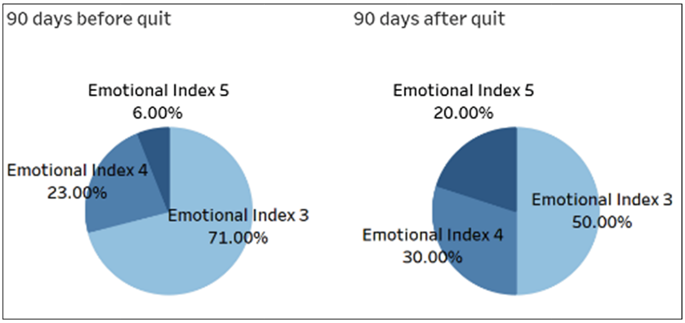


Fig. 3. Emotional Index distribution three months before and after a quit event

Low emotional index after a quit can be due to tobacco users' anger, anxiety, depression, impatience, insomnia and restlessness peak within the first week and last about 2–4 weeks since quit after which they alleviate [25]. The following excerpts 'I do relate with smoking to relieve tension', 'Still angry, tired & upset' and 'let s call it kind of a rage' and 'I am angry at everyone about everything' illustrate the complex and fragile emotions during a quit episode.

In summary, using content to understand behavior change techniques, intercourse themes, and emotional index of successful quitters has offered us valuable insight into specific emotional and informational needs during a quit episode. Specifically, before a quit, when the users are still smoking, they are emotionally weak and their behavioral attributes suggest they plan their quits and are reminding themselves of the specific reason of their quit, how their smoking has affected their family and are anxious to maintain their quit. Immediately after a quit (starting from quit up to 4 weeks), quitters are overwhelmed by the natural withdrawal effects – physically and emotionally. However, well after a quit event (three months after), users seem to have been relieved of their withdrawal effects.

4 Limitations and Future Work

Our analysis is limited to a specific category of users called successful quitters. In the future, we can expand our studies to compare and understand the differences between successful quitters and other groups of users such as relapsers and active smokers. Understanding the characteristic traits of such users can be used to offer targeted interventional support for groups of users that enables sustained quit. The automated classification is performed on a relatively small social media sample and should be extend to voluminous datasets to establish accuracy at scale.

The emotional index computations are performed using an existing corpus, however, this analysis can be further improved when conducted at atomic levels to gain insights into state of mind and respective triggers. This can eventually lead to offering personalized interventions that offer support based on emotional index of a user at a specific point in time. Also, our analysis is limited to understanding semantics in social intercourse, future studies should focus on network analysis which to understand the spread of emotions. Further, this can lead to understanding of behavior-specific group dynamics and community features.

5 Conclusions

This study focuses on analyzing peer-to-peer communication to understand the journey of successful quitters in an online community for smoking cessation. Engagement levels, social communication attributes, behavior change techniques, and emotional features of successful quitters can prove to be useful to design behavior change systems to sustain long term positive health changes. The resulting systems can lead to targeted engagement, nuanced support triggers, and automated recommendation engines, which can help us transform online social platforms into affordable precision wellness technologies.

Acknowledgements. Research reported in this publication was supported by the National Library of Medicine and National Cancer Institute of the National Institutes of Health under Award Numbers R21LM012271 and R21CA220670. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. U.S. Department of Health and Human Services: The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health (2014). Accessed 20 Apr 2017
2. Smokefree.gov. http://smokefree.gov. Accessed 21 Dec 2017
3. Steptoe, A., Wardle, J., Pollard, T.M., Canaan, L., Davies, G.J.: Stress, social support and health-related behavior: a study of smoking, alcohol consumption and physical exercise. J. Psychosom. Res. **41**(2), 171–180 (1996)
4. Glassman, A.H.: HElzer, J.E., Covey, L.S.: Smoking, smoking cessation, and major depression. J. Am. Med. Assoc. **264**(12), 1546–1549 (1990)
5. Kahler, C.W., Spillane, N.S., Busch, A.M., Leventhal, A.M.: Time-varying smoking abstinence predicts lower depressive symptoms following smoking cessation treatment. Nicotine Tob. Res. **13**(2), 146–150 (2011). https://doi.org/10.1093/ntr/ntq213
6. Sarna, L., Bialous, S.A., Colley, M.E., Jun, H.J., Feskanich, D.: Impact of smoking and smoking cessation on health-related quality of life in women in the Nurses' Health Study. Qual. Life Res. **17**(10), 1217–1227 (2008)
7. Mulder, I., Tijhuis, M., Smit, H.A., Kromout, D.: Smoking cessation and quality of life: the effect of amount of smoking and time since quitting. Prev. Med. **33**(6), 653–660 (2001). https://doi.org/10.1006/pmed.2001.0941
8. Piper, M.E., Kenford, S., Fiore, M.C., Baker, T.B.: Smoking cessation and quality of life: changes in life satisfaction over 3 years following a quit attempt. Ann. Behav. Med. **43**(2), 262–270 (2012)
9. Katz, R.C., Singh, N.N.: Reflections on the ex-smoker: some findings on successful quitters. J. Behav. Med. **9**(2), 191–202 (1986)
10. Gonzales, D., Jorenby, D.E., Brandon, T.H., Arteaga, C., Lee, T.C.: Immediate versus delayed quitting and rates of relapse among smokers treated successfully eith varenicline, bupropion SR or placebo. Addiction **105**(11), 2002–2013 (2010)
11. Sridharan, V., Cohen, T., Cobb, N., Myneni, S.: Characterization of temporal semantic shifts in peer-to-peer communication in a health-related online community: implications for data-driven health promotion. In: AMIA Annual Symposium Proceedings, Chicago (2016)
12. Zhang, M.: Social media analytics of smoking cessation intervention: user behavior analysis, classification, and prediction. Drexel University, Pennsylvania (2015)
13. QuitNet LLC. https://quitnet.meyouhealth.com/. Accessed 28 Nov 2017 [WebCite Cache]

14. Michie, S., Richardson, M., Johnston, M., et al.: The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting change interventions. Ann. Behav. Med. **46**(1), 81–95 (2013)
15. Myneni, S., Fujimoto, K., Cobb, N., Cohen, T.: Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks. Am. J. Publ. Health **105**(6), 1206–1212 (2015)
16. Myneni, S., Cobb, N., Cohen, T.: In pursuit of theoretical ground in behavior change support systems: analysis of peer-peer communication in health related online community. J. Med. Internet Res. **18**(2) (2016)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS 2013, pp. 3111–3119 (2013). [21]
18. Widdows, D., Cohen, T.: The semantic vectors package: new algorithms and public tools for distributional semantics. In: Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing, ICSC 2010, pp. 9–15 (2010)
19. Salton, G., McGill, M.J.: Introduction to Modern Information Retrieval. McGraw-Hill, New York (2013)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explor. **11**(1), 10–18 (2009)
21. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
22. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE **6**(12), e26752 (2011). https://doi.org/10.1371/journal.pone.0026752
23. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., et al.: Quantitative analysis of culture using millions of digitized books. Sci. Mag. **331**, 176–182 (2011)
24. Amazon's Mechanical Turk Service. Amazon's Mechanical Turk Service. https://www.mturk.com/. Accessed 24 Oct 2017
25. Hughes, J.R.: Effects of abstinence from tobacco: valid symptoms and time course. Nicotine Tob. Res. **9**(3), 315–327 (2007)

Digilego: A Standardized Analytics-Driven Consumer-Oriented Connected Health Framework

Sahiti Myneni(✉), Deevakar Rogith, and Amy Franklin

The University of Texas School of Biomedical Informatics, Houston, TX, USA
sahiti.myneni@uth.tmc.edu

Abstract. Connected health solutions provide novel pathways to provide integrated and affordable care. Emerging research suggests these connected tools can result improved health outcomes and sustainable self-health management. However, current health technology frameworks limit flexibility, engagement, and reusability of underlying connected health components. The objective of this paper is to develop a data-driven consumer engagement framework, which we call Digilego, to facilitate development of connected health solutions that are targeted, modular, extensible, and engaging. The major components include social media analysis, patient engagement features, and behavioral intervention technologies. We propose implementation of these Digilego components using FHIR specification such that the resulting technology is compliant to industry standards. We apply and evaluate the proposed framework to characterize four individual building blocks (DigiMe, DigiSocial, DigiConnect, DigiEHR) for a connected health solution that is responsive to cancer survivor needs. Results indicate that the framework (a) allows identification of survivor needs (e.g. social integration, treatment side effects) through semi-automated social media analysis, (b) facilitates infusion of engagement elements (e.g. smart health trackers, integrated electronic health records), and (c) integrates behavior change constructs into the design architecture of survivorship applications (e.g. goal setting, emotional coping). End user evaluation with 16 cancer survivors indicated general user acceptance and enthusiasm to adopt the solution for self-care management. Implications for design of patient-engaging chronic disease management solutions are discussed.

Keywords: Connected health · Consumer informatics · Chronic disease

1 Introduction

Connected health ecosystems have revolutionized care delivery and patient engagement in chronic disease management [1]. Several consumer-facing health technologies, such as physical activity trackers, health journals, social logs, have become indispensable components of care coordination playing a crucial role in health and wellness infrastructure. Managing the appropriate integration of these new era consumer-driven data-intensive artifacts into traditional health care ecosystem is vital to exploit the positive effects of connected health, while simultaneously addressing data quality

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 263–273, 2018.
https://doi.org/10.1007/978-3-319-93372-6_30

issues and vulnerability threats. Further, such integration should be flexible and adaptive to consumers' personal preferences and privacy concerns. Although, there are several technology development frameworks in health care that focus on infusing health technologies with theoretical constructs, behavior change techniques, usability features, and device automation [2, 3], the majority of existing frameworks do not address the issue of care management in connected health ecosystem from a health consumer's perspective. The risk of a data deluge from these heterogeneous, siloed, distributed technology components can result in a chaotic health data repository. This resulting complexity can overwhelm a general health consumer and result in suboptimal knowledge management and self-health management. Given the complexity of decision making associated with chronic care management, the new technology revolution facilitated by connected health should ease the burden of personal health management. To address these challenges, we present a novel framework, *Digilego*, for the development of consumer-facing connected health applications that effectively integrates personal demographics, clinical data from electronic health records, and personal health data from wearables and home-based monitoring systems, while accounting for consumers' personal preferences in care management. The *Digilego* framework integrates the Social Media Analysis (SMA) [4, 5], Patient Engagement Framework (PEF) [6], Behavioral Intervention Technology (BIT) model [7], and Fast Healthcare Interoperability Resources (FHIR) Specification [8, 9], a standard for exchanging healthcare information electronically. Research shows the use of SMA effective in the design of targeted health applications that are consumer-centered focus, specifically for the purpose of self-monitoring and behavior change in chronic disease management [4]. SMA facilitates capture of culturally sensitive expressions of consumers' information needs in social platforms at scale [4]. Subsequently, PEF acts as a bridging tool to identify opportunities that can help translate the knowledge abstracted within a disease domain to engagement elements of a consumer-centered application. The BIT model allows instantiation of individual *digilego* blocks as technological constituents of the intended connected health solution. Finally, the FHIR specification facilitates implementation of the *digilego* in a web-based user interactive form. In the next sections of the paper, we describe the framework components along with illustrative examples demonstrating its application to the domain of cancer survivorship. Specifically, we focus on facilitating the development of a connected health solution for cancer survivorship, an important chronic health condition [10] through the *digilego* blocks.

2 Methods

Figure 1 provides a high-level overview of our proposed *Digilego* framework, which facilitates the development and integration of individual *digilego* building blocks to form a consumer-facing connected health solution. The four main components are (1) content-inclusive social network analysis, (2) engagement elements of *digilego* blocks, (3) feature development using the BIT model, and (4) implementation of standard-compliant software features using FHIR specification.

2.1 Digilego Development Framework

Social Media Analysis: As part of our needs analysis for Digilego framework, we analyzed 24,723 publicly available deidentified peer interactions in an online community devoted to cancer survivors. We conducted qualitative analysis of 1000 messages selected at random using a random number generator to gain insights into survivors' sociobehavioral and information needs. The analysis also helped us identify communication topics related to care management of a cancer survivor. We assigned each message to one or more communication topic ranging from treatment discussions to insurance management and medical wills. These qualitative codes were scaled up to the entire dataset by using random indexing approaches found in Semantic Vectors package [11]. The resulting vectors were used as features of machine learning classifiers within Weka [12], an open source text analysis software. Multiple classifiers (Nearest neighbor, J48, Random forest, Naïve Bayes, Support Vector Machine) were compared and we have chosen the best performing classifier based on accuracy metrics (F-measure). The resulting fully annotated dataset was then used to conceptualize individual Digilego blocks.

Digilego Engagement Elements: PEF has been developed by Healthcare Information and Management Systems Society (HIMSS) through cumulative layering of five phases- "inform me," "engage me," "empower me," "partner with me," and "support my e-community." A total of nine features have been specified at the highest engagement level, including 'information and way-finding', 'e-tools', 'forms', 'patient-specific education', 'patient access and use', 'patient-generated data', 'interoperable records', 'collaborative care and community support' [6]. This framework has been used in our study to define the functionality of individual Digilego blocks that form our proposed connected health solution to facilitate self-management, goal setting and reinforcement, peer support, and patient-provider communication. A mapping process was conducted to identify the engagement features which were used to operationalize the insights from SMA in prior step. The lessons learned in the previous step allowed us to understand the characteristics of the information that should be delivered to the survivor through PEF features. We also specified the level of intended survivor engagement to characterize granularity and complexity of the system features.

Digilego Development: We used BIT model to conceptualize digilego blocks (reusable analytics-driven connected health components) to delineate the operational aims, identify behavior change strategies (where applicable), define user interactions, and outline technical aspects for real-time implementation. The BIT model provides a framework for articulating the relationship between intervention aims, elements, characteristics, and workflow [7]. The BIT model was originally proposed to develop behavioral interventions, however, all digilego blocks did not have a behavioral component associated with them (e.g. DigiEHR, DigiMe). For the non-behavioral ones, we still used BIT model to ensure workflow alignment and smoother interfacing among digilego blocks.

Digilego Software Implementation: We developed HTML prototype of implementation of the digilego. The HTML prototype is a responsive design that can scale based on the digilego specific to the user. The prototype was powered by Fast Healthcare Interoperable Resources (FHIR®) compliant server-side app [8]. We used Model View Control (MVC) approach to design the web app. The view comprises of the digilego groups under four domains – Profile (DigiMe), Clinical data (DigiEHR), Sensor and personal health device integration (DigiConnect), and Social engagement (DigiSocial). The controllers were HTTP requests using FHIR specifications to manipulate the data and the view. Specifically for DigiMe and Digi EHR, we used resources that are specified under HL7 FHIR DSTU 2 (version 1.0.2) in alignment with implementation of commercial EHR vendors in the United States. For DigiConnect and DigiSocial that did not have existing FHIR resources, we used data model based on schema.org, compliant with FHIR extensions.

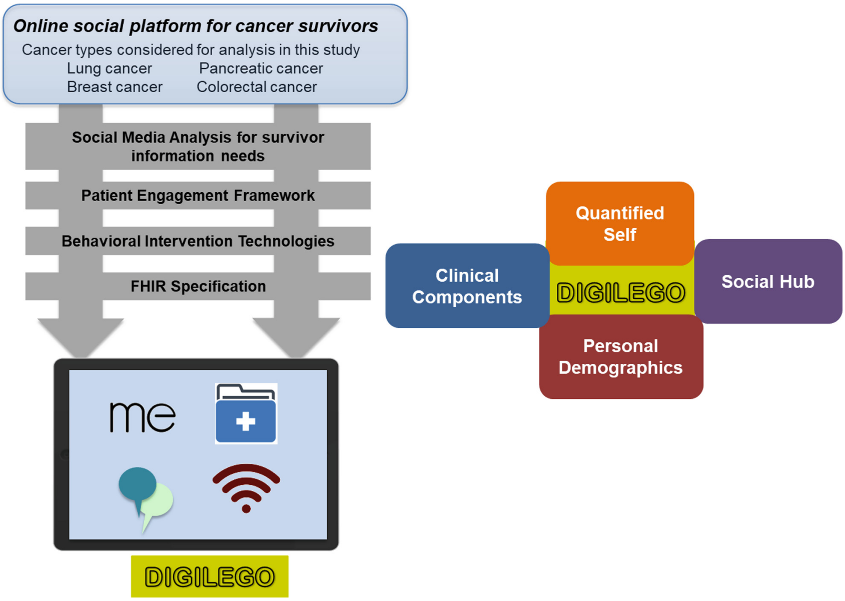

Online social platform for cancer survivors
Cancer types considered for analysis in this study
Lung cancer
Pancreatic cancer
Breast cancer
Colorectal cancer
Social Media Analysis for survivor information needs
Patient Engagement Framework
Behavioral Intervention Technologies
FHIR Specification
me
DIGILEGO
Quantified Self
Clinical Components
DIGILEGO
Social Hub
Personal Demographics

Fig. 1. Proposed development framework for cancer survivor digilego

2.2 Digilego Evaluation

Having developed four Digilego blocks in the context of cancer survivorship, we conducted a preliminary evaluation of user perceptions of the technology and underlying framework. Two focus group sessions were conducted with 16 cancer survivors. The focus group participants were shown a web-based mobile-responsive prototype and were asked open-ended questions to understand the levels of technology acceptance, system usability, and perceived usefulness of the proposed design methodology and Digilego environment.

In the next sections, we describe the application of Digilego framework to develop a connected health solution for cancer survivorship using illustrative examples. Such compartmentalized design architecture allowed us to facilitate customization, harness analytics-mediated knowledge, integrate engagement tactics, and adopt theory-driven techniques that ultimately result in targeted engagement of cancer survivors throughout the cancer care continuum.

3 Results and Discussion

3.1 Application of Digilego Framework to Cancer Survivorship Domain

Social Media Analysis: Initial qualitative analysis revealed 15 communication topics, which are distributed in the study dataset as follows.

(a) 39% of the messages were related to (a) care management and coordination: treatment plan discussions (complications, recovery time), medication questions, and request for guidelines to structure upcoming physician appointments,

(b) 36% of the messages were related to social integration, emotional support, positive mindfulness, and support groups,

(c) 18% of the messages were related to healthy living, remission prevention, and late effects of cancer treatment,

(d) 4% of the questions were related to monetary topics (fundraising, insurance limits, exclusions), feedback on care providers and treatment centers, and

(e) the remaining 3% of the messages were related to medical wills and hospice care.

The automated text analysis methods revealed that Random forest classifier resulted in the optimal F-measure of 0.66. The following automated analysis resulted in a completely annotated dataset. Figure 2 shows the distribution of the communication topics in the entire corpus.

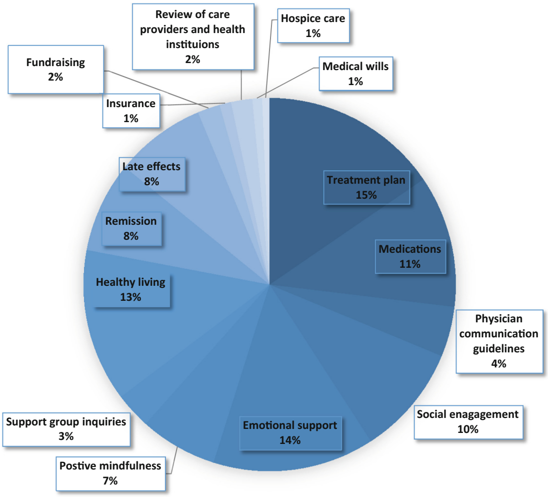


Fig. 2. Distribution of the communication topics in cancer survivorship dataset

Digilego Development, Optimization, and Implementation: Four distinct digilego blocks (see Table 1) were conceived and prototyped. Each of the blocks are further segmented to meet the granular information needs that have manifested in the social media interactions. These Digilego blocks are responsive to 12 of the 15 social media topics extracted using automated text analysis methods described in the earlier sections of the paper. For instance, around 13% of the messages focused on adoption of cancer prevention behaviors (e.g. smoking cessation, physical activity, stress management). The Digiconnect and DigiSocial modules are designed considering these interactions.

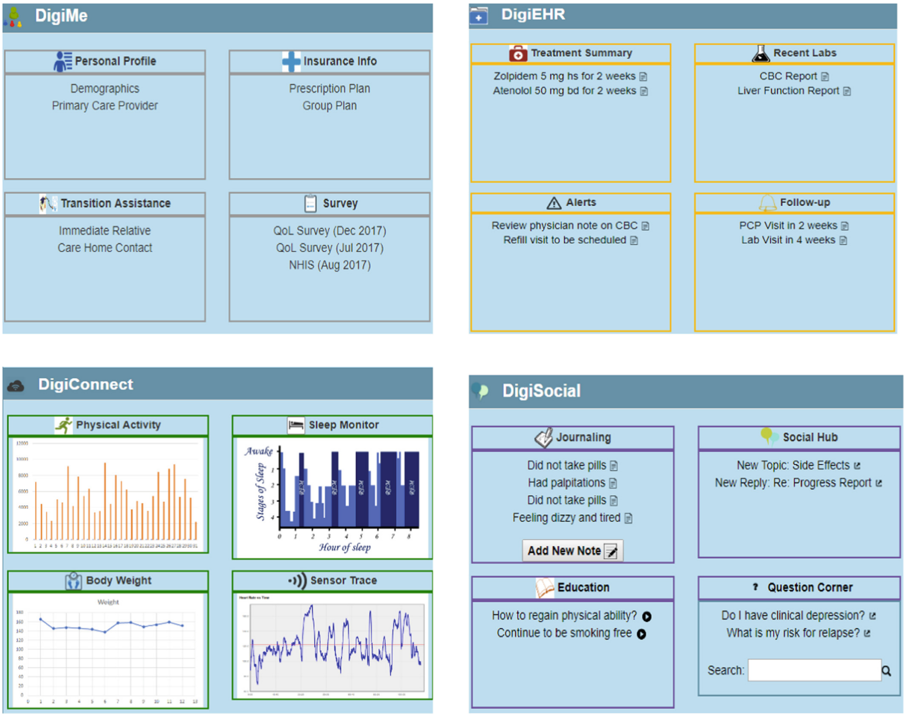
Table 1. Cancer Survivorship Digilego components and related social media topics

Digilego	Digilego	Social media topics
DigiMe Allows management of administrative and personal information pertinent to cancer survivorship	👤	Insurance Personal summary Caregiver profile Agenda for physician appointments Surveys and review submissions
DigiSocial Connects survivors with peers, care providers while also enabling journal writing of their efforts to stay healthy	💬	Physician communication Social engagement Emotional support Positive mindfulness Support group inquiries
DigiEHR Provides a recent copy of survivor health information from physician's EHR as related to the stage of cancer survivorship	✚	Treatment summary Recent labs Phsycian communication
DigiConnect A snapshot of objective sensor measurements from personal health devices	☁	Healthy living Physician communication Positive mindfulness

The engagement elements identified using the PEF framework for feature selection of Digilego blocks are provided in Table 2. For example, the DigiEHR module is fitted with Empower (Level III) Integrated form: EHR. Similarly, DigiConnect employs Engage (Level II) e-Tools through behavioral trackers and external sensors. For the purpose of our case study on cancer survivorship, Digilego functionalities and interrelations are modeled as behavioral intervention technologies by defining the overarching operational intention, which is to promote self-management of cancer care and survivorship through adaptive means, as survivor needs change with the stage of cancer continuum. Examples of the sub-goals (see Fig. 3) for each component include increasing positive health behaviors, informing survivors of the late effects of cancer treatment, and promoting adherence to follow-up care regimen. While the operational intention for each *digilego* is unique depending on the content specialty, the usage intention is same across all components where the aim is to engage survivors in self-management of their health. Henceforth, social media features that connects survivor to care providers and peers, and personalization features (age-specific, cancer-specific, stage-specific educational material) have been integrated to the design framework of *digilego* to promote user engagement. Furthermore, DigiSocial and DigiConnect operationalize multiple theoretical constructs that range feedback and self-monitoring via e-health tools such as health behavioral trackers, social support, observational learning [7]. Mapping these strategies to digilego elements is straight-forward, given the clear formulization of the intentions of each component and strategy using BIT model. Further, the instantiation criteria for each of the digilego are defined in terms of the interaction features and workflow. Notifications, logs, information delivery are the most used interaction elements. For example, notifications and logs are assigned to "DigiMe" for review submissions, agenda generation, and survey responses. Similarly, messaging elements and visualizations are used for Digi Social and DigiConnect respectively to provide users with (a) communication tools to interact with peer and care providers, and (b) consolidated feedback to users on their healthy living indicators. Event-based and time-based workflow criteria are used to derive personalization effects. For instance, consider a transition in the cancer care continuum from diagnosis to treatment, education materials related to the treatment type (e.g. chemotherapy, mastectormy) are delivered depending on cancer stage and cancer type. Similarly, time-based workflow is defined for "DigiMe" and "DigiConnect", where customizations will come to effect depending on time interval since last survey response and goal setting. Figure 3 shows the prototypes for the four Digilego and their underlying information components. These prototypes are FHIR-compliant, extensible, and truly connected from a health consumer's perspective. The four Digilego components integrate clinical, personal, administrative and social facets of survivor care management. The interface supports interactions such as view, edit, search, write, etc., depending on the digilego. The DigiEHR and DigiConnect are read-only interactions, and enable related entries in DigiSocial or DigiMe. For instance, a survivor can navigate from DigiEHR to DigiConnect to resolve medication-related questions by intetracting with peers and care providers.

Table 2. Cancer survivorship Digilego and engagement elements

Engagement phase	Features	Related Digilego
Engage (Level II) • Behavioral trackers • Physiological sensors • Survivor profile • Insurance information • View components of electronic health record	e-Tools Interactive tools Patient access: records	DigiConnect DigiMe DigiEHR
Empower (Level III) • Care experience surveys • Self-management diaries	Patient-generated health data	DigiMe DigiSocial
Partner with me (Level IV) • Home monitoring • Condition-specific self-management tools	Patient-generated data Patient-specific education	DigiConnect DigiSocial
Support my e-community • Online support forums	Community support	DigiSocial



DigiMe

Personal Profile: Demographics; Primary Care Provider

Insurance Info: Prescription Plan; Group Plan

Transition Assistance: Immediate Relative; Care Home Contact

Survey: QoL Survey (Dec 2017); QoL Survey (Jul 2017); NHIS (Aug 2017)

DigiEHR

Treatment Summary: Zolpidem 5 mg hs for 2 weeks; Atenolol 50 mg bd for 2 weeks

Recent Labs: CBC Report; Liver Function Report

Alerts: Review physician note on CBC; Refill visit to be scheduled

Follow-up: PCP Visit in 2 weeks; Lab Visit in 4 weeks

DigiConnect

Physical Activity

Sleep Monitor

Body Weight

Sensor Trace

DigiSocial

Journaling: Did not take pills; Had palpitations; Did not take pills; Feeling dizzy and tired; Add New Note

Social Hub: New Topic: Side Effects; New Reply: Re: Progress Report

Education: How to regain physical ability?; Continue to be smoking free

Question Corner: Do I have clinical depression?; What is my risk for relapse?; Search:

Fig. 3. FHIR-compliant cancer survivorship Digilego prototypes

3.2 Preliminary Evaluation of Cancer Survivorship Digilego

All participants expressed confidence giving in the proposed design methodology giving it a rating of ‘5’, indicating they strongly believe the proposed system will improve the quality of my cancer care management. 75% of the participants indicated that use of social media analysis in the development process increased their belief in the system’s ability to assist them in care management and specifically mentioned it played a role in their acceptance of the proposed technology platform. 87.5% of the participants voted favorably for all the four Digilego bloacks, while the remaining 12.5% felt communication with care provider team and caregivers (family) warrants two separate Digilego blocks. 100% of the participants preferred a mobile platform to a web-based system. Overall, the design philosophy of Digilego that allows cancer survivors to tailor the features of their care assistant technology is favorable received.

In summary, a cluster of connected health components have been conceived, defined, characterized, prototyped using Digilego, an integrative multifaceted standard-compliant framework. Initial evaluation indicated high rates of user acceptance of the prosed development framework.

4 Limitations and Future Steps

Our study incorporated design elements from user perspective alone. Future work should focus on expert advisory board to ensure coverage of information needs that are not captures through social media analysis. The automated text analysis methods can be refined to include background corpus and sophisticated machine learning algorithms and distributional representations to improve accuracy of automated classification system [13]. Future work should integrate ontology models [14] with social media analytics to ensure inclusion of knowledge from existing literature into Digilego architecture. Some digilego blocks may not have existing FHIR resources, and custom FHIR compliant extensions/schemas should be generated and validated. The evaluation is limited to user perceptions of system utility. Further studies should investigate usability, performance, and effectiveness of Digilegos in care management. Finally, the four Digilego components form a preliminary proof–of-concept and do not offer full coverage of chronic disease domains such as cancer. Future research should focus on the development of a Digilego bank for a given disease profile from which a health consumer can build a customized connected health tool for self-health management.

5 Conclusion

Connected health solutions are becoming increasingly popular in healthcare. However, the current ecosystem lacks design methodologies that are theory-driven, standard-compliant while considering health consumer preference during the design and field implementation stages. This paper presents an approach that aims at the development of integrated consumer-facing connected health solutions. Chronic diseases (e.g. cancer, diabetes) are lifelong endeavor for patients, their families, and caregivers. The Digilego framework is a foundational step that will help influence the development of connected health applications with reusable and customizable components as per the needs of the health consumers.

Acknowledgement. Research reported in this publication was supported by the National Library of Medicine and National Cancer Institute of the National Institutes of Health under Award Numbers R21LM012271 and R21CA220670. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Caulfield, B.M., Donnelly, S.C.: What is connected health and why will it change your practice? QJM: Int. J. Med. **106**(8), 703–707 (2013)
2. Harte, R.P., Glynn, L.G., Broderick, B.J., Rodriguez-Molinero, A., Baker, P., McGuiness, B., O'Sullivan, L., Diaz, M., Quinlan, L.R., ÓLaighin, G.: Human centred design considerations for connected health devices for the older adult. J. Personal. Med. **4**(2), 245–281 (2014)
3. Das, A.K., Goswami, A.: A secure and efficient uniqueness-and-anonymity-preserving remote user authentication scheme for connected health care. J. Med. Syst. **37**(3), 9948 (2013)
4. Myneni, S., Fujimoto, K., Cohen, T.: Leveraging social media for health promotion and behavior change: methods of analysis and opportunities for intervention. In: Patel, V.L., Arocha, J.F., Ancker, J.S. (eds.) Cognitive Informatics in Health and Biomedicine. HI, pp. 315–345. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-51732-2_15
5. Myneni, S., Iyengar, S.: Socially influencing technologies for health promotion: translating social media analytics into consumer-facing health solutions. In: 2016 49th Hawaii International Conference on System Sciences (HICSS), pp. 3084–3093. IEEE, January 2016
6. Patient Engagement Framework. www.himss.org/ResourceLibrary/genResourceDetailPDF.aspx?ItemNumber=28305. Accessed 21 Dec 2016
7. Mohr, D.C., Schueller, S.M., Montague, E., Burns, M.N., Rashidi, P.: The behavioral intervention technology model: an integrated conceptual and technological framework for eHealth and mHealth interventions. J. Med. Internet Res. **16**(6), e146 (2014)
8. Mandel, J.C., Kreda, D.A., Mandl, K.D., Kohane, I.S., Ramoni, R.B.: SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J. Am. Med. Inform. Assoc. **23**, 899–908 (2016)
9. Mandl, K.D., Kohane, I.S.: A 21st-century health IT system—creating a real-world information economy. N. Engl. J. Med. **376**, 1905–1907 (2017)
10. DeSantis, C.E., Lin, C.C., Mariotto, A.B., Siegel, R.L., Stein, K.D., Kramer, J.L., Alteri, R., Robbins, A.S., Jemal, A.: Cancer treatment and survivorship statistics. Cancer J. Clin. **64**(4), 252–271 (2014)
11. Widdows, D., Cohen, T.: The semantic vectors package: new algorithms and public tools for distributional semantics. In: 2010 IEEE Fourth International Conference on Semantic Computing (ICSC), pp. 9–15. IEEE, September 2010
12. Holmes, G., Donkin, A., Witten, I.H.: Weka: a machine learning workbench. In: Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems, pp. 357–361. IEEE (1994)

13. Myneni, S., Fujimoto, K., Cobb, N., Cohen, T.: Content-driven analysis of an online community for smoking cessation: integration of qualitative techniques, automated text analysis, and affiliation networks. Am. J. Public Health **105**(6), 1206–1212 (2015)
14. Bickmore, T.W., Schulman, D., Sidner, C.L.: A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology. J. Biomed. Inform. **44**(2), 183–197 (2011)

Pain Town, an Agent-Based Model of Opioid Use Trajectories in a Small Community

Georgiy Bobashev(✉), Sam Goree, Jennifer Frank, and William Zule

RTI International, Durham, NC, USA
bobashev@rti.org

Abstract. We developed a simulation model to illustrate and evaluate the potential effects of opioid-related policies and interventions at the local (e.g., community) level. In the United States, the opioid epidemic was declared a national public health emergency in 2017 because of extremely large numbers of opioid-related overdose deaths. Overprescription of addictive opioid-based painkillers could lead to physical dependence with subsequent dose increase. Some patients switch to heroin to support their drug habit. The use of high doses of prescription opioids, heroin especially in combination with a more powerful synthetic opioid, fentanyl, can sometimes lead to overdose, which can be lethal. A number of prevention and treatment policies have been proposed and some implemented to fight the epidemic. These policies include prescription drug monitoring programs (PDMP), reduced initial opioid dose distribution of naloxone to counter overdose, medication-assisted treatment of problem users, and tamper-proof pills to prevent noncompliant behavior. The model describes the dynamics of opioid prescription and use in an interconnected community of pain patients. The model simulates individual patients' life trajectories with respect to the use of opioids under different policies. The model includes potential policies based on the overdose and mortality rates of prescription opioid users, the overdose and mortality rates of heroin users, and the number of patients who turn to illicit means to acquire their drugs. Simulation study results show strong effects of naloxone use, very marginal short-term effects of PDMP compliance, and few to no positive effects of tamper-resistant medications on non-child opioid use trajectories.

Keywords: Opioid painkillers · Heroin overdose · Agent-based model
Intervention policy analysis

1 Background

The intertwined epidemics of prescription opioid (PO) and illicit opioid (e.g., heroin and fentanyl) misuse in the United States are driving alarming increases in addiction, accidental opioid overdose, and death. Opioid misuse also leads to decreased productivity, increased crime, and overloaded prisons [1, 2]. Opioid-based pain relievers recently became the most prescribed class of medications in the United States [3, 4]. This increase has been accompanied by dramatic increases in misuse and visits to emergency departments (EDs) [5].

The 2016 National Survey on Drug Use and Health found that 11.8 million Americans reported misusing opioids in the past year—11.5 million who misused prescription pain

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 274–285, 2018.
https://doi.org/10.1007/978-3-319-93372-6_31

relievers and 948,000 who used heroin [6]. An estimated 1.8 million people suffer from pain reliever use disorders, and an estimated 626,000 suffer from a heroin use disorder [6].

Opioid-involved drug poisoning deaths have been increasing for many years, and in 2015 the number was more than twice that in 2005 [7]. Drug poisoning deaths related specifically to heroin or synthetic opioids nearly quintupled from 2005 to 2015 [7]. As indicated by the latest report by the Centers for Disease Control and Prevention (CDC), overdose deaths from POs are not decreasing, and heroin overdose deaths continue to increase sharply, indicating growing numbers of users in need of treatment. In addition, the Surgeon General issued a report classifying substance abuse disorders as a public health problem rather than a criminal justice problem [2].

To combat these consequences the U.S. Congress approved more than $1B in funding for opioid prevention and treatment as part of the Comprehensive Addiction and Recovery Act of 2016, S.524. 114th Congress (2015–2016) [8]. These activities have three goals: addressing the overprescribing of opioids, increasing use of naloxone to avoid overdose deaths, and expanding medically assisted treatments [8, 9]. Achieving these goals is challenging for the complex reasons described below.

The opioid epidemic is complex, and combatting it requires a thorough understanding of its many interacting elements. The epidemic comprises several dynamically changing processes, including overprescription of opioid painkillers, diversion of painkillers and methadone, increased use of heroin and distribution across the United States, transition to heroin among painkiller-dependent individuals, and emergence of illegal distribution of high-potency synthetic opioids [10]. Although the percentage of nonmedical use of POs has remained stable over the years [11], the demographics, location, and method of PO use has rapidly changed, specifically the shift from PO to heroin. The supply of treatment and new medication-assisted therapies lags behind [12, 13].

An effective combination of policies requires developing a comprehensive solution strategy. The effects of treatment and prevention policies are not independent of each other. Although prevention is intended to ease the future burden on treatment, policy-resistant behavior could actually increase or alter the expected future treatment load. Ignoring this complexity can lead to waste of government funds and worsen the course of the epidemic. For example, if not supported with adequate treatment interventions, some patients dependent on POs may switch to heroin and thus increase the chance of overdose [10, 14–17]. To prevent this, strategies to reduce the supply of POs must be implemented in parallel with strategies to increase the availability of treatment for opioid use disorders. A comprehensive approach that considers the complexity and reciprocal nature of the opioid epidemic is critical to understand future treatment needs [18]. When policy makers have to respond quickly to a complex problem they often turn to simulation models. This has been done to impact policies on heroin treatment [19–21], response to PO overprescription [22, 23], tobacco control [24], and infectious and chronic diseases [25–28]. Without a forward-looking simulation tool that accounts for the complexity of the opioid epidemic, making reliable projections of future treatment gaps and estimating the economic impact of interventions is problematic. In this paper we present an initial model that focuses on the trajectories of people who are prescribed painkiller opioids. This category of patients is a source of intense discussion among prevention, treatment, and policy professionals mostly because the epidemic is relatively new and little has been done to formally model individual trajectories of such patients. Although a number of alternative theories exist [9], we base the model structure on recent literature with the main concept described by Cicero et al. [14].

2 The Model

The model was implemented in NetLogo [29] and we followed an Overview, Design concepts, and Detail (ODD) protocol [30], which is briefly summarized here.

2.1 Agents, Behavioral Rules, and Processes

Our model describes the dynamics of opioid use in a community populated by patients, physicians, drug dealers, and pharmacies and EDs that distribute opioids. The model includes only these agents to keep it simple and focused specifically on the pathways of patients in pain. Thus the working name "Pain Town". Those agents behave as follows:

Patient: If suffering from chronic pain, patients seek treatment from a physician. If prescribed opioids by a physician, a patient will go to a pharmacy to fill the prescription and take the opioids over the course of a month. Patients develop tolerance (i.e., larger dose needed to achieve the same effect) over time depending on the prescribed dose, and some patients will probabilistically become non-adherent, taking a higher dose than prescribed. If a non-adherent patient's desire for opioids exceeds his or her supply, the patient may instead choose to visit additional physicians, visit an ED, or buy from another patient or a dealer friend, and may switch from pills to heroin if heroin is available. The main progression of patients, based on the pathway described by Cicero et al. [14], is stochastic; its progression is visualized in Fig. 1.

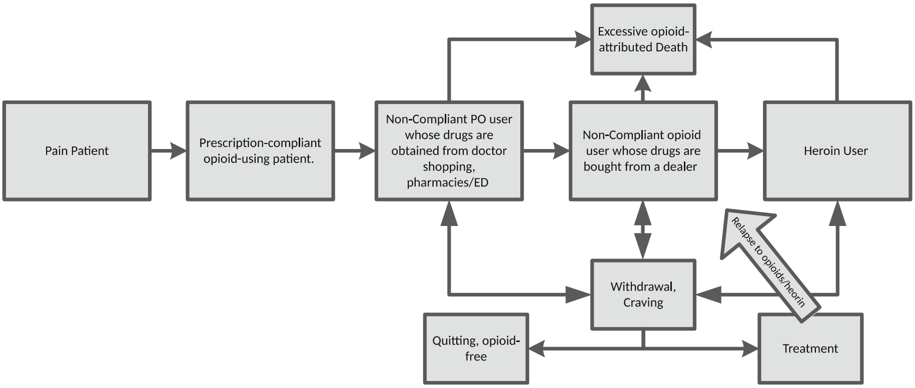

Excessive opioid-attributed Death
Pain Patient
Prescription-compliant opioid-using patient.
Non-Compliant PO user whose drugs are obtained from doctor shopping, pharmacies/ED
Non-Compliant opioid user whose drugs are bought from a dealer
Heroin User
Relapse to opioids/heorin
Withdrawal, Craving
Quitting, opioid-free
Treatment

Fig. 1. Patient pathway summary diagram. A patient can transition between the stages from pain prescription to the use of heroin, adapted from Cicero et al. [14]. At each point of time, an individual can stay in the current state or transit to the next state in the diagram. Most PO patients don't cross to the non-compliance region.

Physician: Can choose to prescribe opioids to patients. Before deciding whether to prescribe, a given physician has some likelihood of checking patients' purchase record in the prescription drug monitoring program (PDMP).

Emergency Department: Can choose to give patients opioids and decides on the dose. A given ED has some likelihood of checking a patient's purchase record in the PDMP (policy compliance). If an ED provides opioids, it adds that information to the PDMP.

Pharmacy: If a patient is prescribed opioids, he or she can get that prescription filled at a pharmacy. The pharmacy can check the patient's purchase record in the PDMP (policy compliance) to decide on dispensing a prescription. If the pharmacy fills the prescription, it adds that information to the PDMP. Pharmacies, based on global tamper-proof pill compliance, may provide tamper-resistant opioid pills instead of the standard variety.

Dealer: A dealer will give a patient the requested dose of opioids or heroin without a prescription, but the dealer has a fixed supply of both and resupplies every 30 days.

Agents are connected via link relations—each patient has links to one or more physicians, one or more friends who may be other patients or nonpatient dealers, and a pharmacy. Each patient and dealer has supplies of POs, tamper-resistant POs, and heroin, which they can give to their friends.

For presentation purposes in the model pain patients are arranged in a circle and a community of patients, physicians, dealers, pharmacies, and EDs is represented in Fig. 2.

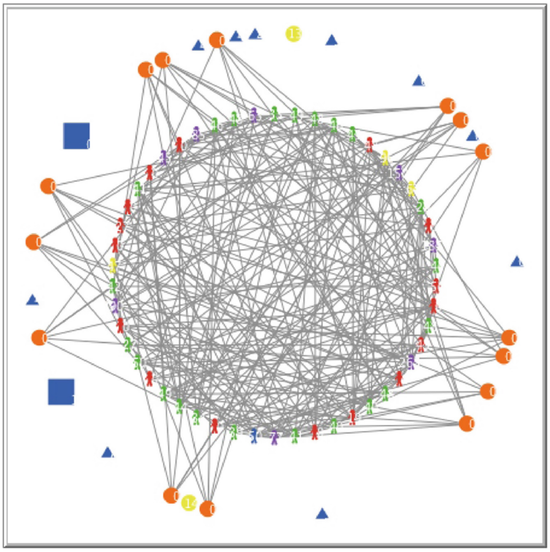


Fig. 2. A screenshot of the model illustrating patients socially connected in a circle with their current dose overlaid. Orange circles represent dealers, yellow circles represent EDs, blue triangles represent physicians, and squares are pharmacies. (Color figure online)

2.2 Time, Scales, and Events

We consider a time horizon of several years (5 years for these specific simulations), with a fixed time step equal to one day. Agents are updated in a random sequential manner (i.e., the next agent is updated based on the information of the previous agent's update, a common setting in NetLogo). For the exploratory simulations we considered a community of 10,000 patients, 70 nonpatient drug dealers, 30 physicians, 10 EDs, and 10 pharmacies. This is intended to be reflective of certain populations in a town community (i.e., this model does not include community members who are not in pain). All patients in the model have access to a physician and are able to switch physicians.

Once a month, patients visit a physician for a refill. Physicians prescribe 30-day supplies of opioids, and dealers can supply between 0 and 30 days, limited by the dealer's own supply, which is replenished every 30 days. Compliant patients take their prescribed dose each day, while noncompliant patients take opioids equal to their desire. In this simplified model, PDMP compliance is defined as checking the amount of opioids the patient has filled in the past month prior to prescribing more.

2.3 Adaptation

Noncompliant patients who are friends with a dealer probabilistically decide to buy opioids or heroin from that dealer instead of their pharmacy or patient friends when they do not have enough opioids to satiate their desires. Additionally, if a patient receives opioids from another patient who is buying them from a dealer, the second patient might introduce the first to the dealer, which allows dealers to increase their connectedness within networks of noncompliant patients. Such introductions are not common because many drug users see value in being middlemen between the dealer and other patients. However, the details of social networking among PO patients have not yet been clearly described.

Patients can physically adapt to the drugs they take. The actual rate of adaptation strongly varies between individuals. According to a 2013 Substance Abuse and Mental Health Services Administration report only 3.6% of nonmedical PO users progress to heroin. We used an approach previously used in Hoffer et al. [31, 32] to describe potential pathways to tolerance and dependence. Patients' satiation level is modeled as a function of the current dose, the length of use, and a random variable to represent between-person variation. After prolonged use the satiation level starts slowly increasing, building a gap between the current satiation level and current dose (tolerance building). At the same time each patient's desire increases to match the growing satiation level and to reach satiation the patient uses more drugs. This process forms a feedback loop: if a patient desires a higher dose than is prescribed the patient can either complain to a physician that the drug is not working as before and thus get an increase in the dose or start using more drugs and look for other opioid sources. The increase in the dose used in turn increases satiation level and desire. Some individuals can develop tolerance quite quickly, but for the vast majority of compliant patients on small doses tolerance builds up slowly.

If patients return to compliant use or abstinence, their satiation level will drop more quickly than their desire, which means they are at high risk of relapse for the first few weeks before their desire drops as well.

3 Model Parameters and Experimental Settings

Model parameters are informed by findings in the literature and expert input. Many values were converted from annual rates to daily rates assuming an equal distribution throughout the year. We considered four experimental interventions: physician PDMP compliance, pharmacy use of tamper-resistant pills, reduced initial doses of opioids, and increased naloxone availability.

The initial dose intervention reduced the average initial dose that patients were prescribed. Although our model does not claim to represent the intricacies of dosing opioids to treat different causes or amounts of pain, we imagine that real-world interventions like trainings to increase physician awareness, better-standardized treatment guidelines, or computer-assisted dosing would lead to reduced average doses in general.

A simplified version of PDMP is used for this model, and compliance is implemented as a two-part process. For purposes of this model, PDMP tracks only the total prescribed opioid dose that a patient receives in a 30-day period. When patients fill their prescriptions at the pharmacy or are dispensed opioids by an ED, their record is added to a PDMP database. When a patient sees a physician or requests opioids from a pharmacy, the physician and the pharmacist will probabilistically check that patient's PDMP record with probability equal to their PDMP compliance. If the patient has recently been prescribed opioids, the physician or pharmacist will turn them away.

Tamper-resistant pill compliance affects the kind of opioids dispensed by pharmacies. If a pharmacy is tamper-resistant pill compliant, it will only dispense tamper-resistant medication. Rather than represent the variety of techniques used to prevent pill tampering (which may involve bottles that dispense fixed numbers of pills, pills that cannot be crushed for injection use, or medications that contain naloxone so they are ineffective when injected), we model tamper-resistant opioids as medication that can only be taken in its prescribed dose. Despite pharmacy tamper-resistant medication, dealers still dispense opioids that can be tampered with, and noncompliant heroin-using patients can combine tamper-resistant drugs with heroin if they desire a greater dose [33–35].

Naloxone availability helps prevent deaths from opioid overdose. Specifically, naloxone is available with probability equal to the naloxone-availability parameter. If naloxone is available when a patient overdoses, if the patient would have died, he or she instead survives with probability equal to the naloxone-effectiveness parameter.

Medication-assisted treatment (MAT) includes three types of drugs: two opioid agonists—methadone, and buprenorphine—and an antagonist—naltrexone. Although our model contains a description of MAT, the treatment process is the subject of much more complex study and was not considered here.

Intervention parameters were by default set to "worst case" values (i.e., zero PDMP, tamper-resistant pill compliance, naloxone availability, and 50 mg initial dose) because that is the CDC threshold for a high opioid dose [36]. The experimental parameters are presented in Table 1.

Table 1. Experiment parameters with tested values.

Intervention	Value range experiment frequency
Initial dose	20–100 mg, tested every 5 mg
Average PDMP compliance	0–1, tested every 0.05
Tamper-resistant pill compliance	0–1, tested every 0.05
Naloxone availability	0–0.9, tested every 0.1

4 Results

Each parameter combination was measured using a mean of 100 iterations to reduce stochastic variation. The effectiveness of each intervention was measured using population counts of the following outcomes: number of heroin users, PO and heroin overdose

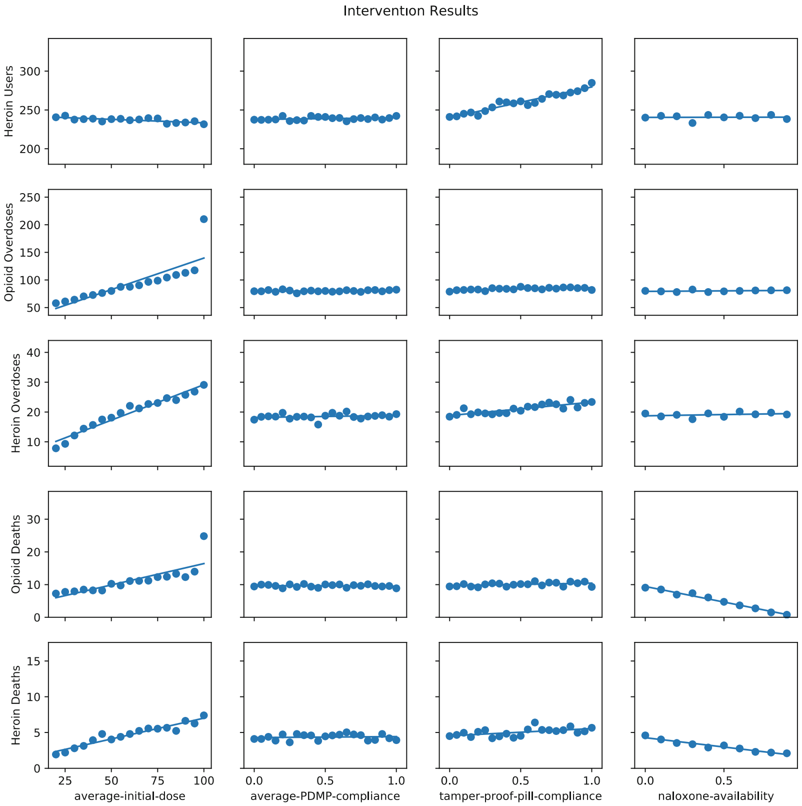


Fig. 3. Experimental results, by varied parameter and metric for simulations run for 5 years. Mean of 100 iterations.

rates, and PO and heroin death rates. We then regressed the five outcomes on the values of the parameters and calculated the strength of the relationships. Each of the parameters was varied with the rest kept at the default value. Thus, the results present only the marginal effects of the interventions. The combination of specific interventions can lead to potentially stronger results.

Simulation experiments find that decreasing average initial dose slightly increases heroin use but decreases rates of overdose and opioid death, tamper-resistant medication increases heroin use and both overdose rates, and naloxone availability decreases opioid and heroin death rates. The results are presented in Fig. 3.

In addition, we experimented with different values of PDMP compliance, which is reflective of expansion of PDMP programs, and the effect on number of opioid overdoses and number of heroin users, simulated for 20 years. The results are presented in Fig. 4. Notably, partial PDMP compliance led to slightly higher rates of heroin usage, but full PDMP compliance significantly reduced PO overdose rates. Each line represents one iteration.

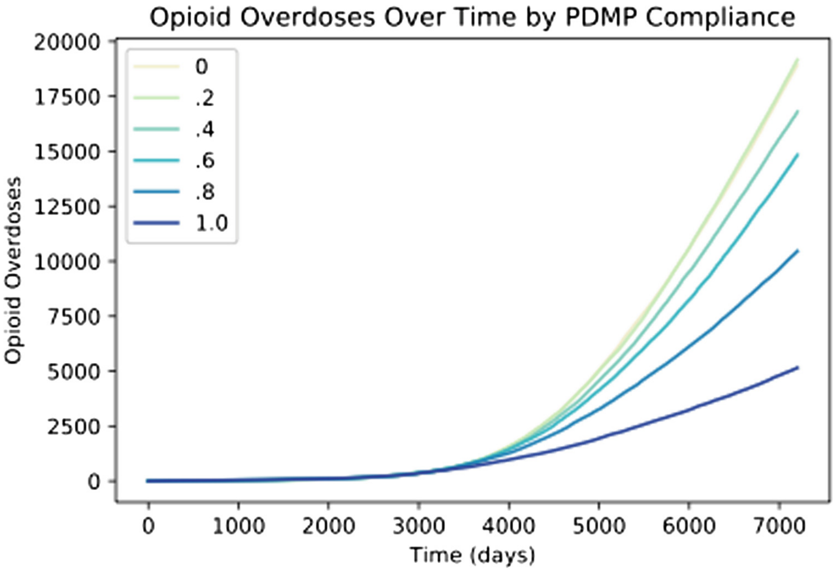


Fig. 4. Opioid overdose totals over time for different PDMP compliance ratios.

5 Discussion

We presented a simulation model and the results of simulated interventions aimed to fight the opioid epidemic. The effects of average initial dose on overdose and death rates are unsurprising. However, its effect on heroin use suggests that in the short term the initial dose does not have much effect, perhaps because patients either switch to PO from a dealer without yet switching to heroin or die from heroin overdose in a relatively short period of time thus depleting the heroin user pool. In this sense, it is critical to evaluate actual long-term trajectories rather than just examine “before and after” short-term snapshots of the population.

Notably, PDMP compliance does not strongly correlate with any of our intervention metrics during a short (5-year) time horizon. PDMP shows a notable effect at a much longer timescale (e.g., 10 years) but in the short term the positive effect of lower doses for newer patients is accompanied by an increase in switching to heroin among those who are already on high enough doses. Patients in our model initiate noncompliant behavior regardless of whether they can maintain it. PDMPs should reduce the uncontrolled flow of drugs through the patient network but dealer opioid supplies are sufficient to ensure that there are enough opioids in the network for all of the patients without direct connections to physicians. This additionally diminishes the effect of the PDMP among noncompliant individuals.

Tamper-resistant medications appear to have negligible impacts across the board. It is possible that patients who experience withdrawal exhibit the emergent behavior of tampering with their pills and take a smaller dose than prescribed to keep their satiation above zero. Because tamper-resistant medications cannot be used in smaller doses in our model, they increase relapse rates among noncompliant patients. This needs to be further examined in terms of the theory and supporting evidence.

Naloxone availability has the biggest impact on death rates of any intervention and has no clear negative consequences. However, naloxone’s only effect is to avert overdose death and has no impact on switching from POs to heroin in our model.

Our model highlighted the need to look closer into the process of building tolerance. Although discussions about patients’ tolerance can be found in the literature and the news, we found little hard data measuring tolerance development. The same goes for factors associated with transitions to heroin and other drug use. The overall pathway from prescribed opioids to overdose death has not been well established without large-scale longitudinal studies similar to the ones used to calibrate individual trajectories for chronic diseases such as cancer or diabetes.

5.1 Limitations and Future Work

Despite preliminary results, this model is still a work in progress. A number of components need further development and accurate calibration. Among such components are representation of pain, patient behaviors, physician behavior surrounding PDMPs, treatment options, availability and coverage, and finally non-overdose harm that results from drug use. For example, we do not capture the differences between chronic and acute pain, evolving pain over time, etc. The real-world downside to decreased doses of opioids is that they do not necessarily treat pain effectively, and our model does not capture that nuance. Patient characteristics and behaviors are greatly simplified in this model. We assume that all patients in pain will seek treatment and react positively to opioid treatment. Further investigation could address the behaviors of chronic versus acute care patients and effects of individual or community protective characteristics.

Physician behavior is another underdeveloped component of our model; our physicians will not encourage patients to get treatment for opioid dependence or learn to turn away nonmedical users. Additionally, we use a simplified model of PDMP. In reality, PDMP consists of multiple policy components that could be tested separately and together. We do not model any of the non-overdose–related dangers of injection drug use, including communicable diseases or infections resulting from unsafe injection or drug-related crime and do not have any interventions related to harm-reduction or treatment. Because these sorts of interventions are commonly discussed as strategies for preventing drug-related deaths, more investigation is warranted.

The other major limitation of our model is lack of data. We do not have clear information about how noncompliant opioid users behave and make several assumptions, in particular about tolerance and desire. An investigation into how this subset of the population uses drugs is necessary for any of our results to give conclusive findings.

6 Conclusions

We presented a model of an unexplored pathway from prescription-compliant opioid use to heroin abuse and explored the effects of several interventions on patient outcomes. Our simulation model produced strong effects of naloxone use, strong effects of low initial prescription dose, and marginal short-term effects of PDMP compliance but much stronger long-term (10 years) effects.

References

1. Rossen, L.M., Bastian, B., Warner, M., Khan, D., Chong, Y.: Drug poisoning mortality in the United States, 1999–2015 (2016). https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/
2. U.S. Department of Health and Human Services (HHS), Office of the Surgeon General: Facing Addiction in America: the Surgeon General's Report on Alcohol, Drugs, and Health. HHS, Washington, DC, November 2016
3. Levy, B., et al.: Trends in opioid analgesic-prescribing rates by specialty, U.S., 2007–2012. Am. J. Prev. Med. **49**(3), 409–413 (2015)
4. Volkow, N.D., et al.: Characteristics of opioid prescriptions in 2009. JAMA, J. Am. Med. Assoc. **305**(13), 1299–1301 (2011)
5. Crane, E.H.: The CBHSQ report: emergency department visits involving narcotic pain relievers. Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality, Rockville, MD (2013)
6. Substance Abuse and Mental Health Services Administration: Key substance use and mental health indicators in the United States: results from the 2016 National Survey on Drug Use and Health (HHS Publication No. SMA 17-5044, NSDUH Series H-52). Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD (2017). https://www.samhsa.gov/data/
7. Ruhm, C.J.: Corrected US opioid-involved drug poisoning deaths and mortality rates, 1999–2015. Addiction (2018)
8. United States Congress, Senate Judiciary: S.524 - Comprehensive addiction and recovery act of 2016. In: 114th Congress, Washington (2016)
9. Kolodny, A., et al.: The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. Ann. Rev. Public Health **36**, 559–574 (2015)

10. Dasgupta, N., Beletsky, L., Ciccarone, D.: Opioid crisis: no easy fix to its social and economic determinants. Am. J. Public Health **108**(2), 182–186 (2018)
11. Substance Abuse and Mental Health Services Administration, Center for Behavioral Health Statistics and Quality: The NSDUH Report: Substance Use and Mental Health Estimates from the 2013 National Survey on Drug Use and Health: Overview of Findings. Substance Abuse and Mental Health Services Administration, Rockville, MD, 4 September 2014
12. Jones, C.M., et al.: National and state treatment need and capacity for opioid agonist medication-assisted treatment. Am. J. Public Health **105**(8), e55–e63 (2015)
13. Maxwell, J.C.: The pain reliever and heroin epidemic in the united states: shifting winds in the perfect storm. J. Addict. Dis. **34**(2–3), 127–140 (2015)
14. Cicero, T.J., et al.: The changing face of heroin use in the United States: a retrospective analysis of the past 50 years. JAMA Psychiatry **71**(7), 821–826 (2014)
15. Mars, S.G., et al.: "Every 'never' I ever said came true": transitions from opioid pills to heroin injecting. Int. J. Drug Policy **25**(2), 257–266 (2014)
16. Inciardi, J.A., et al.: Prescription opioid abuse and diversion in an urban community: the results of an ultrarapid assessment. Pain Med. **10**(3), 537–548 (2009)
17. Inciardi, J.A., Martin, S.S., Butzin, C.A.: Five-year outcomes of therapeutic community treatment of drug-involved offenders after release from prison. NCCD News **50**(1), 88–107 (2004)
18. Dasgupta, N., et al.: Observed transition from opioid analgesic deaths toward heroin. Drug Alcohol Depend. **145**, 238–241 (2014)
19. Zarkin, G.A., et al.: Benefits and costs of substance abuse treatment programs for state prison inmates: results from a lifetime simulation model. Health Econ. **21**(6), 633–652 (2012)
20. Zarkin, G.A., et al.: Benefits and costs of methadone treatment: results from a lifetime simulation model. Health Econ. **14**(11), 1133–1150 (2005)
21. Ritter, A., Shukla, N., Shanahan, M., Van Hoang, P., Cao, V.L., Perez, P., Farrell, M.: Building a microsimulation model of heroin use careers in Australia (2016)
22. Wakeland, W., et al.: Modeling the impact of simulated educational interventions on the use and abuse of pharmaceutical opioids in the United States: a report on initial efforts. Health Educ. Behav. **40**(1 Suppl), 74s–86s (2013)
23. Wakeland, W., Nielsen, A., Geissert, P.: Dynamic model of nonmedical opioid use trajectories and potential policy interventions. Am. J. Drug Alcohol Abuse **41**(6), 508–518 (2015)
24. U.S. Department of Health and Human Services: The Health Consequences of Smoking: 50 Years of Progress. A Report of the Surgeon General. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA, January 2014. Printed with corrections
25. Subramanian, S., Bobashev, G., Morris, R.J.: Modeling the cost-effectiveness of colorectal cancer screening: policy guidance based on patient preferences and compliance. Cancer Epidemiol. Biomarkers Prev. **18**(7), 1971–1978 (2009)
26. Epstein, J.M., et al.: Controlling pandemic flu: the value of international air travel restrictions. PLoS One **2**(5), e401 (2007)
27. Subramanian, S., et al.: Personalized medicine for prevention: can risk stratified screening decrease colorectal cancer mortality at an acceptable cost? Cancer Causes Control **28**(4), 299–308 (2017)
28. Subramanian, S., Bobashev, G., Morris, R.J.: When budgets are tight, there are better options than colonoscopies for colorectal cancer screening. Health Aff. (Millwood) **29**(9), 1734–1740 (2010)

29. Wilensky, U.: NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL (2017). http://ccl.northwestern.edu/netlogo/
30. Railsback, S.F., Grimm, V.: Agent-Based and Individual-Based Modeling. Princeton University Press, Princeton (2012)
31. Hoffer, L.D., Bobashev, G., Morris, R.J.: Researching a local heroin market as a complex adaptive system. Am. J. Community Psychol. **44**(3–4), 273–286 (2009)
32. Hoffer, L., Bobashev, G.V., Morris, R.J.: Simulating patterns of heroin addiction within the social context of a local heroin market. In: Gutkin, B., Ahmed, S. (eds.) Computational Neuroscience of Drug Addiction. NEUROSCI, vol. 10, pp. 313–331. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-0751-5_11
33. Leece, P., Orkin, A.M., Kahan, M.: Tamper-resistant drugs cannot solve the opioid crisis. CMAJ Can. Med. Assoc. J. **187**(10), 717–718 (2015)
34. Cicero, T.J., Ellis, M.S.: Abuse-deterrent formulations and the prescription opioid abuse epidemic in the United States: lessons learned from OxyContin. JAMA Psychiatry **72**(5), 424–430 (2015)
35. Romach, M.K., Schoedel, K.A., Sellers, E.M.: Update on tamper-resistant drug formulations. Drug Alcohol Depend. **130**(1), 13–23 (2013)
36. Centers for Disease Control and Prevention: CDC guideline for prescribing opioids for chronic pain. Center for Disease Control and Prevention, Atlanta, GA (2017)

Assessing Target Audiences of Digital Public Health Campaigns: A Computational Approach

Robert F. Chew[1](✉), Annice Kim[1], Vivian Chen[2], Paul Ruddle[3], and Antonio Morgan-Lopez[1]

[1] RTI International, Research Triangle Park, Durham, NC, USA
rchew@rti.org
[2] Northwestern University, Evanston, IL, USA
[3] Imangi Studios, Raleigh, NC, USA

Abstract. As a larger proportion of society participates in social media, public health organizations are increasingly using digital campaigns to engage and educate their target audiences. Computational methods such as social network analysis and machine learning can provide social media campaigns with a rare opportunity to better understand their followers at scale. In this short paper, we demonstrate how such methods can help inform program evaluation through a case study of FDA's The Real Cost anti-smoking Twitter campaign (@knowtherealcost). By mining publicly available Twitter data, campaigns can identify and understand key communities to help maximize reach of campaign messages to their target audiences.

Keywords: Social network analysis · Machine learning · Social media
Public health · Tobacco control

1 Introduction

Under the 2009 Tobacco Control Act, the U.S. Food and Drug Administration was given the authority to regulate tobacco products and to educate consumers about the health consequences of tobacco use. The FDA Center for Tobacco Products launched its first public education campaign in February 2014 targeting U.S. youth ages 12–17. The objective of The Real Cost campaign is to educate the nearly 10 million at-risk teens who are either open to smoking or already experimenting with cigarettes about the harmful effects of tobacco use. The Real Cost campaign launched ads across multiple media platforms including TV, radio, print, web, and out-of-home (e.g. billboards). The Real Cost also used digital and social media to reach targeted youth, but unlike other media channels [1–5], very little research has examined the reach and effectiveness of the social media components of the campaign.

With the rise of social media, organizations, including the FDA Center for Tobacco Products are utilizing social media platforms like Twitter, Facebook, Instagram, and YouTube to disseminate health messages to target populations. Social media is increasingly important for reaching youth since nearly 92% of them are online daily [6]. A social media presence enables campaigns to extend and amplify the reach of their

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 286–291, 2018.
https://doi.org/10.1007/978-3-319-93372-6_32

campaign messages by building a network of campaign followers and engaging with audiences interactively over time. This strategy is not feasible on other media channels that simply deliver paid advertising. One of the key advantages of using social media is leveraging influencers within existing networks to extend the reach of campaign messages to their followers. Reach and engagement metrics such as number of followers, likes, and comments are available through social media native analytic tools (e.g., Twitter Analytics, Facebook Insights). While these tools provide detailed metrics about the followers' engagement with campaign messages, the data is aggregated for all followers, which may be less useful if campaign followers include social media users who are not part of the intended target audience for the campaign (e.g., other public health agencies). Additionally, native analytic tools do not provide insights into how the network of followers are interconnected, what type of communities make up the followers' network, and which influencers are most connected to the campaign's intended target audience (e.g. youth). Having this information would give campaigns a better assessment of which communities and key influencers they need to target as campaign ambassadors to maximize reach of campaign messages to their intended target audience. Addressing these research questions requires social network analysis, which researchers are increasingly using to examine the reach of health campaigns on social media [7]. Nonetheless, to date, few studies have applied social network analysis to identify the communities among specific campaign followers, with even fewer using machine learning to help understand if they are reaching their target audience at scale.

There are several ways that social network analysis could be helpful for understanding the reach of The Real Cost Twitter campaign (@knowtherealcost): (1) by graphing the network of followers, we can understand how the followers are interconnected and identify the core communities, and (2) once the core network communities are identified, we can examine the age distribution of the members in these networks to determine which communities have higher proportion of youth followers. Twitter only provides age and other demographic information about campaign followers at the aggregate level for all followers. By applying age prediction algorithms that examines the user handle's profile and tweeting behavior, we can predict whether twitter users are youth vs. young adults vs. adults [8].

2 Methods

2.1 Data Acquisition and Preparation

To create the campaign follower networks, we acquired data from the Twitter REST API in the first year of the campaign (August – September 2014). As of the summer of 2014, there were 24,060 accounts following @knowtherealcost. To collect network data, we made an initial request to collect followers of the @knowtherealcost campaign (24,060 users), followed by a request to collect the followers for each account following @knowtherealcost (~27 million users). However, since analyzing network data at this scale often adds noise without a compensating improvement in signal [9] and can bias metrics such as network density [10], we focused on the 1.5-degree ego network around @knowtherealcost – the campaign's direct followers and the connections between them.

Additionally, we collected up to the last 200 tweets for each account shortly after the initial request for follower account information to create features for the age prediction model.

In 2014, we found that nearly half of the followers of the campaign (14,560 followers) had no connections to anyone else in the network; since these nodes do not have edges connecting them to other nodes, they were omitted from the network analysis. While this allowed for a cleaner detection of distinct communities, network and centrality measures should be interpreted in the context of this modification, with the understanding that the modified network is less dense than the full 1.5-degree network. Additionally, because of privacy and data quality concerns respectively, private accounts and accounts without any tweets were also excused from analysis.

2.2 Graphing Network

To understand the relationships between followers of @knowtherealcost, we created a 1.5-degree ego network, including only follower accounts that are connected to at least one other follower account. Such a network extends a 1-degree network (an account and its followers) by also including direct connections between followers. We imported this data into the network analysis software Gephi and graphed the network using the Force-Atlas2 [11] layout algorithm, which can handle large networks well while maintaining useful network representations. Community detection was performed using the Louvain method [12] with the modularity metric in Gephi to produce well-defined communities. We then applied the age prediction algorithm to predict age groups within both the network and key communities.

2.3 Age Prediction

To understand the extent to which the campaign target audience of youth comprised the follower network, we applied an age prediction model to classify @knowtherealcost followers into three distinct age categories of youth (13–17 years of age), young adults (18 to 24 years of age), and adults (25 years or older) [8]. This model predicts a Twitter user's age by using a gradient boosted tree classifier on features derived from a user's metadata and their tweeting behavior (from their 200 most recent publicly available tweets). The model has previously shown to correctly predict the age of Twitter users with precision, recall, and F1-score statistics each of 74%, with superior accuracy measures for predicting youth and young adults [8]. This model was applied to all @knowtherealcost followers with public accounts and at least one tweet at the time of collection in 2014.

3 Results

Figures 1 illustrate the key networks that emerged when we graphed @knowtherealcost followers that had: (1) at least one other connection to another @knowtherealcost follower; (2) a public account; and (3) at least one public tweet. Other smaller networks emerged, but we present the top 4 major network communities. Figure 1 illustrates that in 2014 there were 4 key communities among the @knowtherealcost Twitter followers: ***music*** (2670 accounts), ***Tony Hawk/skating*** (1381 accounts), ***public health*** (1201 accounts) and ***ABC's The Fosters*** (702 accounts). The largest community focused around music and included key nodes such as @SleepSkee, @MoeRockOnline, and @REALMARQUETT. The music network was comprised of 39.7% predicted to be youth, 46.0% predicted to be young adults, and 14.4% predicted to be adults. The second largest network focused around skating and included key nodes such as @tonyhawk, @dustinlynch, and @impactwrestling. The skating network was comprised of 46.7% predicted to be youth, 37.2% predicted to be young adults, and 16.1% predicted to be adults. The third largest network focused around public health and included key nodes such as @Womenshealth, @HHSGov, and @PublicHealth. The public health network was comprised of 28.1% predicted to be youth, 25.9% predicted to be young adults and 46.0% predicted to be adults. The ABC's Foster's Show network, including key nodes such as @TheFostersABCF and @cierraramirez was comprised of 67.8% predicted to be youth, 28.9% predicted to be young adults, and 3.3% predicted to be adults. The remaining 800 account in the network belonged to a collection of miscellaneous communities that were smaller and more disperse in nature.

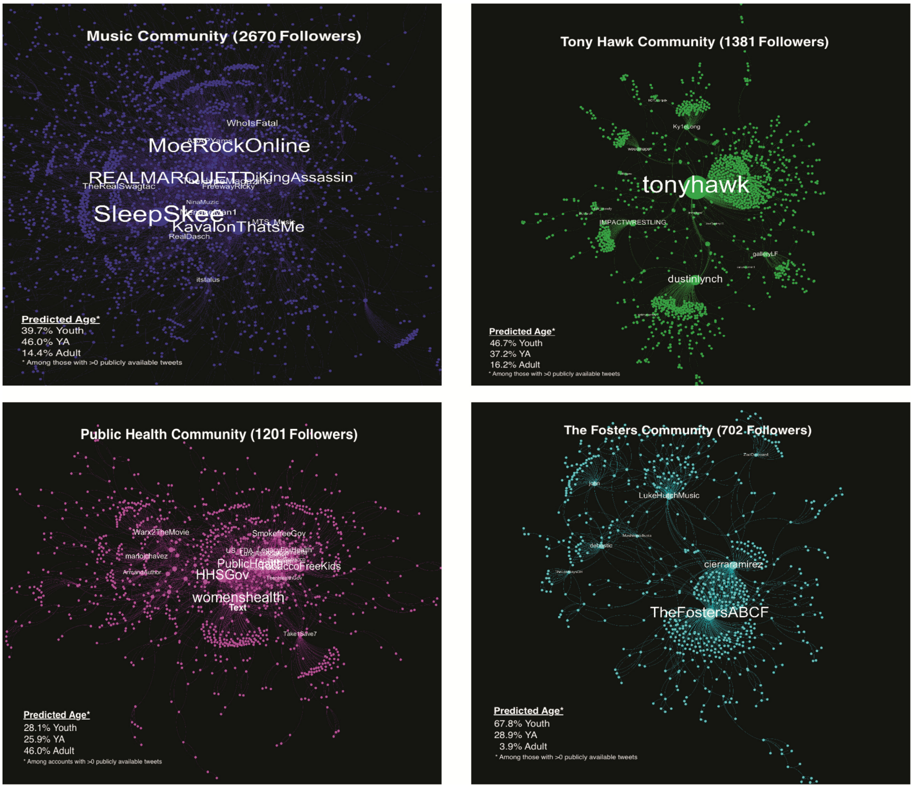


Fig. 1. Four key network communities among @knowtherealcost twitter followers in 2014.

4 Discussion

Using community detection algorithms, 4 key communities (celebrities, musicians, public health organizations, and the ABC Foster Show) were discovered in the 2014 @knowtherealcost Twitter follower network. Agreeing with intuition, the greatest proportion of youth were members of the ABC Foster's Show, followed by the skating and music communities. In contrast, few youth were part of the public health community. This is likely because the Foster's Show was an ABC family drama television series about a family of adopted teens that was popular among youth audiences [13].

The large follower network of celebrities and musicians may in part be explained by preferential attachment, a process in which new members give preference to following nodes that are already well-connected [14]. Youth may be more prone to this behavior, as research has shown that teens interact and engage around celebrities, pop culture, and entertaining trending topics on social media [15]. However, while our analyses showed that celebrities and musicians have larger reach to youth, this doesn't mean they are necessarily more influential than peers or less popular accounts to which teens may have stronger affinity. Studies are needed that examine how youth engage with smoking prevention campaigns delivered on social media and what impact this has on their behaviors.

There are some limitations of the study. First, the network analysis was conducted on accounts that were public, had connection with at least one other handle following @knowtherealcost, and had tweeted at least once during the period of data collection. Nearly half of the @knowtherealcost followers in 2014 did not fit these criteria and therefore, our results are not representative of the entire @knowtherealcost follower network. Second, since the Louvain method is an iterative process that depends on initial seed nodes, there can be variation in communities between different runs on the data. As such, community assignment should be viewed as a stable but not necessarily universal association between accounts. Third, there is a lag from when the initial call to the Twitter API is set-up to collect the network and tweet data, and when the data is finished collecting. As such, there can be drift in the follower network over the course of data collection period, and these results may not generalize to @knowtherealcost Twitter follower networks examined at other time points. Fourth, the age prediction algorithm has a 74% accuracy rate across age groups on a test set [8], so there may be some level of misclassification in the distribution of youth, young adults, and adults in these networks. Fifth, we only examined the predicted age of followers, and while this provides some indication of whether @knowtherealcost is reaching its target audience of teens, we were not able to examine other characteristics of FDA's campaign target, i.e. whether these were teens who had experimented with smoking or were susceptible to smoking. It is possible that teens who experiment with or are susceptible to smoking may be clustered in different network communities than the ones we identified in our analysis. Finally, in this study, we only examined the follower networks so these results may not generalize to teens who are not actively following @knowtherealcost handle but may nevertheless be exposed to these messages via other influencers.

While community detection and age prediction models provide some indication of whether target audiences are being reached, this view alone may not provide agencies with enough programmatic insight to inform interventions. Future work could identify influential members of each community through use of various centrality metrics, to better help agencies focus their resources effectively. Additionally, methods such as dynamic community detection could help campaigns understand how community membership and composition change over time, providing potential insights for program evaluation. Lastly, creating a network of tweeting behavior about the campaign could provide additional insights into important influencers and richer context into the follower engagement.

References

1. Duke, J.C., Alexander, T.N., Zhao, X., Delahanty, J.C., Allen, J.A., MacMonegle, A.J., Farrelly, M.C.: Youth's awareness of and reactions to the real cost national tobacco public education campaign. PLoS ONE **10**(12), e0144827 (2015)
2. Duke, J.C., Farrelly, M.C., Alexander, T.N., MacMonegle, A.J., Zhao, X., Allen, J.A., Delahanty, J.C., Rao, P., Nonnemaker, J.: Effect of a national tobacco public education campaign on youth's risk perceptions and beliefs about smoking. Am. J. Health Promot. 0890117117720745 (2017). http://journals.sagepub.com/doi/abs/10.1177/0890117117720745
3. Huang, L.L., Lazard, A.J., Pepper, J.K., Noar, S.M., Ranney, L.M., Goldstein, A.O.: Impact of the real cost campaign on adolescents' recall, attitudes, and risk perceptions about tobacco use: a national study. Int. J. Environ. Res. Public Health **14**(1), 42 (2017)
4. Farrelly, M.C.: Association between the real cost media campaign and smoking initiation among youths—United States, 2014–2016. MMWR Morb. Mortal. Wkly Rep. **66**, 47–50 (2017)
5. Zhao, X., et al.: Youth receptivity to FDA's the real cost tobacco prevention campaign: evidence from message pretesting. J. Health Commun. **21**(11), 1153–1160 (2016)
6. Lenhart, A., Duggan, M., Perrin, A., Stepler, R., Rainie, H., Parker, K.: Teens, social media & technology overview 2015, pp. 04–09. Pew Research Center [Internet & American Life Project] (2015)
7. Cobb, N.K., Graham, A.L., Byron, M.J., Niaura, R.S., Abrams, D.B., Participants, W.: Online social networks and smoking cessation: a scientific research agenda. J. Med. Internet Res. **13**(4), e119 (2011)
8. Morgan-Lopez, A.A., Kim, A.E., Chew, R.F., Ruddle, P.: Predicting age groups of Twitter users based on language and metadata features. PLoS ONE **12**(8), e0183537 (2017). https://doi.org/10.1371/journal.pone.0183537
9. Hai-Jew, S.: Querying Social Media with NodeXL. Scalar Publishing (2015). http://scalar.usc.edu/works/querying-social-media-with-nodexl/what-is-content-network-analysis
10. Cook J.: Social Networks, Lecture 5: Thinking About Ego Networks. http://www.umasocialmedia.com/socialnetworks/lecture-5-thinking-about-ego-networks/
11. Jacomy, M., Venturini, T., et al.: ForceAtlas2, a continuous graph layout algorithm for handy network visualization design for the Gephi software. PLoS ONE **9**(6), e98679 (2014). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098679
12. De Meo, P., Ferrara, E., Fiumara, G., Provetti, A.: Generalized Louvain method for community detection in large networks (2012). https://arxiv.org/pdf/1108.1502.pdf
13. The Fosters: Wikpedia. https://en.wikipedia.org/wiki/The_Fosters_(2013_TV_series). Accessed 19 Dec 2017
14. Newman, M.: Clustering and preferential attachment in growing networks. Phys. Rev. E **64**, 025102 (2001)
15. Boyd, D.: It's Complicated: The Social Lives of Networked Teens. Yale University Press, New Haven (2014)

Evaluating Semantic Similarity for Adverse Drug Event Narratives

Hameeduddin Irfan Khaja[1](✉), Marie Abate[2], Wanhong Zheng[3],
Ahmed Abbasi[4], and Donald Adjeroh[1]

[1] Department of Computer Science and Electrical Engineering,
West Virginia University, Morgantown, WV 26506, USA
hkirfan@mix.wvu.edu, don@csee.wvu.edu

[2] School of Pharmacy, West Virginia University, Morgantown, WV 26506,
USA

[3] School of Medicine, West Virginia University, Morgantown, WV 26506, USA

[4] McIntire School of Commerce, University of Virginia,
Charlottesville, VA 22904, USA

Abstract. We propose a method to evaluate adverse drug event (ADE) narratives using biomedical semantic similarity measures. Automated drug surveillance systems have used social media as a prime resource to detect ADEs. However, the problem of language usage over social media has been a challenge in evaluating the performance of such systems. We address this key issue by using semantic similarity measures and the biomedical vocabularies from the Unified Medical Language System. This is important in comparing results of social media driven approaches against standard reference documents from regulatory agencies.

Keywords: Semantic similarity · Adverse drug events · Social media

1 Introduction

High morbidity and mortality rates are associated with adverse drug events (ADEs), and hence, pharmacovigilance serves a critical task in post marketing surveillance. Recent advancements have shown a good potential for the detection of ADEs using social media much earlier than the traditional reporting systems [1–6]. Unfortunately, most of the work on detecting ADEs through social media have not emphasized the issue of language usage. The language used in expressing issues by healthcare consumers on social media forums and microblogging websites like Twitter is often very casual and informal [7]. On the other hand, warning labels and notifications from official regulatory agencies (such as the Food and Drug Administration (FDA) in the US) are formal documents and usually described in a language that is very carefully selected by biomedical experts. This raises a major concern as the words detected from social media channels by the surveillance systems do not exactly match with the contents of a typical FDA Black Box Warning (BBW) label or alert notification.

For many pairs of terms, there is a potential to miss the semantic similarity between social media extracted ADE terms and terms from FDA notification when two sets of

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 292–299, 2018.
https://doi.org/10.1007/978-3-319-93372-6_33

terms do not share exact text. More specifically the problem is as follows: given a formal FDA ADE narrative: $X = \{x_1, x_2, \ldots x_n\}$, and an informal ADE narrative from social media $Y = \{y_1, y_2, \ldots y_m\}$, determine the semantic similarity between X and Y. The three major issues related to semantic similarity in automated drug surveillance are: (1) How to measure semantic similarity between social media narratives and official formal documents, (2) How to use semantic similarity to evaluate the accuracy of detected ADEs, and (3) How to use semantic similarity to improve ADE signal detection. This work focuses on the first two problems. In general, X and Y could represent any two documents with words. Thus, semantic similarity can also have applications in other fields like medical appliances, ecommerce, etc.

Previously, Yang et al. [4] attempted to address the problem of health consumers' language over the Internet by generating ADE lexicons using Consumer Health Vocabulary (CHV) [7]. But, this did not address the issue comprehensively, as there are over 200 biomedical vocabularies in just UMLS (Unified Medical Language System), which also includes CHV [8]. Here, we use UMLS-Similarity program developed by McInnes et al. [9], for computing semantic similarity. It incorporates well-known semantic similarity and semantic relatedness measures. The prominent ones include path finding measures (such as Rada et al. [10], and Wu and Palmer [11]) as well as information content (IC) measures (such as Jiang and Conrath [12], and Sánchez et al. [13]). In prior work, Park et al. evaluated vocabularies from UMLS based on diabetes-related terms extracted from social media [14]. However, it confines itself to only one subset of the vast healthcare domain. We aimed at evaluating all the measures listed in UMLS-Similarity and vocabularies in UMLS to determine the best combination of measures and vocabulary in computing semantic similarity for ADE narratives.

2 Materials and Methods

Our methodology follows the procedure: (1) Identify the best vocabulary configurations (VCs) to use, (2) Determine the best combination of VCs and similarity measurement algorithms (SMAs) via joint optimization, and (3) Perform semantic similarity measurement using VC and SMA on given narratives.

2.1 Datasets

Problem Domain Terms. To evaluate VCs and SMAs, we used a list of terms grouped into anatomy and reaction categories. This dataset was earlier used by Adjeroh et al. [2] to study ADEs using social media data. The dataset has 105 anatomy terms and 202 reaction terms (called clusters). Each cluster was expanded with words having similar meanings, resulting in a new list with 178 anatomy terms, and 417 reaction terms.

Human Ratings. Language is a major concern in evaluating the signals generated from social media, hence, the testing on SMAs and VCs should be based on the ratings obtained from general healthcare consumers along with healthcare professionals. Thus, we used human ratings as the standard to compare the performance of each combination of SMA and VC. Initially, we had 178 anatomy terms and 417 reaction terms, and forming pairs with all these terms would lead to over 100,000 pairs and that would have been impossible for the respondents to rate the similarity. Thus, we randomly selected 30 anatomy terms forming a set of 435 [(30 * 29)/2] anatomy pairs and 40 reaction pairs forming a set of 780 [(40 * 39)/2] reaction pairs. Further, to rate these 1215 pairs we contacted 6 computer science graduate researchers having appreciable knowledge of biomedical vocabulary usage over social media. Finally, based on their ratings a template with a set of 100 pairs was designed comprising 50 anatomy pairs and 50 reaction pairs. This template had rating options 0, 0.25, 0.5, 0.75 and 1 indicating levels from non-similar to very similar. We obtained 130 user ratings across the United States. This consists of 54 individuals coming from 5 different universities with health sciences and engineering background, and 76 from Amazon Mechanical Turk users having at least US Bachelor's degree. Further, we selected 117 ratings by excluding the outliers that had a negative correlation with the mean. We also analyzed the inter-rater agreement in terms of average correlation between raters. We filtered the ratings to achieve the benchmark of 80% average correlation and this resulted in a total of 107 ratings.

FDA BBW. To evaluate our work, we used FDA black box warning (BBW) labels as gold standard references and extracted ADE terms from the labels from January 2008 to April 2015. (http://www.fda.gov/safety/medwatch/safetyinformation/). This included 107 BBWs, on 90 drugs over the seven-year period.

2.2 Selection of Vocabulary Configurations (VCs)

Since the biomedical terms are found in multiple vocabularies it becomes a challenging question to decide which vocabulary to be used. The harder part is to find how good a given vocabulary is, in terms of covering all terms in a given problem domain.

Initial Selection. UMLS has a huge collection of over 200 biomedical vocabularies which serves as a good resource for our work. However, we cannot use all the vocabularies in UMLS-Similarity due to performance and computational issues (see [15] for example). For our domain-specific social media extracted ADE terms, we followed the discussions in Park et al. [14], and selected vocabularies represented by source abbreviation (SAB): SNOMEDCT_US, CHV, MSH, LCH_NW, LNC, RXNORM, NCI_FDA, VANDF, and MTHSPL from UMLS [8]. The work in [14] was based on terms extracted from social media using queries for terms related to diabetes. For a more comprehensive treatment, we have considered some additional vocabularies where the content is closely related to ADE terms; namely, FMA, MDR, UWDA, WHO, NCI_NICHD, NCI_CTCAE, NDFRT_FDASPL, ICD10CM, MTHHH, and GS. Thus, given our specific problem domain of analyzing ADEs over social media channels, we had a total of 19 vocabularies to start our study.

Refining the VCs Selection. Our next task is to reduce the list to get the best possible VCs based on the concepts. We considered the following features:

1. Total CUI's: Total # of concept unique identifiers (CUIs) listed for the vocabulary;
2. Terms detected: number of problem domain terms detected in the vocabulary;
3. Concept coverage: number of concepts (CUI's) listed for problem domain terms;
4. Unique concepts: number of unique CUIs listed for each vocabulary; and
5. Clusters detected: number of clusters which had at least one term detected as CUI.

For our purpose, a good vocabulary is expected to have higher values for these features.

2.3 Similarity Measurement Algorithms (SMAs)

For automated evaluation of semantic similarity, vocabulary is just one piece of the puzzle. Another key piece is the specific SMA to be used to measure the similarity using the identified VC. Thus, having narrowed down the VCs we now turn to the problem of selecting the SMAs. Interestingly, the match performance can also be influenced by the vocabulary used. Thus, the final choice of vocabulary cannot be made in isolation but must consider the specific SMA being used. We used all SMAs in UMLS-Similarity except the *vector* measure which is meant to compute relatedness (see Table 1).

Joint Selection of VC and SMA. We computed similarity values for the problem domain terms using each combination of selected VCs and the SMAs. To select the best SMA and VC, we compared their results with those of human observers in two steps: (1) using Pearson correlation against the mean rating from human observers, and (2) using information retrieval measures, where we grouped the problem domain term pairs into 3 classes: **similar pairs**, **unknown pairs**, and **non-similar pairs**. Let $S(x, y)$ be the semantic similarity value between term pair (x, y), as returned by a given algorithm. We then used two thresholds τ_1 and τ_2 $(\tau_1 \geq \tau_2)$ to classify a word pair (v_1, v_2):

Table 1. Similarity measurement algorithms in UMLS-Similarity. *References for each can be found in [9].

#	UMLS-Similarity notation	Type	#	UMLS-Similarity notation	Type
1	***lch***	Path finding	9	***lin***	IC based
2	***wup***	Path finding	10	***jcn***	IC based
3	***zhong***	Path finding	11	***vector***	Context vector
4	***path***	Path finding	12	***pks***	Path finding
5	***upath***	Path finding	13	***faith***	IC based
6	***cdist***	Path finding	14	***cmatch***	Feature based
7	***nam***	Path finding	15	***batet***	Feature based
8	***res***	IC based	16	***sanchez***	IC based

$$Class(S(v_1, v_2)) = \begin{cases} similar, S(v_1, v_2) > \tau_1 \\ unknown, \tau_1 \geq S(v_1, v_2) \geq \tau_2 \\ non\text{-}similar, S(v_1, v_2) < \tau_2 \end{cases} \quad (1)$$

We used traditional information retrieval measures, namely, Precision (Pr), Recall (Rc), and F-measure (Fm) to evaluate the performance of combinations of VCs and SMAs across the three classes.

3 Experiments and Results

3.1 Filtering Vocabularies

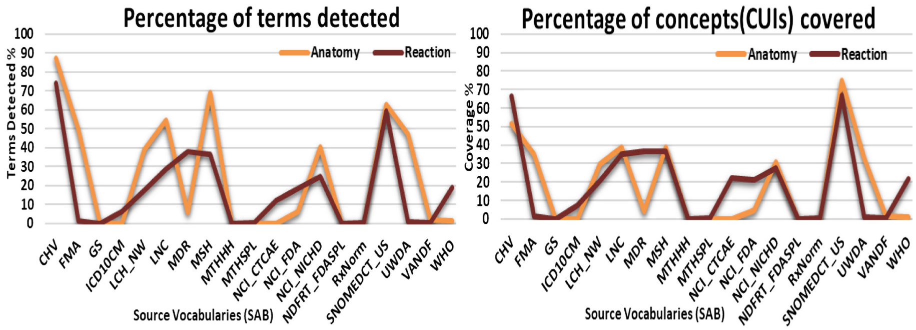
Using programs from the UMLS-Interface [9], we listed all the Concept Unique Identifiers (CUIs) for vocabularies configured with various relations defined in UMLS [8]. Interestingly, some vocabularies have concepts but are not connected by any relations. Additionally, we obtained the CUIs for all the problem domain terms to evaluate each vocabulary based on various features discussed in Sect. 2.2. Figure 1 shows some of the features used to describe the vocabularies. We observed that the top 5 vocabularies for anatomy category are SNOMEDCT, CHV, LNC, MSH, and FMA. The top 5 vocabularies for reaction category are SNOMEDCT, CHV, MDR, MSH, and LNC. However, we found that, CHV has no relations defined between CUIs which restricts its use independently. Thus, we used CHV in combination with other VCs as it has more coverage of terms, and has been shown to improve performance [4, 14].

3.2 Joint Selection of VC and SMA

Correlation Analysis. If the significance level is $\leq 5\%$ (i.e., p-value $\leq$ 0.05) and the corresponding correlation coefficient is positively high for any VC and SMA, then we say that the SMA or VC is favored. From Table 2 and Fig. 2, we can see that for anatomy category the SMAs which frequently appear to be good are *cmatch, jcn* and *sanchez* with VCs CHV-SNOMEDCT and CHV-LNC. For reaction category, we did not get significant p-value to favor any of the algorithms. However, it has been observed that *nam* has very high correlation coefficient with vocabularies CHV-MDR and CHV-MSH and undefined value for CHV-LNC. This behavior is because of the similarity values being −1.0 for most term pairs, resulting in less variability. Overall, the correlation analysis suggests that CHV-SNOMEDCT and CHV-MDR are the best VCs for working on reaction category terms (see Fig. 2(b)).

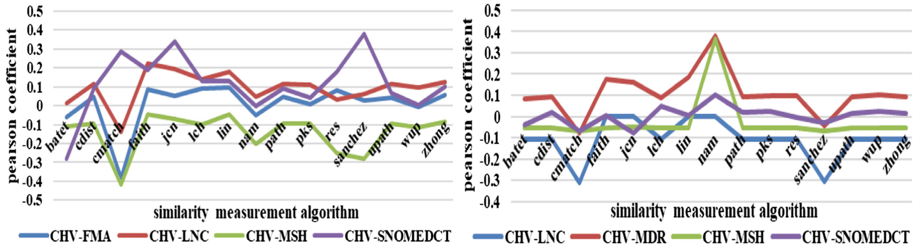
Table 2. Outcomes of Pearson correlation

SMA favored		VC favored	
Anatomy	Reactions	Anatomy	Reactions
cmatch, jcn, sanchez	*nam*	CHV-SNOMEDCT, CHV-LNC	CHV-SNOMEDCT, CHV-MDR



(a) Terms detected (b) Concepts Identified

Fig. 1. Features for filtering vocabularies based on problem domain terms



(a) Anatomy (b) Reaction

Fig. 2. Correlation of computed similarity values with human rating

Information Retrieval Factors. For the median of human ratings, we chose thresholds τ_1 as 0.75 and τ_2 as 0.3 to classify them into similar pairs, unknown pairs, and non-similar pairs. Similar to human ratings, for the SMA-VC obtained similarity values we chose τ_1 ranging from 0.5 to 0.95 and τ_2 ranging from 0.05 to 0.45 with a step size of 0.05. We selected the top 5 SMA-VCs based on F-measure against human rating statistic. For anatomy terms (Table 3), we found that the SMAs *jcn, faith, lin, cmatch* and *sanchez* with CHV-SNOMEDCT VC are having high F-measure values with respect to human ratings. For reaction category (Table 4), the SMAs *wup, lin, pks, cmatch* with CHV-SNOMEDCT, and *res* with CHV-MDR VC performed well. Interestingly, we observe that *sanchez* has good F-measure for both CHV-SNOMEDCT and CHV-MDR.

Table 3. Top 5 SMA/VC (Similar pairs–Anatomy)

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
jcn	0.8	0.5	0.3	CHV-SNOMEDCT	0.89	0.62	0.73
faith	0.7	0.5	0.2	CHV-SNOMEDCT	0.89	0.62	0.73
lin	0.8	0.45	0.35	CHV-SNOMEDCT	0.89	0.62	0.73
cmatch	0.5	0.45	0.05	CHV-SNOMEDCT	0.72	0.62	0.67
sanchez	0.8	0.5	0.3	CHV-SNOMEDCT	0.72	0.62	0.67

Table 4. Top 5 SMA/VC (Similar pairs–Reaction)

Measure	τ_1	τ_2	τ_{diff}	Configuration	Pr	Rc	Fm
pks	0.55	0.35	0.2	CHV-SNOMEDCT	1	0.3	0.46
res	0.8	0.3	0.5	CHV-MDR	1	0.3	0.46
sanchez	0.5	0.4	0.1	CHV-MDR	1	0.3	0.46
wup	0.75	0.3	0.45	CHV-SNOMEDCT	1	0.3	0.46
sanchez	0.85	0.3	0.55	CHV-SNOMEDCT	0.75	0.3	0.43

3.3 Evaluating Narratives in ADE Surveillance Systems

Considering both the information retrieval metrics and the correlation analysis, our results suggest the following: for anatomy term pairs, we should use *jcn, cmatch,* or *sanchez* SMA, with CHV-SNOMEDCT VC. For reaction term pairs, we should use *sanchez*, *wup*, or *res* SMA, with CHV-SNOMEDCT or CHV-MDR VC. A key observation is the need for a combination of vocabularies (typically, CHV with some others), rather than one single vocabulary as has been used in prior work, such as [4]. Prior work also did not consider the impact of the SMA on the results. We evaluated suggested ADE narratives from social media based on the method described in [16] using the BBW dataset (refer Sect. 2.1). We considered four cases: (1) exact match, i.e., not using semantic similarity; and the other 3 cases with SMA *sanchez* along with VC (2) CHV, (3) SNOMEDCT and (4) combination of CHV-SNOMEDCT (See Table 5). Clearly, our suggested approach using combination of CHV and SNOMEDCT performed better than others.

Table 5. Evaluating social media ADE narratives for BBW data

Approach	Anatomy			Reaction		
	Pr	Rc	Fm	Pr	Rc	Fm
exact match	0.048	0.176	0.076	0.022	0.140	0.038
CHV	0.048	0.176	0.076	0.024	0.141	0.041
SNOMEDCT	0.181	0.395	0.249	0.155	0.402	0.224
CHV-SNOMEDCT	**0.197**	**0.452**	**0.275**	**0.175**	**0.465**	**0.255**

4 Discussion and Conclusion

In this work, we chose UMLS-Similarity as it is built on UMLS which provides access to multiple vocabularies. The human ratings we used had a good representation of doctors, health professionals, health science students, engineering graduates and general graduate students. As the participants were familiar with social media as a significant source of healthcare information, we believe our dataset best fits the testing. We followed a-step-by-step approach evaluating all vocabularies and measures exhaustively, to get the best suitable VC-SMA combination for the ADE terms. Our results showed that, CHV-SNOMEDCT is the best VC for anatomy terms using the intrinsic IC-based measures *sanchez* or *jcn*. It is also observed that CHV-MDR and CHV-SNOMEDCT VCs work well for reaction category terms with *sanchez*. However, our results also indicate that using biomedical ontologies and the similarity measures is not sufficient for reaction category terms. The major reason is that reaction terms are more general and are not as specific when compared to anatomy category terms. Thus, we believe that using general English vocabularies such as WordNet [17] along with UMLS would improve the semantic similarity for reaction category terms. We plan to evaluate this in further studies. Our findings also show that the vocabulary MedDRA–Medical Dictionary for Regulatory Activities (abbreviated as MDR in UMLS) has a good representation of reaction category problem domain terms. This can be considered in the light of the fact that SIDER, a well-known dataset for representing side effects uses MedDRA to generate side effect names [18].

References

1. Sarker, A., Ginn, R., Nikfarjam, A., O'Connor, K., et al.: Utilizing social media data for pharmacovigilance: a review. J. Biomed. Inform. **54**, 202–212 (2015)
2. Adjeroh, D., Beal, R., Abbasi, A., Zheng, W., et al.: Signal fusion for social media analysis of adverse drug events. IEEE Intell. Syst. **29**(2), 74–80 (2014)
3. Abbasi, A., Adjeroh, D., Dredze, M., Paul, M.J., et al.: Social media analytics for smart health. IEEE Intell. Syst. **29**(2), 60–80 (2014)
4. Yang, C.C., Yang, H., Jiang, L.: Postmarketing drug safety surveillance using publicly available health-consumer-contributed content in social media. ACM Trans. Manag. Inf. Syst. **5**(1), 1–21 (2014)
5. Correia, R.B., Li, L., Rocha, L.M.: Monitoring potential drug interactions and reactions via network analysis of Instagram user timelines. In: Proceedings of the Pacific Symposium on Biocomputing 2016, vol. 21, pp. 492–503 (2016)
6. Abbasi, A., Li, J., Abbasi, S., Adjeroh, D., et al.: Don't mention it? Analyzing user-generated content signals for early adverse drug event warnings. In: Proceedings of the Workshop on Information Technologies and Systems (WITS), Dallas, TX, pp. 1–16 (2015)
7. Zeng, Q.T., Tse, T.: Exploring and developing consumer health vocabularies. J. Am. Med. Inform. Assoc. **13**(1), 24–29 (2006)
8. National Library of Medicine (US), UMLS® Reference Manual. National Library of Medicine (US) (2009)
9. McInnes, B.T., Pedersen, T., Pakhomov, S.V.S.: UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In: Proceedings of the AMIA Annual Symposium, pp. 431–5 (2009)
10. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Trans. Syst. Man Cybern. **19**(1), 17–30 (1989)
11. Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138 (1994)
12. Jiang, J.J., Conrath, D.W.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of International Conference Research on Computational Linguistics, Taiwan (1997)
13. Sánchez, D., Batet, M., Isern, D., Valls, A.: Ontology-based semantic similarity: a new feature-based approach. Expert Syst. Appl. **39**(9), 7718–7728 (2012)
14. Park, M.S., He, Z., Chen, Z., Oh, S., Bian, J.: Consumers' use of UMLS concepts on social media: diabetes-related textual data analysis in blog and social Q&A sites. JMIR Med. Inform. **4**(4), e41 (2016)
15. Re: [UMLS-Similarity] Practical large coverage configuration. https://www.mail-archive.com/umls-similarity@yahoogroups.com/msg00334.html. Accessed 15 Mar 2018
16. Khaja, H.I.: Signal fusion and semantic similarity evaluation for social media based adverse drug event detection. MS Thesis, Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA (2018)
17. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
18. Kuhn, M., Letunic, I., Jensen, L.J., Bork, P.: The SIDER database of drugs and side effects. Nucleic Acids Res. **44**(D1), D1075–D1079 (2016)

Military and Intelligence Applications

Framing Shifts of the Ukraine Conflict in pro-Russian News Media

Sultan Alzahrani[1(✉)], Nyunsu Kim[1], Mert Ozer[1], Scott W. Ruston[2],
Jason Schlachter[3], and Steve R. Corman[2]

[1] School of Computing, Informatics and Decision Systems Engineering,
Arizona State University, Tempe, AZ, USA
{ssalzahr,nkim30,mozer}@asu.edu

[2] Hugh Downs School of Human Communication, Arizona State University,
Tempe, AZ, USA
{scott.ruston,steve.corman}@asu.edu

[3] Informatics Laboratory, Lockheed Martin Advanced Technology Laboratories,
Kennesaw, GA, USA
jason.schlachter@lmco.com

Abstract. An important aspect of information operations (IO) are influence campaigns where a state actor or organizations under its control attempt to shift public opinion by framing information to support a narrative that facilitate their goals. If there is a playbook in operation, then in principle it should be possible to detect its signatures in mainstream media and to potentially provide early warning of malicious intent. This paper describes the results of a proof-of-concept effort where our goal was to detect framing shifts during the Ukraine conflict in pro-Russian news media surrounding the 2014 annexation of Crimea. Our results show significant framing shifts exceeding a smaller peak of 2010, in November 2013, and sharply spiking and trending again in Dec 2013, three-four months ahead of Crimea's annexation by the Russian Federation.

Keywords: Framing analysis · Time series data
Framing shifts detection

1 Introduction

Analysts recognize that the Russian government uses information operations (IO) as a tactic in its strategic efforts to reclaim territory in former Soviet states (it's so-called "near-abroad" [24]). For example, in 2008 Russia sent troops into South Ossetia, Georgia in response to an attack on the semi-autonomous region by Georgian forces. The speed and decisiveness of the Russian invasion and their subsequent extension of the invasion into Georgia proper caught Western leaders by surprise.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 303–314, 2018.
https://doi.org/10.1007/978-3-319-93372-6_34

Russia had promoted ethnic conflict in Georgia to maintain influence there,[1] and provided extensive support to South Ossetian and Abkhazian separatists [17]. Russia also exchanged old Soviet passports for new Russian ones in both South Ossetia and Abkhazia [3] so-called "passportization"- creating a pretext for intervention to protect "Russian citizens," and to take de facto control. Less than six years later, the West was again surprised when Russia used the same techniques to support annexation of Crimea in Ukraine. Joint Chiefs Chairman General Martin Dempsey said of Vladimir Putin, "he's got a playbook that has worked for him now two or three times" [18].[2] What is in this playbook?

> Case officers for the intelligence community operate without official cover, [and] recruit sources and assess the battlefield. Then, small units of special operations forces sneak in, sometimes blending in with the populace, ready to make trouble. Then, special forces units that specialize in "information operations" designed to induce anxiety and outrage among local populations follow a strategy that comes from the top of the government. The idea is to generate genuine indigenous protest movements. Using these protest movements as evidence of "human rights violations," Russia intervenes [16].

It is widely believed that Russia aims to repeat this performance in other ethnically Russian areas, especially the Gaugazia region of Moldova [20]. The Baltics are also a potential target. Three years ago, a Russian Foreign Ministry official echoed playbook tactics when he warned that ethnic discrimination there "may have far-reaching, unfortunate consequences" [21].

If there is a playbook in operation, then in principle it should be possible to detect its IO signature, stimulated by Russian propaganda and other 'gray zone' activities, in mainstream media, to potentially provide early warning of another invasion in other near-abroad states. This paper describes results of a proof-of-concept effort by the ASU's Center for Strategic Communication and Lockheed Martin Advanced Technology Laboratory. Our goal was to detect shifts in framing surrounding the 2014 annexation of Crimea using natural language processing of Russian propaganda articles and machine classifiers trained to recognize framing.

Our corpus comprised of over 100,000 news articles from 372 news sources dated between 2010 and 2017. Our methods and contributions can be summarized as follows:

[1] Archives of the CSCE, Georgia Files, Com. No. 408, Prague, Stockholm, 11 December 1992; Ibid, N.41, Prague, 2 February 1993; Bruce Clark, 'Russian Army blamed for Inflaming Georgian War,' The Times, 6 October 1992; Fiona Hill and Pamela Jewett, 'Back in the USSR: Russia's Intervention in the Internal Aairs of the Former Soviet Republics and the Implications for United States Policy toward Russia,' Cambridge, MA.: Harvard University JFK School of Government, Strengthening Democratic Institutions Project, January 1994.

[2] A playbook indicates a set of plans, approaches or strategies that aim to be equipped with a play ready catalog stating proposed actions and responses worked out ahead of time.

- We recruited a pair of area experts to classify top 200 news sources as either pro-Russian or other. We were able to train a classifier which achieved 90% F1-score to discriminate between propaganda vs. other articles.
- We worked with subject matter experts (SMEs) from ASU Center for Strategic Communication (CSC) to inductively develop a code book comprising five categories of Russian strategic frames used in Ukraine. Four student coders were trained to map sentences in randomly selected articles to one (or none) of these framing categories. After multiple rounds of training, coders achieved a inter-coder reliability (a.k.a Krippendorff ratio) of $\alpha = 0.83$ [19], which we judged as acceptable.
- We used coded sentences to train a text classifier which achieved 77% F1-score in labeling unseen sentences with the correct frame (or "no frame").
- The propaganda and framing classifiers were used on the news corpus to produce a daily time series of framing density vectors for articles classified as Russian propaganda. We computed Jensen-Shannon [4] divergence between framing density vectors of consecutive days. Results show significant framing shifts exceeding a smaller peak of 2010, in November 2013, and sharply spiking and trending again in Dec 2013, three-four months ahead of Crimea's annexation by the Russian Federation – which took place between 20 February 2014 and 19 March 2014. The war has been ongoing in the Donbass region of Ukraine since 6 April 2014 until the present day.

The rest of the paper is organized as follows. Section 2 presents a review of related works. Section 3 summarizes our data sources and approach. Sections 4 and 5 present the codebook of Russian strategic framing induced from propaganda articles and our sentence coding procedure. Sections 6 and 7 present text classifiers for frame detection, time series analysis of daily framing density vectors and significant framing shifts. Section 8 concludes the presentation with discussions and future work.

2 Related Work

Framing analysis has roots in mass media studies and several frameworks for assisting human identification and coding of frames were developed. Notable works include: Odijk et al. [14] where they developed a two-phase approach: (1) a systematic questionnaire for human coders to evaluate the nature (i.e. conflict, economic consequence, human interest, morality) and aspects of framing, (2) an ensemble of classifiers trained to detect frame presence in text using the coders questionnaire responses. Baumer et al. [9] compared performance effects of different types of features (i.e. lexical, grammatical and manual dictionary-based) for detecting frames in news. Their findings suggest that lexical n-gram features combined with grammatical part-of-speech (POS) tags result in significant improvements in frame detection. We also employed lexical frequent discriminative bi-grams alongside grammatical (subject, verb, object) based generalized triples [11] as features in our framework. Our experiments resulted in an accuracy of 41% average F1-score with bi-grams alone, and an average F1-score of

77% with combined features including bi-grams, generalized triples and other lexical features.

The temporal analyses of framing are also relevant since they can offer indications for detecting framing shifts. Several works were developed for spike detection in noisy time series data based on raw signal smoothing [15] and wavelet transforms [22] for different types of data (e.g. seismic analysis, disease epidemiology, and stock market prediction, etc.). Weng et al. [26] proposed an event detection framework in messages based on detecting correlated bursts of keywords that are expressed during events. To identify related keywords, they apply wavelet transformations on time series of keyword frequencies and measure cross-correlations between keywords and events. Next, they employ modularity-based graph clustering to detect keyword groups signaling events. In our paper, we utilized Jensen-Shannon divergence [4] to measure the daily variations of framing densities in pro-Russian international news. We checked the overlaps of their framing shifts and trends over time with significant phases of the Ukraine crisis to draw our conclusions.

3 Approach

Our analysis is based on detecting strategic framing [13,25] in news articles. Framing is accomplished when a choice of words, phrases, metaphors, images, and other rhetorical devices favor one interpretation of a set of facts, and discourage other interpretations. A special case is adversarial framing, which "is typically competitive, fought between parties or ideological factions, and [where issues] are debated and framed in opposing terms" [12]. A domestic example of adversarial framing is Republicans in the 1990s referring to the US estate tax as a "death tax"- connoting the long arm of the government taxing you even beyond the grave - while their political opponent Democrats referred to the same tax policy conventionally, as an "estate tax" - suggesting that only the super wealthy are subject to the tax.

Similar techniques are used by Russia with respect to the near abroad countries it threatens. One signature behavior is the framing of an ethnic issue as dealing with "human rights." In May 2014, the Russian Foreign Ministry released a white book detailing what it said were large-scale human rights violations in Ukraine [1], including discrimination against religious and ethnic minorities. In an earlier speech to the Russian Parliament, Vladimir Putin complained, "we hoped that Russian citizens and Russian speakers in Ukraine, especially its southeast and Crimea, would live in a friendly, democratic and civilized state that would protect their rights in line with the norms of international law. However, this is not how the situation developed" [2].

Framing is also undertaken by ethnic groups in the countries where Russian incursions are a threat. In 2012, a Latvian referendum rejected Russian as an official national language. Residents of Eastern regions where Russian is the primary language framed this act as a violation of rights. One such resident was quoted as saying: "[Latvian] society is divided into two classes - one half has full rights and the other half's rights are violated"[5].

Our approach, therefore, sought to identify and detect strategic framing before and after the 2014 invasion of Crimea. To do so we (i) collected mainstream media texts from Russian propaganda sources dealing with Ukrainian ethnic and political issues for the period between 2010–2017, (ii) inductively developed a set of framing categories, (iii) trained human coders to reliably identify sentences invoking these frames in sample texts, (iv) used these coded sentences to train machine classifiers to recognize all other framing instances in the corpus, (v) generated vectors representing the daily densities of these frames in news articles classified as propaganda, and (vi) conducted time-series analysis to identify shifts in framing densities and (vii) locate these shifts within significant phases of the Ukraine conflict.

3.1 News Corpus

This project was supported by Lockheed Martin Advanced Technology Laboratories and used news feeds extracted from Lockheed Martin's ICEWS system. ICEWS is a program of record in the U.S. Department of Defense used by component agencies to track conflict events. During its operation, ICEWS collects and archives English-language and translations of foreign language articles from mainstream media sources and websites worldwide. We queried the ICEWS database for articles between 2010 and 2017, which mentioned Ukraine, and further constrained this dataset to stories which contained keywords believed to be associated with Russian propaganda (i.e. anti-facist, discrimination, second-class citizens, etc.). This resulted in a news corpus containing 103,912 articles.

To focus our analysis on Russian propaganda sources, we recruited two area experts to classify the top 200 sources in our corpus (in terms of article frequency) as either pro-Russian or other. Next we extracted bigrams (i.e. pairs of two consecutive words after text preprocessing) and generalized concepts [11] from these sources and we trained a sparse logistic regression text classifier to discriminate between propaganda vs. other type of articles. A ten-fold cross-validation evaluation showed that the propaganda detection classifier has a an average F1-score of 90% and an F1-score of 86% for the smaller Russian 'propaganda' category. We ran this classifier on the news corpus, yielding 30,845 texts classified as Russian propaganda. These texts formed the basis of our coding and framing analysis.

4 Codebook

A codebook is survey research approach to provide a guide for framing categories and coding responses to the categories definitions. Using the notion of the playbook described in the introduction, we randomly selected articles from our Russian propaganda sources with high counts of discriminative propaganda-related keywords. Two subject matter experts, who are co-authors of this paper, from ASU's Center for Strategic Communication (CSC) read these texts and identified the following five framing categories inductively:

Fascist vs. anti-fascist struggle (denoted by: fascist). There are frequent accusations that leadership/society of a target country support "fascists" or "Nazis," and take actions to harass "anti-fascists" or hinder their efforts to protest and take other actions against the fascists. Essentially, the Nazis/fascists are the "bad guys" from the Russian point of view, and the anti-fascists are the "good guys." Almost any use of "Nazi," "fascist," or "anti-fascist" qualifies as framing, because it interprets the people involved and their actions as part of an ideological struggle between the two sides.

Discrimination against Russian minorities: (denoted by: discrim). This frame addresses discrimination against groups, usually ethnic groups; any such group having its rights trampled on, being marginalized or abused or similar affronts constitutes this frame. Russian information operations seek to convince members of the Russian speaking community in target countries that they are being victimized, discriminated against, and their rights are being violated. This might include references to general or human rights, or specific references to rights like voting, freedom of speech, and political participation. They also claim that there are efforts to stamp-out use of the Russian language, to suppress Russian culture, and to discriminate against Russian speakers in the job market and other domains. Lack of citizenship or denial of citizenship is a form of discrimination.

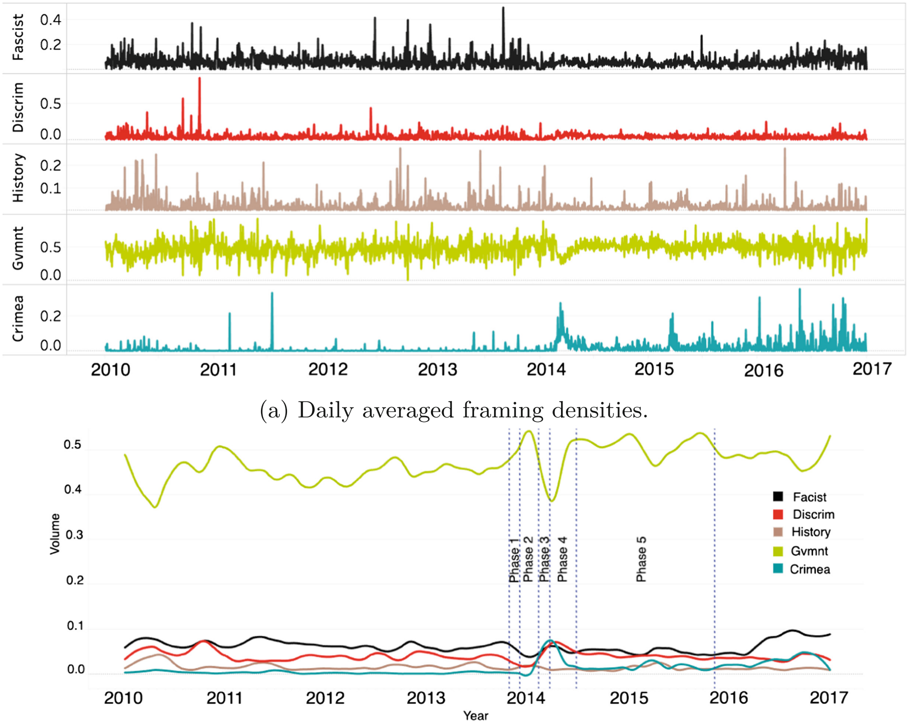
Assault on Soviet history (denoted by: history). Russian information operations seek to condemn the subversion or suppression of Soviet history. This can take the form of complaining about the removal of statues and memorials commemorating the Soviet role in World War II, changing names of Soviet-era streets and other geographical landmarks, or trying to change the historical narrative about the Union of Soviet Socialist Republics (USSR) and its role in former Soviet states.

Criticism of government (denoted by: gvmnt). Russian information operations seek to criticize the governments of target countries, in terms of functioning, procedures, and results (including economic results), as well as corruption among government officials. The frame implies that government is ineffective, not functioning properly, and acting in ways that are detrimental to good governance. The "government" includes legislative, executive and judicial branches at the national, provincial and municipal levels; it includes the police; it includes semi-synonymous terms like "the authorities". The frame applies when the national, provincial or municipal government of a target country is criticized (such as Ukraine, Latvia, Georgia, Lithuania, Estonia, Moldova, Poland, etc.)

Invasion of Crimea (denoted by: crimea). Russian information operations seek to justify and create support for their annexation of Crimea. This can involve discussions of sovereignty, discussion of the area's future, and statements supporting the annexation. The annexation is often framed as a moral imperative or a righteous act, and subsequent opposition by Ukraine, EU, and the international community are immoral, hypocritical, etc. Select this frame when the annexation of Crimea is clearly the context of some sort of justification, not when it could be the subject of the justification.

5 Frame Coding

Computer-aided techniques of frame coding essentially use two approaches: (I) dictionary/keyword lists based (e.g. [10]) or supervised learning approaches (e.g. [23]) trained with human coded sentences. In this project four student coders were trained to assign sentences in randomly selected propaganda texts to one (or none) of the five framing categories described above. Coders would first work independently, assigning each sentence to one (or none) of the coding categories. We would then calculate reliability, and identify disagreements between coders. Coders would then discuss these disagreements as a group, and we would refine category definitions in the codebook as necessary. After seven rounds of training, coders achieved a inter-coder reliability (a.k.a Krippendorff ratio) of $\alpha = 0.83$ [19], which we judged acceptable. Subsequent coding was performed by two randomly assigned coders per text, who discussed and resolved disagreements to arrive at a final set of codes. They coded texts until we had a large enough set of coded sentences, where adding more coded sentences no longer significantly boosted the overall accuracy of the best text classifier model. The final number of coded sentences in each category was: *crimea*, 162; *discrim*, 196; *fascist*, 307; *gvmnt*, 334; *history*, 187, and those sentences were used as the labeled training dataset.



(a) Daily averaged framing densities.

(b) Smoothed daily averaged framing densities.

Fig. 1. Daily averaged framing and Smoothed densities.

6 Frame Detection Model

We used coded sentences described above alongside a random collection of sentences that were not mapped to any framing category from coded articles to train five classifiers - one classifier for each frame category. We used one-vs.-all (OvA) strategy which involves training a single classifier per frame, with the samples of that frame as positive samples and all other samples as negatives. We extracted four sets of features from each sentence: keywords, frequent bigrams, whether the sentence contained a quote, and its matching generalized semantic triplets. Generalized semantic triplets (GST) are merged collections of subjects, verbs, and objects that co-occur together in similar contexts. The details of the GST features can be found in an earlier paper [7,8]. We evaluated several text classifiers using ten-fold cross-validation. The best overall performance was obtained with a linear SVC (L1) classifier yielding the following F1-scores: history, 74%; crimea, 87%; discrim, 76%; fascist, 75%; gvmnt, 73%; average, 77%. The rest of the results are shown in Table 1.

Table 1. Frame detection accuracies

Classifier	Frame														
	fascist			discrim			history			gvmnt			crimea		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
Ridge classifier	.82	.68	.74	.78	.67	.72	.83	.62	.71	.73	.63	.68	.87	.8	.83
Perceptron	.78	.65	.71	.71	.77	.74	.77	.73	**.75**	.76	.62	.68	.84	.86	.85
Passive-aggressive	.8	.65	.72	.79	.69	.74	.81	.67	.73	.75	.71	**.73**	.89	.8	.84
LinearSVC (L2)	.79	.68	.73	.76	.68	.72	.83	.67	.74	.74	.69	.72	.89	.78	.83
SGDClassifier (L2)	.8	.69	.74	.71	.69	.7	.79	.67	.73	.72	.64	.68	.88	.86	**.87**
LinearSVC (L1)	.79	.71	**.75**	.81	.71	**.76**	.8	.7	**.75**	.72	.74	**.73**	.85	.79	.82
SGDClassifier (L1)	.75	.65	.7	.73	.67	.7	.78	.72	**.75**	.7	.65	.68	.85	.82	.84
SGDClassifier (Elastic-Net)	.73	.65	.69	.75	.58	.66	.79	.71	.74	.78	.63	.7	.84	.83	.84

7 Time Series Analysis of Daily Framing Densities

The set of frame classifiers were applied to each sentence to produce real-valued confidence scores. The classifier which reported the highest confidence score was considered to be the dominant frame category for each sentence. We applied this technique to all sentences in each article one-by-one in order to produce a vector of framing density values for each article. These vectors were averaged daily to yield a vector of daily averaged frame densities shown in Fig. 1. Since the time series were noisy, first we performed Gaussian smoothing, shown in Eqs. 1 and 2 (where σ, w are $2, 10$ respectively, acting as low-pass filter) to remove high frequency noise. The smoothed time series are shown in Fig. 1. Next, in order to reveal framing shifts, we computed Jensen-Shannon [4] divergence, a statistical distance measure, between the daily framing density vectors of consecutive days. The resulting divergence plot is shown in Fig. 2.

$$N(x; \mu = 0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{-2\sigma^2}} \tag{1}$$

$$S(t) = \sum_{i=t-w/2}^{t+w/2} O(i) N(t-i) \tag{2}$$

Knowing that KL is the Kullback-Leibler divergence $KL(p;q) = p_i \ln \frac{p_i}{q_i}$, Jensen-Shannon divergence can be expressed in term of KL as follows

$$JS(v_1, v_2) = KL(v_1, \frac{v_1 + v_2}{2}) + KL(v_2, \frac{v_1 + v_2}{2}) \tag{3}$$

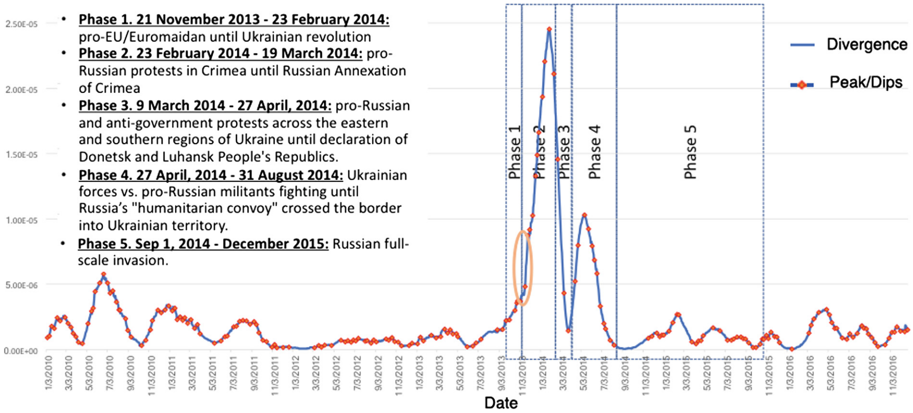


Fig. 2. Daily Jensen-Shannon divergence-vertical lines demarcating the significant phases of the Ukraine conflict timeline determined by the CSIS (CCIS: Center For http://ukraine.csis.org/)

Prior to Phase 1, corresponding to the period between pro-EU Euromaidan protests until the Ukrainian revolution, divergence remains at relatively low levels, except for some small peaks during 2010–2011. As the pro-EU/Euromaidan protests begun in November 2013, the divergence signal begins to rise, exceeding all previous highs in November 2013, followed by a sharp rise in Dec 2013. Divergence increases sharply during the pro-Russian protests well into the midst of Phase 2 which terminates with the annexation of Crimea by the Russian Federation on March 19, 2014. Following that, divergence sharply falls to its baseline levels. During Phase 3, the signal spikes once again as pro-Russian and anti-government protests took place across the eastern and southern regions of Ukraine until the declaration of Donetsk and Luhansk People's Republics. The signal declines again in Phase 4 which marks the Ukrainian forces vs. pro-Russian militants fighting a war. The signal meets zero-line during the initial days of Phase 5 marking the Russian full-scale invasion which was framed as an "humanitarian convoy" crossing into the Ukrainian territory. Following that, the signal remains at its baseline levels with no more major breakouts.

8 Discussions and Future Work

A question arises: could Russian propaganda framing shifts forecast the onset of hostilities leading to an invasion? In the Ukraine case, the divergence signal's early rise, exceeding all previous highs in Nov. 2013 followed by the sharp rise in Dec 2013 provides a signal of interest three-four months ahead of Crimea's annexation. If the premise is accepted that information operations are intended to "soften-up" the target area and provide a pretext for active conflict, then shifts in strategic framing might provide an early warning before the onset of pro-Russian protests, militant action and invasion under the guise of an "humanitarian convoy".

Our future work involves various tasks. Since our classifiers achieved an average 77% F1-score only, we plan to experiment with additional syntactic and semantic (framenet, wordnet, verbnet, LIWC18)[3] features, and other features such as named entity types to improve performance.

Next, we believe it might be possible to automatically surface framing categories to help spot newly emerging framing categories. We aim to synthesize narrative graphs incorporating co-occurrence patterns [11] of discriminant bi-grams, their adverbs, adjectives, named entities (i.e. people, places, organizations and locations) and apply dynamic graph clustering algorithms [6] to detect newly emerging clusters for SME's attention. Our initial experiments indicate that we can surface expert induced framing categories developed in the Ukraine codebook with a Normalized Mutual Information (NMI) score of 56% and purity of 68%.

Finally, we plan to evaluate this framework in other historical contexts; such as the Transnistria War in November 1990 between Moldovan troops and pro-Transnistria forces supported by elements of the Russian Army and the Russo-Georgian War between Georgia, Russia and the Russian-backed self-proclaimed republics of South Ossetia and Abkhazia in August 2008.

Acknowledgements. The authors would like to thank Kristin Fleischer, Madison Roselle, Sean West, and Zebulon Stampfler who participated in sentence coding support.

[3] https://framenet.icsi.berkeley.edu/, https://wordnet.princeton.edu/, https://verbs.colorado.edu/verbnet/, https://liwc.wpengine.com/.

References

1. TASS: Russia - Russian Foreign Ministry presents White Book on human rights abuses in Ukraine. http://tass.com/russia/730463
2. Transcript: Putin says Russia will protect the rights of Russians abroad - The Washington Post. https://goo.gl/bacjU5
3. TSG IntelBrief: Russia's Passport Imperialism—The Soufan Group. http://www.soufangroup.com/tsg-intelbrief-russias-passport-imperialism/
4. Jensen-Shannon divergence and Hilbert space embedding. In: 2004 Proceedings of the International Symposium on Information Theory, ISIT 2004 (2004)
5. Latvians reject Russian as official language—World news—The Guardian (2012). https://goo.gl/wXR8K4
6. Aktunc, R., Toroslu, I.H., Ozer, M., Davulcu, H.: A dynamic modularity based community detection algorithm for large-scale networks: DSLM. In: ASONAM, IEEE/ACM (2015)
7. Alashri, S., Alzahrani, S., Tsai, J.-Y., Corman, S.R., Davulcu, H.: "Climate change" frames detection and categorization based on generalized concepts. Int. J. Semant. Comput. **10**(02), 147–166 (2016)
8. Alzahrani, S., Ceran, B., Alashri, S., Ruston, S.W., Corman, S.R., Davulcu, H.: Story forms detection in text through concept-based co-clustering. In: 2016 IEEE SocialCom (2016)
9. Baumer, E.P.S., Elovic, E., Qin, Y.C., Polletta, F., Gay, G.K.: Testing and comparing computational approaches for identifying the language of framing in political news. In: ACL, pp. 1472–1482 (2015)
10. Benoit, K., Laver, M.: Estimating Irish party policy positions using computer wordscoring: the 2002 election. Ir. Polit. Stud. **18**, 97–107 (2003)
11. Ceran, B., Kedia, N., Corman, S.R., Davulcu, H.: Story detection using generalized concepts and relations. In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (2015)
12. Chong, D., Druckman, J.N.: A theory of framing and opinion formation in competitive elite environments. J. Commun. **57**(1), 99–118 (2007)
13. Entman, R.M.: Framing: toward clarification of a fractured paradigm. J. Commun. **43**(4), 51–58 (1993)
14. Odijk, D., Burscher, B., Vliegenthart, R., de Rijke, M.: Automatic thematic content analysis: finding frames in news. In: Jatowt, A., et al. (eds.) SocInfo 2013. LNCS, vol. 8238, pp. 333–345. Springer, Cham (2013). https://doi.org/10.1007/978-3-319-03260-3_29
15. Tapley, B.D., et al.: GRACE measurements of mass variability in the Earth system. Science **305**(5683), 503–505 (2004)
16. Resneck, J.: In tiny Moldova, Russia is repeating its Ukraine playbook—Public Radio International. https://goo.gl/PQR5XA
17. Graham, J.: Russia and Ethnic Conflict in Georgia - On This Day. https://www.onthisday.com/russia/georgia.php
18. Kifield, J.: How to Prevent War with Russia - POLITICO Magazine. https://goo.gl/ocLo55
19. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. Sage, Beverly Hills (2004)
20. Ambinder, M.: Russia masters the art of clandestine warfare against Ukraine. https://goo.gl/dG3EzD

21. Seddon, M.: Russia Warns of "Unfortunate Consequences" Over Ethnic Tension in Baltic States. https://goo.gl/rfA39V
22. Zoran, N., Burdick, J.W.: Spike detection using the continuous wavelet transform. IEEE Trans. Biomed. Eng. **52**(1), 74–87 (2005)
23. Quinn, K.M., Monroe, B.L., Colaresi, M., Crespin, M.H., Radev, D.R.: How to analyze political attention with minimal assumptions and costs. Am. J. Polit. Sci. **54**(1), 209–228 (2010)
24. Safire, W.: On Language: The Near Abroad. The New York Times, May 1994
25. Scheufele, D.A., Tewksbury, D.: Framing, agenda setting, and priming: the evolution of three media effects models. J. Commun. **57**, 9–20 (2007)
26. Weng, J., Yao, Y., Leonardi, E., Lee, F.: Event detection in Twitter. Development, pp. 401–408 (2011)

Turning Narrative Descriptions of Individual Behavior into Network Visualization and Analysis: Example of Terrorist Group Dynamics

Georgiy Bobashev[1], Marc Sageman[2], Amanda Lewis Evans[1], John Wittenborn[3], and Robert F. Chew[1](✉)

[1] RTI International, Research Triangle Park, NC, USA
{bobashev,rchew}@rti.org
[2] Sageman Consulting, Rockville, MD, USA
[3] National Opinion Research Center (NORC), Chicago, IL, USA

Abstract. Social networks play a critical role in the formation of criminal and radical groups. However, understanding of these formations relies on difficult to collect data. We present an approach where narrative data from the trial of the 1995 Paris Metro and RER bombings was used to extract actors, places, groups and actions that led to the formation of the radical group. This data was dynamically visualized and allowed one to follow the process of terrorist group formation. An important part of the approach is the inclusion of the individuals who were parts of the social network of the radicalized individuals but who did not get radicalized (e.g. members of a soccer team). We emphasize the importance of the Natural Language Processing (NLP) in timely information extraction followed by dynamic visualization.

Keywords: Terrorist networks · Social network analysis
Network visualization

1 Introduction

Over the past two decades, social and statistical sciences have made substantial progress in the development of analysis methodology for social network data [1–13]. However, outside of a few research teams [3, 14–16], there is limited published network modeling work that capture the dynamic characteristics of terrorist networks. A major barrier for conducting such analyses is the lack of quality data and measures that could be collected, abstracted, analyzed, and interpreted in a systematic manner.

In an influential article discussing the role of social network analysis for studying terrorism, Ressler [17] argues that the chief contributing factor preventing such work is the disconnect between the research communities primarily focusing on methodical data collection on terrorist organizations ("Data Collectors") and those developing complex network models of terrorist networks ("Modelers"). In particular, he argues that though Data Collectors [18–20] have subject matter expertise and often richer data (e.g., individual biographies of terrorist operatives) their work is more descriptive or conceptual

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 315–328, 2018.
https://doi.org/10.1007/978-3-319-93372-6_35

in nature, lacking a statistical analytical framework that could help account for certain data limitations and dependencies. Alternatively, he argues that while Modelers [21–23] are doing impressive work at the intersection of social network analysis, agent based modeling, and text analytics, they often lack a foundation in terrorist studies and do not always have access to the best data, resulting in models that can be potentially misleading or that fail to incorporate important behavioral and contextual issues.

This paper draws inspiration from both communities by providing a case study of dynamic network visualization created from rich descriptive data of the 1995 Paris Metro and RER bombings trial. By converting these narrative texts into measurable entities, we can visualize and analyze temporal development of individual and group paths to radicalism and provide an illustration of how curated textual data can be incorporated into a computational social network framework to help understand the process of radicalization.

1.1 Social Network Theories of Radicalization

Understanding the process of how people radicalize and adopt violent extremist ideologies is an essential component of many counter-terrorism initiatives, as it suggests a means of stemming the recruitment of new extremists and de-radicalizing individuals who are at high risk of violent extremism or performing terrorist acts. Though many theories of radicalization have been proposed, prior research suggests that no single theory or explanatory pathway can apply universally well to the diversity of observed individuals and group behaviors [24]. As a result, many complementary (and sometimes competing) theories of radicalization exist, though most are conceptual rather than empirical in nature [25]. In their review of the literature, Crossett and Spitaletta [26] identified sixteen overarching theories of radicalization, drawing from sociological, psychological, psychoanalytic, and cognitive perspectives. Our work focuses on assisting one of these family of theories, the social network theories of radicalization, through use of careful data capture and dynamic network visualization.

The social network theories of radicalization are characterized by the belief that the structure of an individual's social network strongly effects their opportunities, choices, and influences in life. Wheeler [27] promotes this idea by stating that radical groups benefit from smaller, denser networks without many loose ties outside the main network, as more open network structures leave room for exposure to new ideas and influences outside of the core membership. From an individual perspective, Sageman [19] and Post [28] have shown the impact of isolation and lack of social connections on the adoption of radical views and successful recruitment into terrorist networks. Also related to recruitment, Perliger and Milton [29] report that approximately three-fourths of Islamic State (IS) or al-Qaeda members join in groups instead of individually, with many group members consisting of previously-formed social networks. Jasko et al. [30] found evidence that having radicalized friends (but not family members) in an individual's social network increased their likelihood of performing acts of violent extremism. Perliger and Pedahzur [31] theorize that radicalization forms when communities holding totalistic ideologies are confronted by powerful outsiders, and furthermore, that the shift to violence is bolstered by the framework of isolated, close-knit social networks within the broader radicalized community. Social network theories even extend to the process of deradicalization, as Stern [32] provides evidence of Saudi Arabia using family connections and arranged marriages to position former radical extremists back into conventional society.

While social network theories of radicalization have aided in the understanding of the radicalization process, few researchers have had the data or expertise necessary to visualize how terrorist networks evolve in order to refine their theory, let alone rigorously empirically test their assumptions through network modeling. We present a process of building dynamic social network visualizations from narrative data as a first step in informing and assessing social network radicalization theories. A better understanding of the process that leads groups to radicalize and engage in political violence is essential for designing effective evidence-based policy to deter terrorist events and diminish the effect of hostile actions. Such understanding can help define the pathways to violence and the group characteristics and environmental context that facilitate it. Furthermore, such an understanding can allow us to develop and evaluate interventions that can be used to interrupt the pathways to violence before a group engages in such behavior [33–36].

1.2 Related Work

Other analysis approaches for studying radicalization have generally relied on statistical analysis of data sets that often lack details about internal group evolution leading to radicalization and violence. For example, some of the most commonly used radicalization data sets, such as Minorities At Risk Organizational Behavior (MAROB), Global Terrorism Database (GTD), Big Allied And Dangerous (BAAD), contain either post-factum recording of violent events or general descriptions of existing and established organizations [3, 33, 34]. Analyses of such data sets inherently miss critical information about internal evolution of social interaction, and thus, leave a large gap in understanding the few individuals and groups who become radicalized and why.

This paper instead is focused on the reconstruction of a detailed network leading to radicalization. Specifically, we created an agent-based visualization to evaluate the development of radical groups involved in the 1995 Paris Metro and RER bombings. This work complements the literature and software built to visualize dynamic networks, such as ORA [5], DyNet [37], SoNIA [38], Gephi [39], and R packages like *ndtv* [40]. Though not pertaining to radicalization, other researchers have also utilized similar types of narrative data in the form of transcripts of court proceedings [41] and judicial sentencing comments [42], to build social networks characterizing crimes and criminal organizations.

In the following sections, we discuss the data used for the network visualization (Sect. 2), the process of converting the narrative data into an event matrix, and assumptions underlying the visualization modeling (Sect. 3). We also discuss the technical challenges and limitations of taking this approach, with extensions to future work (Sect. 4). Finally, we conclude with a short discussion (Sect. 5).

2 Data

Over the last several years, members of our research team have collected a unique narrative text database consisting of three elements: (1) a comprehensive chronology of the major global neo-jihadi terrorist plots in the West; (2) a collection of the biographies of all people loosely connected to these plots, including those who chose not to pursue violence, and; (3) a matrix abstracting from the above data to show the connections among people within a specified time element.

The subset of the database considered in this study contains narrative descriptions of relevant biographical events for subjects involved in the 1995 Paris Metro and RER bombings. All data come from public sources and most was abstracted from the official verdicts of two major trials on the bombings in Paris. Further information was obtained from investigative research published in news magazines like *Le Point* and *Le Monde*. Additionally, one of the authors personally interviewed friends and relatives of people who became violent. All interviewees provided their consent prior to being interviewed. Most subjects in the data set were convicted of belonging to a terrorist organization. They served their terms and are now free with the exception of a very few, who received life sentences at other trials. While this study focuses on international neo-jihadi networks, further research could extend to data sets including additional varieties of terrorism by including non-religious actors.

The file is structured as individual chapters containing biographical events for each individual in the trial. Exhibit 1 contains an example of the narrative. A member of our research team, a native French speaker and terrorism researcher familiar with the 1995 Paris Metro and RER bombings case, translated the court documents from French to English and coded the events. This process ensured an expert review of the documents, though additional coders could have provided further assurances of inter-coder reliability. Given the events and actors coded were concrete events rather than subjective ratings or assessments of latent measures, we anticipated less reliability concerns and need for additional coders for this task.

The data set also includes people who do not turn to violence, thus allowing one to compare this de-facto control group with terrorists. This specificity should be viewed as an evolving pattern of activity over time. A time series of events and social connections might contribute to the understanding of the evolution of networks that eventually adopt terrorist activities to pursue their goals. Most relationships between people are temporary and cannot be assumed to be permanent as most static social network analyses (SNA) assume.

The data capture the complex and often subtle relationships among the full range of actors engaged in recent political violence events, and have proven to be useful for improving our understanding of the individuals and groups who radicalize and engage in violence. This understanding has been extracted from the data through the laborious review and analysis of the narrative data by conflict policy experts in order to inform the development of policy at multiple levels of the US government.

Exhibit 1. Example of the Original Narrative. (The actual names have been redacted)

> *"In October 1994, T shared with the friends his project to create a structure facilitating gun running to Algeria and B was tasked to modify a truck to smuggle weapons there. However, they never got going on this plan because T kept asking them to do small chores for him. Thus, B went to London and Manchester to get an envelope that contained 5,000 lb and German money. Someone named H had given him the envelope at the arrival of the train in London and sent him to Manchester. He then took him back to London and had given him a return plane ticket to France. When B got back, T counted the bills and then called someone on the telephone. Part of this money financed J and V trip to Holland".*

3 Converting Narratives into an Event Matrix

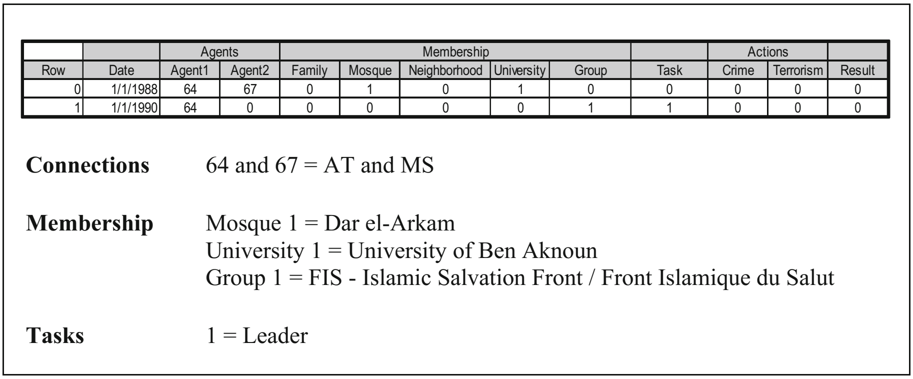
3.1 Definition of the Events and Entities

Entities and events were initially defined through a brief review of the text and then were updated as more types of events and relationships were revealed during detailed reading. The following list gives examples of events list with definitions:

- **Connection:** A simple connection/relationship where two or more people were documented as being together or talking to each other by phone.
- **Membership:** Belonging to families, places of worship or other gathering places, neighborhoods, educational institutions, and other groups/organizations.
- **Task:** The person's role or "job" within the circle of associates; examples include financier, spy, weapons appropriation, bomber, etc.
- **Action:** An activity that can be classified as a crime or terrorist act;
- **Result:** The resulting status for a person after a series of events or the end of the narrative provided; results can include arrest, death, being wanted or at large, in prison or jail, etc.

It should be noted that persons can and do have multiple connections, memberships, tasks, actions, and results (see Fig. 1 and Table 1). It should also be noted that connections can dissolve over time or due to a specific negative event, and tasks can increase in significance as a person becomes more involved in the organization.

Table 1. Data entry form and definitions



		Agents		Membership						Actions		
Row	Date	Agent1	Agent2	Family	Mosque	Neighborhood	University	Group	Task	Crime	Terrorism	Result
0	1/1/1988	64	67	0	1	0	1	0	0	0	0	0
1	1/1/1990	64	0	0	0	0	0	1	1	0	0	0

Connections 64 and 67 = AT and MS

Membership Mosque 1 = Dar el-Arkam
University 1 = University of Ben Aknoun
Group 1 = FIS - Islamic Salvation Front / Front Islamique du Salut

Tasks 1 = Leader

Table 2 shows an example of the event log file that contains the timeline, events, and actors (agents) involved in these actions. This log file serves as an input to the visualization system that illustrates the dynamic of links formation and dissolution.

Table 2. Example of events table

Date	Agents		Membership						Actions			Result
	Agent 1	Agent 2	Family	Mosque	Neighbor-hood	University	Group	Task	Move	Crime	Terrorism	
3/1/93	**30**	**20**	0	0	0	0	**2**	0	0	0	0	0
3/1/93	**30**	**49**	0	0	0	0	**2**	0	0	0	0	0
3/1/93	**49**	**36**	0	0	0	0	**4**	0	0	0	0	0
3/1/93	**56**	0	0	0	0	0	0	0	0	0	0	**6**
3/26/93	**46**	0	0	0	0	0	0	0	0	**1**	0	**2**
4/1/93	**30**	0	0	0	0	0	**−2**	0	0	0	0	0
4/1/93	**55**	**34**	0	0	0	0	0	0	0	0	0	0

"On March 1, 1993, NH (agent 30) met with SB and AM (agents 20 and 49). Also, AM (agent 49) has met with JJ (agent 36). They have established contact/associations with groups 2(FAF) and 4(Al Ansar). On the same day RR (agent 56) was sentenced (code 6) in Algeria for bombing. On March 26, 1993, SMi (agent 46) committed a crime (1) resulting in his arrest (code 2). On April 1, 1993, NH (agent 30) left FAF (−2) because it was dissolved. On the same day SR (agent 55) was in contact with HH (agent 34)".

3.2 Visualization Table Developed from Events Input

The visualization prototype produced two main summary measurements. One is a set of lists that provide main model definitions. These include: names, roles, locations, activities, schools, and families. All people mentioned in the narrative were modeled as "agents" and are assigned an agent ID. Similarly, organizations, roles, locations, schools, and activities were classified and assigned an ID. Some of these entities were added at the beginning of the process as part of initial understanding of the problem. Others were added at later stages when the coder encountered an entity not belonging to previously defined categories.

The second product is a list of events organized as a timeline. The events are described in chronological order, and the events that occur simultaneously are recorded with the same time. So far, the links between the agents are defined in a pair-wise manner, i.e., if three agents meet, this meeting is described as three pair-wise meetings occurring at the same time. The main advantage of this approach compared to other approaches is that the connections are coded either as generic (no information about the nature of a contact) or as relational, (i.e., whether the individuals are from the same neighborhood, belong to the same family, play sports together, go to the same school, mosque, etc.) (see Table 3). The relationships could be also directional, i.e., agent 1 gives orders to agent 2. Thus, tables could be analyzed with respect to the types of these relationships.

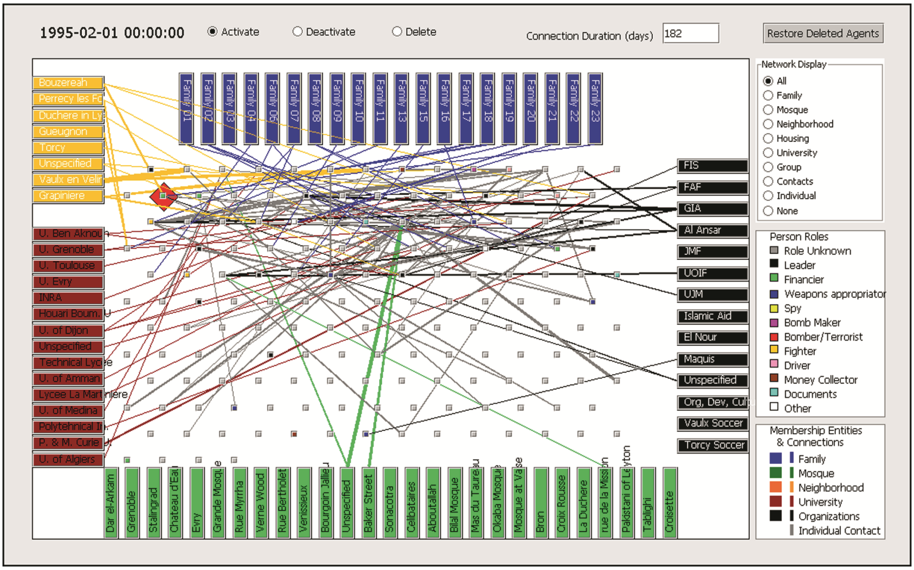
Table 3. Example of the relationship table between individuals in the group at a particular point in time (the actual names have been redacted).

	27 D-I	28 G-I	29 H-A	30 H-I	31 H-Am	32 H-As
14 B-J		1				
15 B-M	4					
16 B-D		1				
17 B-S						
18 B-A			3			
19 B-R						
20 B-As			6	4	3	3
21 B-An						
22 B-Am						
23 B-F		6				
24 C-E						
25 C-R	6	5				

3.3 Assumptions Used in Visualization Model

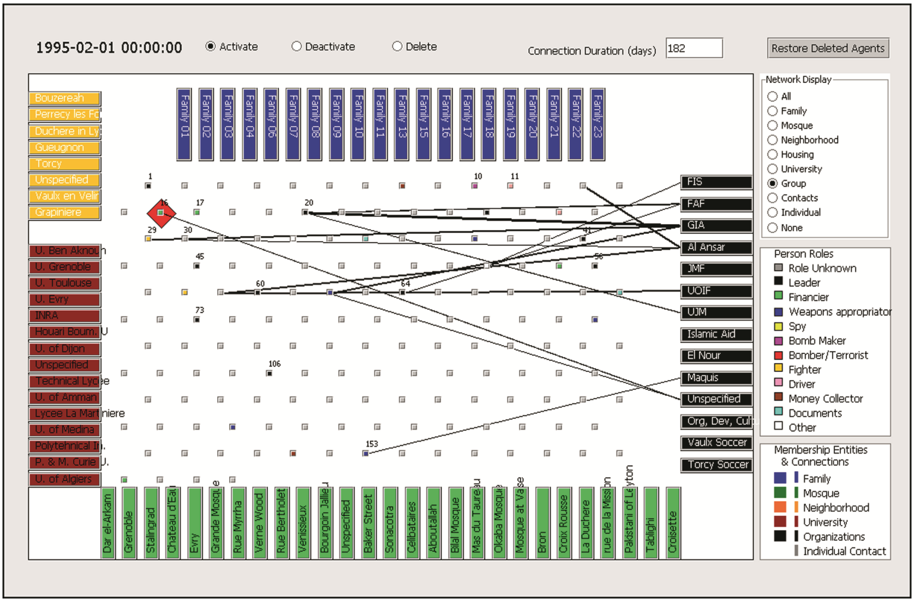
The visualization model assumes that individuals establish links to one another and these links are dynamic, i.e., they can appear, disappear, and increase in value. Although we redacted details of the events and did not record who contacted whom and for what reason, the main events have comments in the notes section. If the event of connection has been established, a link between the two individuals or entities is established. If no additional events happen during 6 months, then the link disappears. If new activities occur within a 6-month period, then the width of the link widens. Removal of the links after some time of inactivity (6 months by default) is another innovation we added to the system. If links are kept fixed, after some time the visualization becomes too busy because of the number of new people added to the connected network. If inactive links are dropped, only active links remain and the visualization shows how the dynamics of contacts change in time and who remains a consistent part of the network. Though 6 months is used in the default model, the framework allows for sensitivity analysis on this and other parameters to understand the effects of various modeling assumptions on the outcome.

Figure 1 shows an example of a screenshot denoting various activities and contacts that occurred within a 6-month period before February 1, 1995. On the sides of the screen are neighborhoods, universities, mosques, organizations, and families that were mentioned in the narratives and that have associations with the agents. The agents


1995-02-01 00:00:00 ● Activate ○ Deactivate ○ Delete Connection Duration (days) 182 Restore Deleted Agents

Network Display
● All
○ Family
○ Mosque
○ Neighborhood
○ Housing
○ University
○ Group
○ Contacts
○ Individual
○ None

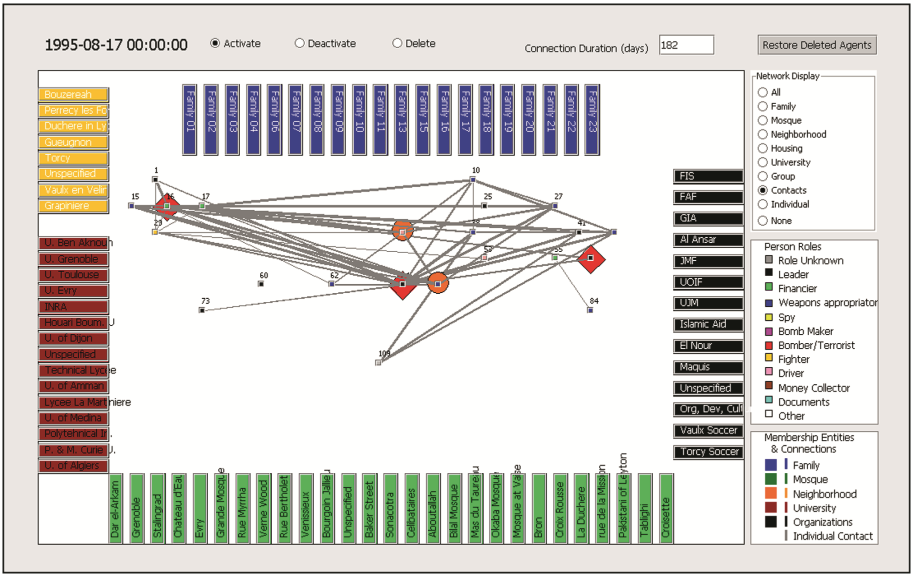
Fig. 1. Screenshot with unfiltered information caption on feb. 1, 1995 (Color figure online)


1995-02-01 00:00:00 ● Activate ○ Deactivate ○ Delete Connection Duration (days) 182 Restore Deleted Agents

Network Display
○ All
○ Family
○ Mosque
○ Neighborhood
○ Housing
○ University
● Group
○ Contacts
○ Individual
○ None

Fig. 2. Screenshot with filtered information caption on feb. 1, 1995, showing only links between individuals and organizations (Color figure online)

themselves are marked as small squares in the middle of the screen. If no role is known for a particular individual, the square color remains gray. If an agent performs a particular task or plays a specific role, then the color is changed. For example, black corresponds to the leader in a task, green connotes providing funds, red means being a bomber, etc. The number of individuals is quite large, and the graphics are very complex so that displaying the names becomes a problem. One can use filters to show specific relationships and add clarity to the analysis. Figures 2 and 3 show how information filtering can assist visualization and building social theories.


1995-08-17 00:00:00 ● Activate ○ Deactivate ○ Delete

Connection Duration (days) 182

Network Display
○ All
○ Family
○ Mosque
○ Neighborhood
○ Housing
○ University
○ Group
● Contacts
○ Individual
○ None

Fig. 3. Screenshot with filtered information caption on aug. 17, 1995, showing links between individuals who have either more than 3 connections or explicitly terrorism-related functions (e.g., weapons appropriation) (Color figure online)

Filtering connections with terrorist organizations allows one to identify leaders and trace their connections to the terrorist organizations. This filter identifies AT, SB, BB, RR, KK, SBa, AA, HH, ABr, MK, EG, ZS, and RM (Fig. 2). As time unfolds, the activity of SB, AD, and AB increases (Fig. 3). Removal of agents who do not have intensive contacts and do not perform important tasks results in the identification of the core group. Comparison of the results of our filtering procedure with the end points for the entire network of considered individuals shows that such filtering captures the main players involved in the bombing.

The major limitation of this procedure is that the data were collected after the fact and the court records were naturally focused on the main perpetrators of the terrorism rather than on the surrounding individuals. Thus, the filtering procedure based on contact intensity is aimed towards the identification of the key individuals. Alternatively, the performance-based filtering that keeps only those who conducted critical tasks is less dependent on the survey bias. The purpose of the prototype was to show that the model and approach can reproduce the knowledge that has been validated and, thus, the model has a potential to produce correct results on a new data set where the results are not as clear.

4 Limitations and Future Work

To extend and scale this approach, future research must address several theoretical and application challenges that we have encountered in this project.

Conversion of Narrative Data into Numeric Tables. Though the textual data used for this research is relatively rich in comparison to similar data sets, better narrative descriptions would help when converting text into numeric tables. Vague and conflicting language can put the annotator in the position of making important decisions about the data, thereby creating the potential for incorrect interpretation. Missing or incomplete event dates can also effect the usefulness of the data; ideally, the narrative will explicitly list the date of significant events using the month, day, and year. Leaving staff to estimate vague or missing dates can produce inaccuracies, which depending on the temporal context, can bias result of dynamic network analysis.

Additionally, cultural insight is critical for correct understanding and interpretation of textual data. For analysts new to the cultural context being discussed in the annotations, a cultural "dictionary" or ontology explaining specificity of name structures and relationships among people will save substantial amount of time for resolving ubiquities and quality control. For example, explanation of different name forms for married and unmarried persons, relationships between names and place of birth or fathers and tribal belongings could let one quickly recognize and allocate the person.

Data Entry. So far, data entry has been done manually, i.e., an assistant reads and understands the text, identifies the components (i.e., actors, roles, actions, links), and enters them into an Excel spreadsheet. As the assistant reads more of the text, he/she encounters and records more actors, event types, and locations. This creates delays and the need to reread the text for quality control purposes from the beginning after the initial pass is done. Future work can explore the use of nature language processing (NLP) techniques, such as semantic parsing, named entity recognition, and parts of speech tagging, to assist coders in identifying actors, roles, and relationships between entities. Given the opportunity here for significant advancement and rapid advances in the field, future use of NLP to assist annotation is discussed further detail in the Conclusion.

Analysis of Dynamic Data Sets. An analysis tool extending this work should be able to produce tables that could then be analyzed using general and specific statistical tools. For example, conventional SNA such as R packages, ORA, UCINET, etc., can import these social contact matrices and produce statistical analysis [5, 8, 11, 12].

Use Theories from Social Sciences. So far, the presented system has not focused on testing any specific theory but rather aimed to provide a proof-of-concept that a theory could be tested. Additionally, data mining of the narratives from different sources could provide an insight into potential hypotheses that could be tested in the future through novel data collection approach.

5 Conclusion

Understanding the process of radicalization and drivers of political violence is essential in deterring terrorist events and diminishing their effects. However, theories of radicalization have traditionally been conceptual in nature and difficult to test, validate, or falsify. This paper outlines an approach to systematically structure and visualize agents interacting over time, capturing dependences and relationships that may contribute to violent acts of terrorism. This model is a step towards a systematic approach for describing formation and interaction of human social groups. It provides a framework for examining existing theories of radicalization against structured data and the first step in the development of an agent-based network model from narrative data.

This approach is particularly useful for informing theories that emphasize social networks as an important component of radicalization, such as the work of Sageman [19, 43] and Wiktorowicz [44]. As an illustrative example, we explore the notion of “cognitive opening” [44] which states that radicalization is first primed by a person undergoing a transformative event (death of a family member, persuasion by activists, etc.) followed by exposure to a network of radicals. To test this assumption, we could compare individuals who were radicalized and committed violent acts with those who were in contact with the same network but not radicalized and see if they differed in having undergone a “cognitive opening” prior to their first (or first few) exposures. By formally modeling these interactions and assumptions, we can move to a quantified view of the radicalization process rather than merely a theoretical one. Furthermore, such a framework gets us closer to being able to evaluate interventions for disrupting violence and modeling how hypothetical scenarios may manifest, given prior data and theories of change.

Developing structured event information from narratives requires intensive manual labor. A huge opportunity for accelerating this type of work is incorporating advances in natural language processing (NLP) and information extraction (IE) to help streamline the annotation process. A range of tools and methods exist for helping extract entities and relational data from text, some even specifically designed for use in dynamic network analysis. The most directly relevant is the Automap text mining tool [14, 45], which generates meta-network information from text documents (i.e., classifying concepts into ontological categories and determining connections between these classified concepts) for use in future network analyses. More broadly, systems such as DeepDive [46], that use a combination of relation extraction and entity linking techniques for knowledge base construction, could be used in a similar fashion to help extract entities and their roles and membership in organizations from narrative texts. Current state-of-the-art in relation extraction rely heavily on machine learning techniques, with particular momentum gaining in the use of either distance supervision [47] (using relationships from existing databases to create labels) or weak supervision [48] (using domain heuristics to create labelling functions) to create noisy training data, combined with deep neural networks architectures to create refined structured classification models

[49–51]. While assessing and incorporating existing NLP and IE models was outside the scope of this study, work that relies on a combination of subject matter expertise and model generated information extraction for developing social network data will likely be a fruitful area of future research. More specifically, models that can account for temporal changes in relationships and entity types will be of particular use for the types of dynamic network analysis and visualizations presented in this study.

Lastly, while the presented study focuses on the analysis of terrorism-related data, we believe our approach could apply more broadly to any human social groups of interest, such as gangs or organized crime. Additionally, conversations on social media or forums provide a rich new source of narrative data in which this methodology could be applied. For example, codifying and analyzing social media activity of teens could help test social cognitive theories of cyberbullying or mining user networks could help understand the spread and adoption of fake news.

By providing a means to capture and visualize network dynamics, this paper contributes to the network formation and evaluation literature by providing an important first step in helping researchers assess existing theory and develop more realistic agent based models.

References

1. Morris, M.: Local rules and global properties: modeling the emergence of network structure. In: Breiger, R., Carley, K., Pattison, P. (eds.) Dynamic Social Network Modeling and Analysis, pp. 174–186. National Academy Press, Washington, DC (2003)
2. Morris, M., Handcock, M.S., Hunter, D.R.: Specification of exponential-family random graph models: terms and computational aspects. J. Stat. Softw. **24**(4), 1548 (2008)
3. Carley, K.M.: Destabilizing dynamic terrorist networks. In: Proceedings of the 8th International Command and Control Research and Technology Symposium. Conference held at the National Defense War College, Washington DC. Evidence Based Research, Track 3, Electronic Publication (2003). http://www.dodccrp.org/events/2003/8th_ICCRTS/pdf/021.pdf
4. Carley, K.M., Pfeffer, J., Liu, H., Morstatter, F., Goolsby, R.: Near real time assessment of social media using geo-temporal network analytics. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 517–524. IEEE (2013)
5. Carley, K.M.: ORA: a toolkit for dynamic network analysis and visualization. In: Alhajj, R., Rokne, J. (eds.) Encyclopedia of Social Network Analysis and Mining, pp. 1219–1228. Springer, New York (2014). https://doi.org/10.1007/978-1-4614-6170-8
6. Borgatti, S.P.: Identifying sets of key players in a network. Comput. Math. Organ. Theory **12**, 21–34 (2006)
7. Borgatti, S.P., Mehra, A., Brass, D.J., Labianca, G.: Network analysis in the social sciences. Science **323**(5916), 892–895 (2009)
8. Cranmer, S.J., Leifeld, P., McClurg, S.D., Rolfe, M.: Navigating the range of statistical tools for inferential network analysis. Am. J. Polit. Sci. **61**(1), 237–251 (2017)
9. De Marchi, S.: Computational and Mathematical Modeling in the Social Sciences. Cambridge University Press, Cambridge (2005)

10. Gunturi, V.M., Shekhar, S., Joseph, K., Carley, K.M.: Scalable computational techniques for centrality metrics on temporally detailed social network. Mach. Learn. **106**(8), 1133–1169 (2017)
11. Handcock, M.S., Hunter, D.R., Butts, C.T., Goodreau, S.M., Morris, M.: statnet: software tools for the representation, visualization, analysis and simulation of network data. J. Stat. Softw. **24**(1), 1548 (2008)
12. Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M.: ergm: a package to fit, simulate and diagnose exponential-family models for networks. J. Stat. Softw. **24**(3), nihpa54860 (2008)
13. Wei, W., Carley, K.M.: Measuring temporal patterns in dynamic social networks. ACM Trans. Knowl. Discov. Data (TKDD) **10**(1), 9 (2015)
14. Carley, K.M.: A dynamic network approach to the assessment of terrorist groups and the impact of alternative courses of action. In: Visualising Network Information (pp. KN1-1–KN1-10). Meeting Proceedings RTO-MP-IST-063, Keynote 1. RTO, Neuilly-sur-Seine (2006). http://www.rto.nato.int/abstracts.asp
15. Reed, B.J., Segal, D.R.: Social network analysis and counterinsurgency operations: the capture of Saddam Hussein. Sociol. Focus **39**(4), 251–264 (2006)
16. de Bie, J.L., de Poot, C.J., Freilich, J.D., Chermak, S.M.: Changing organizational structures of jihadist networks in the Netherlands. Soc. Netw. **48**, 270–283 (2017)
17. Ressler, S.: Social network analysis as an approach to combat terrorism: past, present, and future research. Homel. Secur. Aff. **2**(2) (2006). https://www.hsaj.org/articles/171
18. Rodríguez, J.A., Rodríguez, J.A.: The March 11 th terrorist network: in its weakness lies its strength (2005)
19. Sageman, M.: Understanding Terror Networks. University of Pennsylvania Press, Philadelphia (2004)
20. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
21. Dombroski, M., Fischbeck, P., Carley, K.: Estimating the shape of covert networks. In: Proceedings of the 8th International Command and Control Research and Technology Symposium, June 2003
22. Carley, K.M.: Estimating vulnerabilities in large covert networks. Institute of Software Research Internat, Carnegie-Mellon University, Pittsburgh (2004)
23. Diesner, J., Carley, K.M.: Using network text analysis to detect the organizational structure of covert networks. In: Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference, vol. 3. NAACSOS, July 2004
24. Borum, R.: Psychology of terrorism. Department of Mental Health Law and Policy, University of South Florida, Tampa (2007)
25. Githens-Mazer, J., Lambert, R.: Why conventional wisdom on radicalization fails: the persistence of a failed discourse. Int. Aff. **86**(4), 889–901 (2010)
26. Crossett, C., Spitaletta, J.: Radicalization: relevant psychological and sociological concepts. The John Hopkins University (2010)
27. Wheeler, S.J.: Complex environments—an alternative approach to the assessment of insurgencies and their social terrain, part 1: identifying decisive factors. National Ground Intelligence Center Assessment (2009)
28. Post, J.M.: The Mind of the Terrorist: The Psychology of Terrorism from the IRA to Al-Qaeda. St. Martin's Press, New York (2007)
29. Perliger, A., Milton, D.: From cradle to grave: the lifecycle of foreign fighters in Iraq and Syria. US Military Academy-Combating Terrorism Center West Point United States (2016)
30. Jasko, K., LaFree, G., Kruglanski, A.: Quest for significance and violent extremism: the case of domestic radicalization. Polit. Psychol. **38**(5), 815–831 (2017)

31. Perliger, A., Pedahzur, A.: Counter cultures, group dynamics and religious terrorism. Polit. Stud. **64**(2), 297–314 (2016)
32. Stern, J.: Mind over martyr: how to deradicalize Islamist extremists. Foreign Aff. **89**, 95–108 (2010)
33. Asal, V., Rethemeyer, R.K.: Dilettantes, ideologues, and the weak: terrorists who don't kill. Confl. Manag. Peace Sci. **3**, 244–263 (2008)
34. Asal, V., Rethemeyer, R.K.: The nature of the beast: terrorist organizational characteristics and organizational lethality. J. Polit. **70**(2), 437–449 (2008)
35. Lustick, I.O., Miodownik, D.: Abstractions, ensembles, and virtualizations: simplicity and complexity in agent-based modeling. Comp. Polit. **41**(2), 223–244 (2009)
36. MacKerrow, E.P.: Understanding why: dissecting radical Islamic terrorism with agent-based simulation. Los Alamos Sci. **28**, 184–191 (2003)
37. Fernández, B., Murty, V.V., Chang, W.R.: DyNet: Dynamic Network, User Manual (1989)
38. Moody, J., McFarland, D., Bender-deMoll, S.: Dynamic network visualization. Am. J. Sociol. **110**(4), 1206–1241 (2005)
39. Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. ICWSM **8**, 361–362 (2009)
40. Bender-deMoll, S.: NDTV: Network Dynamic Temporal Visualizations. R Package Version 0.10.0 (2016)
41. Baker, W.E., Faulkner, R.R.: The social organization of conspiracy: illegal networks in the heavy electrical equipment industry. Am. Sociol. Rev. **58**, 837–860 (1993)
42. Bright, D.A., Hughes, C.E., Chalmers, J.: Illuminating dark networks: a social network analysis of an Australian drug trafficking syndicate. Crime Law Soc. Change **57**(2), 151–176 (2012)
43. Sageman, M.: Leaderless Jihad: Terror Networks in the Twenty-First Century. University of Pennsylvania Press, Philadelphia (2008)
44. Wiktorowicz, Q.: Radical Islam rising: Muslim extremism in the West. Rowman & Littlefield Publishers, Lanham (2005)
45. Carley, K.M., Columbus, D., Azoulay, A.: Automap user's guide 2012 (No. CMU-ISR-12-106). Institute of Software Research Internat, Carnegie-Mellon University, Pittsburgh (2012)
46. Niu, F., Zhang, C., Ré, C., Shavlik, J.W.: DeepDive: web-scale knowledge-base construction using statistical learning and inference. In: VLDS, vol. 12, pp. 25–28 (2012)
47. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2, pp. 1003–1011. Association for Computational Linguistics, August 2009
48. Ratner, A., Bach, S.H., Ehrenberg, H., Fries, J., Wu, S., Ré, C.: Snorkel: rapid training data creation with weak supervision. arXiv preprint arXiv:1711.10160 (2017)
49. Kumar, S.: A survey of deep learning methods for relation extraction. arXiv preprint arXiv:1705.03645 (2017)
50. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 2124–2133 (2016)
51. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1753–1762 (2015)

Terrorist Network Monitoring with Identifying Code

Arunabha Sen, Victoria Horan Goliber, Chenyang Zhou, and Kaustav Basu(✉)

Arizona State University, Tempe, AZ 85287, USA
{asen,czhou24,kaustav.basu}@asu.edu, victoria.goliber@gmail.com

Abstract. On multiple incidences of terrorist attacks in recent times across Europe, it has been observed that the perpetrators of the attack were in the suspect databases of the law enforcement authorities, but weren't under active surveillance at the time of the attack due to resource limitations on the part of the authorities. As the suspect databases in various European countries are very large, and it takes significant amount of technical and human resources to monitor a suspect in the database, monitoring all the suspects in the database may be an impossible task. In this paper, we propose a scheme utilizing *Identifying Codes* that will significantly reduce the resource requirement of law enforcement authorities, and will have the capability of uniquely identifying a suspect in case the suspect becomes *active* in planning a terrorist attack. The scheme relies on the assumption that, when an individual becomes active in planning a terrorist attack, his/her friends/associates will have some inkling of the individuals plan. Accordingly, even if the individual is not under active surveillance by the authorities, but the individual's friends/associates are, *the individual planning the attack can be uniquely identified*. We applied our technique on two terrorist networks, one involved in an attack in Paris and the other involved in the 9/11 attack. We show that, in the Paris network, if 5 of the 10 individuals were monitored, the attackers most likely would have been exposed. If only 15 out of the 37 individuals involved in the 9/11 attack were under surveillance, specific individuals involved in the planning of the 9/11 attack would have been exposed.

Keywords: Terrorist network · Identification code
Computational complexity · Approximation algorithm

1 Introduction

Terrorist attacks are on the rise, all across the world. The problem is more acute in Europe, in general, and France in particular. Since the devastating attack in multiple locations in Paris in November 2015, there have been at least six other incidences of terrorist attacks in France [1]. It is now known that on multiple occasions, the perpetrators of the attack were in the suspect database of the law enforcement authorities, but they weren't under active surveillance at the time of the attack due to resource limitations on the part of the law enforcement authorities.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 329–339, 2018.
https://doi.org/10.1007/978-3-319-93372-6_36

According to reports, the two suspects in the attack against French police on April 20, 2017, on the Champs-Elysees, were known to French anti-terrorism authorities. The suspected gunman, 39-year-old French national Karim Cheurfi, had been released from prison the previous year for an earlier attempt to shoot police, after being caught in a stolen car. On that day, he succeeded in killing one officer and injuring two others before being shot dead. The November 2015 attacks in Paris that claimed a total of 130 lives, involved a small network of ISIS-linked terrorists in France and Belgium. Of the 10 individuals involved, several were known to authorities. When 12 people were killed at the Paris headquarters of Charlie Hebdo, a satirical magazine, all three of the terrorists had been under close watch. Cherif Kouachi, Said Kouachi and Amedy Coulibaly were under police surveillance for three years, but eventually dropped in the summer of 2014 only months before the deadly January 2015 attack.

The news organization Politico, [2] reported in October 2016 that the French authorities were monitoring around 15,000 individuals who were suspected of being radical Islamists. The Politico report was based on an earlier publication in the French journal, La Journal du Dimanche. The ABC news affiliated TV station WJLA in Washington D.C., reported in 2017 that, the list has tripled over the last two years [3]. The database is managed by France's Counter-Terrorism Coordination Unit. Obviously, the resources and manpower needed to keep all the terror suspects under surveillance is enormous and often are way beyond the available resources of any local law enforcement authority. It has also been reported that the French Centre for the Analysis of Terrorism has determined that it takes as many as 20 agents per suspect to conduct 24-h surveillance.

In order to address this problem, we propose a scheme utilizing *Identification Code* that will significantly reduce the resource requirement of the law enforcement authorities, and will uniquely identify a suspect in case, the suspect becomes *active* in planning a terrorist attack. The scheme relies on the assumption that, when an individual becomes active in planning a terrorist attack, his/her friends/associates will have some inkling of the individuals plan. Accordingly, even if the individual is not under active surveillance by the authorities, but the individual's friends/associates are, *the individual planning the attack can be uniquely identified.* We applied our technique on the terrorist network involved in the Paris attack and the one involved in the 9/11 attack. In the Paris network, if 5 of the 10 individuals were monitored, the attackers most likely would have been exposed. We show that if only 15 out of the 37 individuals involved in the 9/11 attack were under surveillance, specific individuals in the planning of the 9/11 attack would have been exposed.

2 Identifying Codes

The notion of *Identifying Codes* [4] has been established as a useful concept for optimizing sensor deployment in multiple domains. In this paper, we use Identifying Code of the *simplest form* and define it as follows. *A vertex set V' of a graph $G = (V, E)$ is defined as an Identifying Code Set (ICS) for the vertex* *set V, if for all $v \in V$, $N[v] \cap V'$ is unique where, $N[v] = v \cup N(v)$ and $N(v)$ represents the set of nodes adjacent to v in $G = (V, E)$.* The *Minimum Identifying Code Set* (MICS) problem is to find the Identifying Code Set of *smallest cardinality.* The vertices of the set V' may be viewed as *alphabets* of the code, and the *string* made up with the alphabets of $N[v]$ may be viewed as the unique "code" for the node v. For instance, consider the graph $G = (V, E)$ shown in Fig. 1. In this graph $V' = \{v_1, v_2, v_3, v_4\}$ is an ICS as it can be seen from Table 1 that $N[v] \cap V'$ is *unique* for all $v_i \in V$.

Two nodes $u, v \in V$ are said to be "twins" if $N[v] = N[u]$. It may be noted that Identifying Code for a graph $G = (V, E)$ does not exist if the graph has "twins".

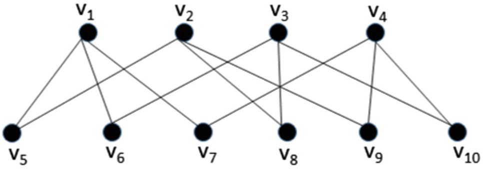


Fig. 1. Graph with Identifying Code Set v_1, v_2, v_3, v_4

Table 1. $N[v] \cap V'$ results for all $v \in V$ for the graph in Fig. 1

$N[v_1] \cap V' = \{v_1\}$	$N[v_2] \cap V' = \{v_2\}$
$N[v_3] \cap V' = \{v_3\}$	$N[v_4] \cap V' = \{v_4\}$
$N[v_5] \cap V' = \{v_1, v_2\}$	$N[v_6] \cap V' = \{v_1, v_3\}$
$N[v_7] \cap V' = \{v_1, v_4\}$	$N[v_8] \cap V' = \{v_2, v_3\}$
$N[v_9] \cap V' = \{v_2, v_4\}$	$N[v_{10}] \cap V' = \{v_3, v_4\}$

In the last few years a number of researchers have studied Identifying Codes and its applications in sensor network domains. Karpovsky *et al.* [4] introduced the concept of Identifying Codes in [4] and provided results for Identifying Codes for graphs with specific topologies, such as binary cubes and trees. Using Identifying Codes, Laifenfeld and Trachtenburg studied covering problems in [5] and joint monitoring and routing in wireless sensor networks in [6]. Ray *et al.* in [12] generalized the concept of Identifying Codes, to incorporate robustness properties to deal with faults in sensor networks. Charon *et al.* in [7,8], studied complexity issues related to computation of minimum Identifying Codes for graphs and showed that in several types of graphs, the problem is NP-hard. Approximation algorithms for computation of Identifying Codes for special types of graphs are presented in [10,11]. Auger [9] shows that the problem can be solved in linear time if the graph happened to be a tree, but even for a planar graph the problem remains NP-complete.

`MICS computation as a Graph Coloring with Seepage (GCS) Problem:` The MICS computation problem can be viewed as a novel variation of the standard Graph Coloring problem. We will refer to this version as the *Graph Coloring with Seepage (GCS)* problem. In the standard graph coloring problem, when a color is *assigned* (or injected) to a node, only that node is colored. The goal of the standard graph coloring problem to use as few distinct colors as possible such that (i) every node receives a color, and (ii) no two adjacent nodes of the graph have the same color. In the GCS problem, when a color is assigned (or injected) to a node, not only that node receives the color, the color also *seeps* into all the adjoining nodes. As a node v_i may be adjacent to two other nodes v_j and v_k in the graph, if the color red is injected to v_j, not only will v_j become red, but also v_i will become red as it is adjacent to v_j. Now if the color blue is injected to v_k, not only will v_k become red, but also, the color blue will seep in to v_i as it is adjacent v_k. Since v_i was ready colored red (due to seepage from v_j), after color seepage from v_k, its color will be a *combination of red and blue, i.e., purple.* At this point all three nodes v_i, v_j and v_k have a color and all of them have distinct colors (red, blue and purple). The goal of the GCS problem is to inject colors to as few nodes as possible, such that (i) every node receives a color, and (ii) no two nodes of the graph have the same color.

Suppose that the node set V' is an ICS of of a graph $G = (V, E)$ and $|V'| = p$. In this case if p distinct colors are injected to the nodes of V' (one distinct color to one node of V'), then as by the definition of ICS for all $v \in V$ if $N[v] \cap V'$ is unique, all nodes of $G = (V, E)$ will be colored and no two nodes will have the same color. Thus computation of MICS will be equivalent to solving the GCS problem.

3 Monitoring Terrorist Networks with Identifying Codes

In this section, we present two terrorist networks. Figure 2 is the 10 man Paris Network, involved in the devastating attack across multiple locations in Paris, in November 2015 and Fig. 3 is the 37 man 9/11 Network. Our goal is to use the concept of Identifying Code to monitor terror networks.

For the Paris network, the minimum Identifying Code set is $2, 4, 6, 7, 8$, i.e., 5 colors are needed to uniquely monitor this network. In Table 2, the first column denotes the nodes, the following five columns denote the colors received by the nodes and the final column is the number of such colors received by a particular node. 1^* indicates the color injection points and 1 denotes the seepage, of a particular color, into adjacent nodes. Since 5 colors are needed to uniquely monitor the Paris network, we will have 5 injection points. As such, nodes $2, 4, 6, 7, 8$ have colors A, B, C, D, E injected into them respectively. Looking at the table, node 1 has received the color B, and thus, the total color is 1. Node 2 has been injected with color A, and this color has seeped into its adjacent nodes, $6, 7, 8$. Colors C, D and E seep into node 2 from nodes 6, 7 and 8 respectively, and hence, the total color, of node 2, is 4. The rest of the table is constructed similarly.

The unique assignment of colors of the nodes in the Paris network, can be easily verified using Table 3. The left hand side of the table denotes the colors

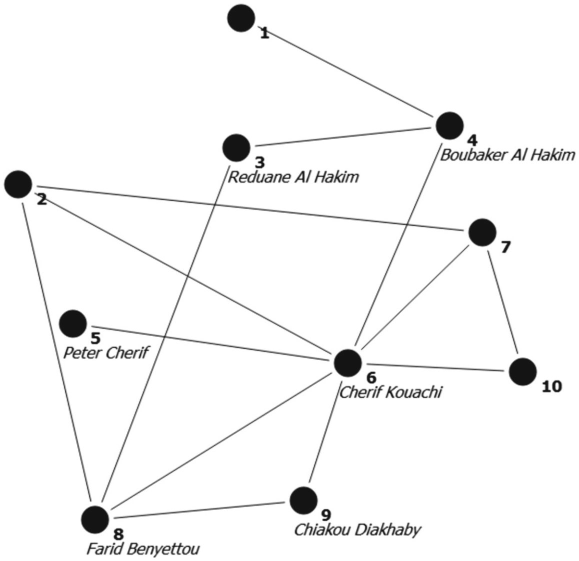


Fig. 2. Paris terrorist network

received by every node. The right hand side of the table contains the colors sorted lexicographically, to denote that each color/color combination has occurred only once. In other words, every node has a unique color string.

Figure 3 is the 9/11 terrorist network. 15 nodes namely, $2, 4, 6, 9, 12, 16, 19, 20, 21, 22, 23, 26, 31, 36, 37$, are required to uniquely monitor the entire network. In this network, nodes 25 and 33 are "twins". As mentioned in Sect. 2, Identifying Codes do not exist for graphs with "twins". Hence, we have collapsed nodes 25 and 33 into a single node, denoted by the number 25. The implication of this in the real world is that, Abdussattar Shaikh and Osama Awadallah know the exact same people in the network. Thus, we can monitor one or the other, and row 33 in Table 4 has been filled with X. Similar to Tables 2 and 4 denotes the color and total color each node has received. Table 5 can be used to verify that 15 colors, namely alphabets A through O, are needed to uniquely monitor this network.

4 Fault Tolerant (Robust) Identifying Codes

Previously, we assumed that when a suspect becomes active in planning a terrorist attack, all of the suspects' friends/associates will have some inkling regarding the intent of the suspect. Based on this assumption, we decided on the individuals to monitor so that if any one in the network is planning an attack, that individual can be uniquely identified. However, this assumption may be too strong and in reality may not always be true. If a signal (intent of a attack) does not reach the monitor (policeman watching a friend/associate) and as a

Table 2. Color assignment at nodes after seepage in the Paris Network

Node no.	A	B	C	D	E	Total color
1		1				1
2	1*		1	1	1	4
3		1			1	2
4		1*	1			2
5			1			1
6	1	1	1*	1	1	5
7	1		1	1*		3
8	1		1		1*	3
9			1		1	2
10			1	1		2

Table 3. Node color assignment in the Paris Network

Node	String	Node	String	String	Node	String	Node
1	B	6	ABCDE	ABCDE	6	BC	4
2	ACDE	7	ACD	ACD	7	BE	3
3	BE	8	ACE	ACDE	2	C	5
4	BC	9	CE	ACE	8	CD	10
5	C	10	CD	B	1	CE	9

consequence, the monitor cannot convey the information to the control center, the individual planning an attack cannot be uniquely identified. This scenario is equivalent to the scenario where the signal correctly reaches the monitor, but the monitor (due to some malfunction) fails to convey the information to the control center. Accordingly, it can be concluded that the system discussed in earlier sections does not have any fault-tolerant capability, in the sense that, if a monitor (sensor) fails to convey any suspicious behavior to the control center, the individual planning the attack cannot be uniquely identified. However, the inability to uniquely identify an individual planning an attack can be overcome by designing a more robust or fault-tolerant system. In this context, a system will be considered more *robust* if it can uniquely identify the individual planning the attack, in spite of failure of one or more monitors to report any suspicious activity to the control center. The failure of the monitor to report any suspicious activity may be either due to signal not reaching the monitor or due to a failure of the monitor to convey the correctly received signal to the control center. An Identifying Code set that can tolerate up to k monitor failures will be referred to as *k Fault-tolerant (or Robust) Identifying Code*. Since the term *Robust Identifying Code* has already been used to describe a different problem in [13], we will refer to problem under discussion in the paper as the *k Fault-tolerant*

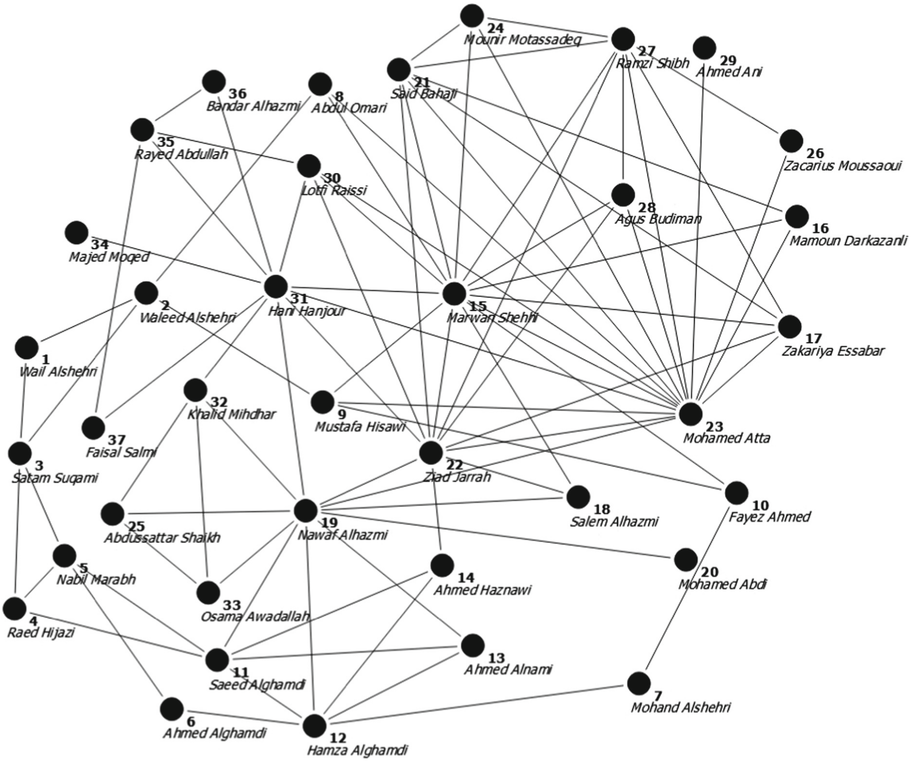


Fig. 3. 9/11 terrorist network

Identifying Code. In the following subsection we establish a *lower bound* on the number of monitors that will be needed to design a *k Fault-tolerant Identifying Code* based system. It may be noted that although the authors in [12] also considered robustness issues in Identifying Code context, the results presented in this paper are different from the ones presented in [12]. In the following section, we show that a lower bound (n_k) on the size of k-fault tolerant Identifying Code, for a graph with n nodes is $n_k \geq \lceil log\ (N+1) \rceil + d - 1$, where $d \geq 2k+1$. No such lower bound result was presented in [12].

4.1 Lower Bound on the Size of k Fault-Tolerant Identifying Codes

Suppose that the graph contains N nodes. In order to be uniquely identifiable, each node must have a *unique signature/code* (or color/string) associated with it. With n bits, 2^n unique bit strings can be generated. However, one of these string comprises of all 0 bits, which cannot be the valid signature for a node as a string comprising of all 0 s also represent a scenario where no node produces a signal. Accordingly, a lower bound on the size of Identifying Code for a N node system will be the smallest n such that $2^n \geq N+1$ or $n \geq \lceil log\ (N+1) \rceil$.

Table 4. Color assignment at nodes after seepage in the 9/11 Network

Node no.	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Total color
1	1															1
2	1*			1												2
3	1	1														2
4		1*														1
5		1	1													2
6			1*		1											2
7					1											1
8	1										1					2
9	1			1*							1					3
10				1												1
11		1			1		1									3
12			1		1*		1									3
13					1		1									2
14					1					1						2
15				1		1			1	1	1		1			6
16						1*			1		1					3
17									1	1	1					3
18							1			1						2
19					1		1*	1		1	1		1			6
20							1	1*								2
21						1			1*	1	1					4
22							1		1	1*	1		1			5
23				1		1	1		1	1	1*	1	1			8
24									1		1					2
25							1									1
26											1	1				2
27									1	1	1	1				4
28										1	1					2
29											1					1
30										1	1		1			3
31							1			1	1		1*	1	1	6
32							1						1			2
33	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
34													1			1
35													1	1	1	3
36													1	1*		2
37													1		1*	2

Table 5. Node color assignment in the 9/11 Network

Node	String	Node	String	String	Node	String	Node
1	A	20	GH	A	1	G	25, 33
2	AD	21	FIJK	AB	3	GH	20
3	AB	22	GIJKM	AD	2	GIJKM	22
4	B	23	DFGIJKLM	ADK	9	GJ	18
5	BC	24	IK	AK	8	GJKMNO	31
6	CE	25	G	B	4	GM	32
7	E	26	KL	BC	5	IJK	17
8	AK	27	IJKL	BEG	11	IJKL	27
9	ADK	28	JK	CE	6	IK	24
10	D	29	K	CEG	12	JK	28
11	BEG	30	JKM	D	10	JKM	30
12	CEG	31	GJKMNO	DFGIJKLM	23	K	29
13	EG	32	GM	DFIJKM	15	KL	26
14	EJ	33	G	E	7	M	34
15	DFIJKM	34	M	EG	13	MN	36
16	FIK	35	MNO	EGHJKM	19	MNO	35
17	IJK	36	MN	EJ	14	MO	37
18	GJ	37	MO	FIJK	21		
19	EGHJKM			FIK	16		

Although with $n = \lceil log\ (N+1) \rceil$, unique codes for all N nodes can be generated, minimum *Hamming Distance* [14] between a pair of codes in this case will be 1. However, with minimum code separation distance being equal to 1, it may be impossible to distinguish between two nodes, even when just one monitor is faulty. Consider two nodes v_1 and v_2 whose unique code/signature is the strings/colors A and AB. In this scenario, if the monitor B malfunctions, there will be no way for the control center to distinguish between the nodes v_1 and v_2. In order to be able to distinguish between two nodes when at most one monitor is faulty, the minimum code separation distance must be equal to 3. This is true, as the minimum code separation distance being equal to 2 will not be sufficient to distinguish between two nodes. Consider two nodes v_1 and v_2, whose unique code/signature is the strings/colors AB and AC. The Hamming distance between the codes is 2. However, in this case, if the control center receives a signal A, it will not be able to distinguish if its for node v_1 with monitor B being faulty or v_2 with monitor C being faulty. In order to design a k fault-tolerant system, the minimum separation (Hamming) distance d between a pair of codes must be at least $2k + 1$, i.e., $d \geq 2k + 1$, or $k \leq \lfloor (d - 1)/2 \rfloor$,

Theorem 1. *A lower bound on the size of Identifying Code for a N node system, that guarantees the minimum code separation distance of at least d, is the smallest n_k, such that $n_k \geq \lceil log\ (N+1) \rceil + d - 1$, where $d \geq 2k+1$.*

Proof. Suppose that to uniquely distinguish N nodes in a k fault-tolerant system, each node must have a *unique signature/code* (or color/string) of n_k bits. Suppose that the bit string is represented by $b_{n_k-1}, b_{n_k-2}, \ldots, b_1, b_0$. With string length n_k, 2^{n_k} strings can be generated, but their minimum code separation distance will not be d (it will be 1). Two strings (codes) associated with two nodes will be at least distance d apart, only if at least in d positions of the corresponding strings, the bits are *inverse* of one another. One example of two such strings will be $b_{n_k-1}, b_{n_k-2}, \ldots, b_d, b_{d-1}, b_{d-2}, \ldots, b_1, b_0$ and $b_{n_k-1}, b_{n_k-2}, \ldots, b_d, \bar{b}_{d-1},\ \bar{b}_{d-2}, \ldots, \bar{b}_1, \bar{b}_0$. By plugging-in any value (0 or 1), in the partial string $b_{n_k-1}, b_{n_k-2}, \ldots, b_d$ (partial string of length $n_k - 1 - d + 1 = n_k - d$), 2^{n_k-d} string can be generated. Each such string when concatenated with partial strings $b_{d-1}, b_{d-2}, \ldots, b_1, b_0$ and $\bar{b}_{d-1}, \bar{b}_{d-2}, \ldots, \bar{b}_1, \bar{b}_0$, will create $2^{n_k-d} \times 2 = 2^{n_k-d+1}$ strings of length n_k, whose minimum code separation distance will be d. Accordingly, a lower bound on the size of Identifying Code for a N node system that guarantees the minimum code separation distance at least d will be the smallest n_k, such that $2^{n_k-d+1} \geq N+1$ or $n_k - d + 1 \geq \lceil log\ (N+1) \rceil$, or $n_k \geq \lceil log\ (N+1) \rceil + d - 1$, or $n_k \geq \lceil log\ (N+1) \rceil + 2k$.

It may be noted that the lower bound result presented in Theorem 1 is independent of the network topology. If the graph $G = (V, E)$ is "twin-free", then if N different colors are injected at N different nodes of the graph, every single node will have a unique color. Thus, N can be viewed as an upper bound of the size of the independent code set. Just like the lower bound result of Theorem 1, this upper bound is also independent of the network topology.

5 Conclusion

This paper discusses the use of Identifying Code to monitor terrorist networks. We have discussed the concept, using two terror networks, the 9/11 and the Paris network. It has been shown that law enforcement authorities need not monitor the entire network but a subset of the network, which significantly reduces resources and costs involved in allocating these resources. In the Paris network, if 5 of the 10 individuals were monitored, the attackers most likely would have been exposed. If only 15 out of the 37 individuals involved in the 9/11 attack were under surveillance, specific individuals in the planning of the 9/11 attack would have been exposed. To the best of our knowledge, this is the first work to discuss the concept of using Identifying Code for monitoring purposes.

References

1. Wikipedia Article: List of Terrorist Incidents in France. Accessed 26 Dec 2017
2. Cooper, H.: 15,000 on French terror watchlist: report. Politico (2016)
3. Bernstein, L.: From France to the U.S. terrorists 'known to authorities' carry out deadly attacks. WJLA (2017)
4. Karpovsky, G.M., Chakraborty, K., Levitin, L.B.: On a new class of codes for identifying vertices in graphs. IEEE Trans. Inf. Theory **44**(2), 599–611 (1998)
5. Laifenfeld, M., Trachtenburg, A.: Identifying codes and covering problems. IEEE Trans. Inf. Theory **54**, 3929–3950 (2008)
6. Laifenfeld, M., Trachtenburg, A., Cohen, R., Starobinski, D.: Joint monitoring and routing in wireless sensor networks using robust identifying codes. Mob. Netw. Appl. **14**, 415–432 (2009)
7. Charon, I., Hudry, O., Lobstein, A.: Identifying and locating-dominating codes: NP-completeness results for directed graphs. IEEE Trans. Inf. Theory **48**, 2192–2200 (2002)
8. Charon, I., Hudry, O., Lobstein, A.: Minimizing the size of an identifying or locating-dominating code in a graph is NP-hard. Theor. Comput. Sci. **290**, 2109–2120 (2003)
9. Auger, D.: Minimal identifying codes in trees and planar graphs with large girth. Eur. J. Comb. **31**, 1372–1384 (2010)
10. Xiao, Y., Hadjicostis, C., Thulasiraman, K.: The d-identifying codes problem for vertex identification in graphs: probabilistic analysis and an approximation algorithm. In: Chen, D.Z., Lee, D.T. (eds.) COCOON 2006. LNCS, vol. 4112, pp. 284–298. Springer, Heidelberg (2006). https://doi.org/10.1007/11809678_31
11. Suomela, J.: Approximability of identifying codes and locating-dominating codes. Inf. Process. Lett. **103**, 28–33 (2007)
12. Ray, S., Starobinski, D., Trachtenburg, A., Ungrangsi, R.: Robust location detection in emergency sensor networks. IEEE J. Sel. Areas Commun. **22**, 1016–1025 (2004)
13. Honkala, I., Karpovsky, M.G., Levitin, L.B.: On robust and dynamic identifying codes. IEEE Trans. Inf. Theory **52**(2), 599–612 (2006)
14. Baer, J.-L.: Computer Systems Architecture. Computer Science Press, Rockville (1980)

Implicit Terrorist Networks: A Two-Mode Social Network Analysis of Terrorism in India

Rithvik Yarlagadda[1], Diane Felmlee[2], Dinesh Verma[3], and Scott Gartner[2](✉)

[1] University of Maryland, College Park, MD, USA
ryarlaga@terpmail.umd.edu
[2] Pennsylvania State University, University Park, PA, USA
{dhf12,ssg13}@psu.edu
[3] IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
dverma@us.ibm.com

Abstract. Recent studies examine factors that lead to the emergence of terrorism and why some locations are more frequently targeted than others. However, much of the research assumes that terrorist incidents and groups are independent. We show that the assumption of independence is not always valid. Instead, we identify the conditions under which terrorist groups share choices over target locations, forming Implicit Terrorist Networks. We demonstrate the utility of this approach by examining Islamic terrorism in India (1990–2015). Using a two-mode network approach, we find that violent target locations are not independent of each other, but instead have a tendency to occur in clusters. The results highlight the patterns by which India has been targeted by a number of active, Islamic terrorist organizations over a 25-year period. More generally, our study: (1) demonstrates the utility of employing an Implicit Network approach to understanding terrorism, (2) shows that cluster analysis can assist in identifying terror group aliases, (3) identifies unexpected locations for violence that may indicate the involvement of external factors, providing leads for counter terrorism efforts, and (4) provides a tool for identifying the structures underlying patterns of global terrorism.

Keywords: Social network analysis · Terrorism · India

1 Introduction

Criminal groups, such as terrorists and organized crime, sometimes make formal agreements with each other [1], such as the 2015 alliance between Boko Haram and ISIS. There are, however, obvious deterrents to such agreements: they are unenforceable, generate evidence, and illegal organizations tend to be unstable and have rapidly changing alliances. Groups can collaborate in other, less formal but more observable ways, by operating in the same way or in similar locations, thus forming ***Implicit Terrorist Networks*** [2,3]. An implicit terrorist network

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 340–347, 2018.
https://doi.org/10.1007/978-3-319-93372-6_37

represents a congruence of violent behavior from two or more terrorist groups that may or may not result from their decision to act together. In implicit networks, terrorist groups seemingly or actually cooperate together without mechanisms for direct collaboration. We examine implicit networks among Islamic terrorist groups in India between 1990 and 2015.

2 Theory

The locations of terrorism vary widely [4]. We examine three key factors motivating the selection of targets. First, terrorist groups favor targets that result in high civilian casualties [5]. Thus, those targets that likely lead to high casualties are more attractive and more likely to be selected by extremist groups for violence. The key here is crowds and cities versus individuals and rural countryside. Second, terrorism often remains the only option available for actors unable to project force through more conventional military means against a stronger foe [6]. Terror groups thus select less protected, soft targets given the tremendous asymmetry of power. A third factor concerns the symbolic importance of a state, city, or building. Symbolic targets are selected by terrorists to garner attention and spread their political message [7]. We explore these factors and the insights they provide on implicit networks with the use of two-mode social network analysis. We investigate both the network centrality of terrorist groups and their attack locations, and the subgroup formation of terrorist attacks within India.

3 Methods

3.1 Data

We employ the Global Terrorism Database/GTD [8], an open-source database maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START); and South Asian Terrorism Portal/SATP [9]. SATP provides data on 15 Indian Islamist groups. GTD data consist of 51 locations and 154 events, representing all terrorist attacks, bombing or explosions in India conducted by groups between 1990 and 2015. We considered multiple strikes in a given location on the same day as one single attack. Also, an attack claimed by multiple terrorist groups at a specific location was coded as an incident for both groups. Despite the high level of activity (7th highest in the world, GTI 2016), terrorism within India using social network approaches has been rarely studied [10,11].

3.2 Two-Mode Network Analysis

Two-mode networks consist of two types of nodes: "actors"and "events," where direct ties exist only between nodes belonging to different types [3,12–14]. We constructed a 51 by 15 two-mode, or bipartite, network, where rows represent the 51 locations that experienced terrorist incidents and the 15 columns represent the extremist organizations. An entry in the $m \times n$ matrix cell represents the number of attacks in location m by actor n (see Fig. 1). For the centrality analysis, this matrix was dichotomized to a binary matrix, with a *one* indicating the occurrence of any attack, and *zero* otherwise. In order to perform the analyses, the two-mode network is converted into two one-mode, symmetric networks, with one of them representing target locations and the other representing extremist organizations. The ties in the two resulting one-mode networks indicate at least one incident of an attack by common terrorist groups between any two pairs of locations, and at least one commonly attacked location between any two pairs of organizations. We focus on degree centrality (the most direct measure of groups' connection to violence) and include three other common measures as a robustness check.

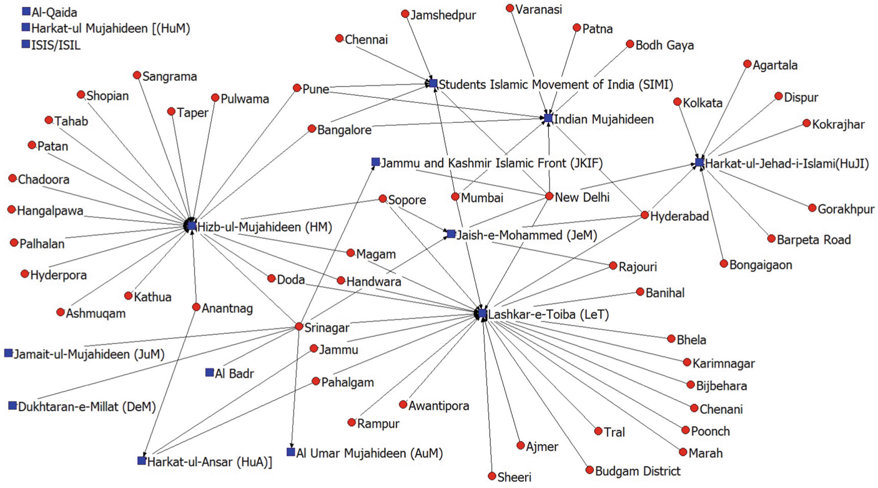


Fig. 1. Two-mode network (blue rectangles/groups; red ovals/locations attacked) (Color figure online)

4 Results

4.1 Centrality of Organizations and Locations [15,16]

As seen in Table 1, Lashkar-e-Taiba (LeT) exhibits the highest normalized degree centrality of all the terrorist organizations, with a score of 0.471. This reflects the fact that LeT perpetrated the most number of attacks in the dataset. Following LeT in degree centrality is Hizb ul-Mujahidin (HM), which has a slightly lower score of 0.392. Each terrorist group is ranked according to their level of centrality ('1' = highest rank of centrality) on four different network centrality measures (Table 1). The extremist groups, LeT and HM, achieve the two highest ranks across all four centrality measures. The high level of centrality of these two groups is evident in the two-mode network graph, as well, with numerous edges emanating from each of these groups towards multiple target locations (see Fig. 1). Note that there are three isolate organizations in the data set, with centrality scores of '0', which likely operate at a distance.

The normalized two-mode degree centrality measure for geographical targets indicates that Srinagar, the capital of Jammu & Kashmir, is highly conflictual and violent (in analysis not shown here). With the highest level of degree centrality (0.533), Srinagar is also the most central and highly attacked location. It is followed in degree centrality by the capital of India, New Delhi (0.400). These findings mean that Srinagar was attacked by more than 50 percent of the Islamic terrorist organizations and New Delhi by 40 percent. Places such as Agartala (0.067) and Taper (0.067), on the other hand, exhibit the lowest degree centrality, indicating that they were struck by just one group. It is not the case, however, that all state capitals are highly central or particularly susceptible to violence.

Table 1. Two mode centrality for terrorist groups.(*Degree centrality* is the number of ties directly connected to an actor regardless of the direction of those ties, identifying the groups carrying out the most attacks and the locations targeted most frequently. *Eigenvector* measures the influence of a node in the overall network (Newman 2008), reflecting the groups sway over other terrorist groups. *Closeness centrality* is the average length of the shortest path between the node and all other nodes in the graph, indicating how close a specific group is to the sites or activities of other groups. *Betweenness centrality* measures how often a node lies in the shortest path between other nodes, capturing how a group affects the overall connections between nodes and locations.)

Group	Degree	Eigen	Between	Closeness	D-Rank	E-Rank	B-Rank	C-Rank
LeT	0.471	0.768	0.622	0.5	1	1	1	1
HM	0.392	0.493	0.523	0.381	2	2	2	2
HuJI	0.176	0.122	0.405	0.197	3	6	7	3
I Mujah	0.157	0.205	0.467	0.117	4	4	4	4
SIMI	0.118	0.152	0.457	0.075	5	5	5	5
JeM	0.098	0.234	0.473	0.039	6	3	3	6
JKIF	0.039	0.111	0.457	0.01	8	7	6	7
HuA	0.059	0.072	0.374	0.006	7	8	12	8
Al Badr	0.02	0.059	0.385	0	9	9	8	9
AUM	0.02	0.059	0.385	0	10	10	9	10
DeM	0.02	0.059	0.385	0	11	11	10	11
JuM	0.02	0.059	0.385	0	12	12	11	12

4.2 Subgroups

For cluster analysis we constructed two sets of edge lists and two one-mode networks derived from the two-mode matrix. These one-mode networks are dichotomized, but we use the number of attacks to inflate node sizes in the network graphs. We employed a Clauset-Newman-Moore hierarchical clustering algorithm based on modularity to uncover the clustering of actors and events [17].

As depicted in Fig. 2, we identify three primary target subgroup clusters (G1, G2, G3). State capitals are clustered together in G2 (except for Srinagar), suggesting they represent a strategic target type. Locations outside of the conflicted state of Jammu & Kashmir (J&K) in India tend to be clustered in one of the subgroups. The geographic locations within J&K are clustered in G1 and G3. Two exceptions are Karimnagar and Ajmer, which are geographically located outside of J&K, but placed within G3, suggesting these cities act as transit points for terrorist groups active in J&K. An intervention strategy that increases state vigilance in Karimnagar and Ajmer may have an impact in reducing terrorism in J&K. The clustering of many cities within J&K together into subgroups underscores the notion that geographical location, as well as regional conflict, are critical for making an area attractive as a terrorist target. The findings also suggest that the two clusters, G1 and G3, are "in-effect"the same, forming an implicit network among themselves that may be analyzed as a single group.

Clustering also occurs among the terrorist groups (not shown). There were two main clusters, one consisting of the domestic organizations and the other international groups. The general clustering pattern for the non-isolates highlights the tendency of domestic groups and international entities to target similar locations within categories of organization (domestic/international), and to strike at locations that differ from those of the other category. Clustering can also indicate the presence of implicit terrorist networks.

Implicit terrorist networks have major theoretical and policy importance. Theoretically, they expand our notion of a terrorist network, connecting network-actors and behavior more closely together. Empirically, they allow us to measure networks through observable visible phenomena rather than employing questionable intelligence. There are also four critical counter-terrorism policy impacts.

4.3 Counter-Terrorism Policy

It is said that it takes a network to take down a network [18]. An addendum is that it takes an understanding of implicit networks to target and destroy terrorism networks. Terrorist groups often pick on specific locations that are referred to as hot spots, which tend to occur in clusters. Counter-terrorism resources should be focused on areas most likely to be targeted, such as state capitals, national capitals and other symbolic cities or areas. Extremist groups act in ways that are not independent, creating implicit networks. Counter-terrorists need to understand the interaction of groups and contexts [19] and recognize that violence does not inoculate an area but indicates the opposite - a high likelihood of more violence in the same area.

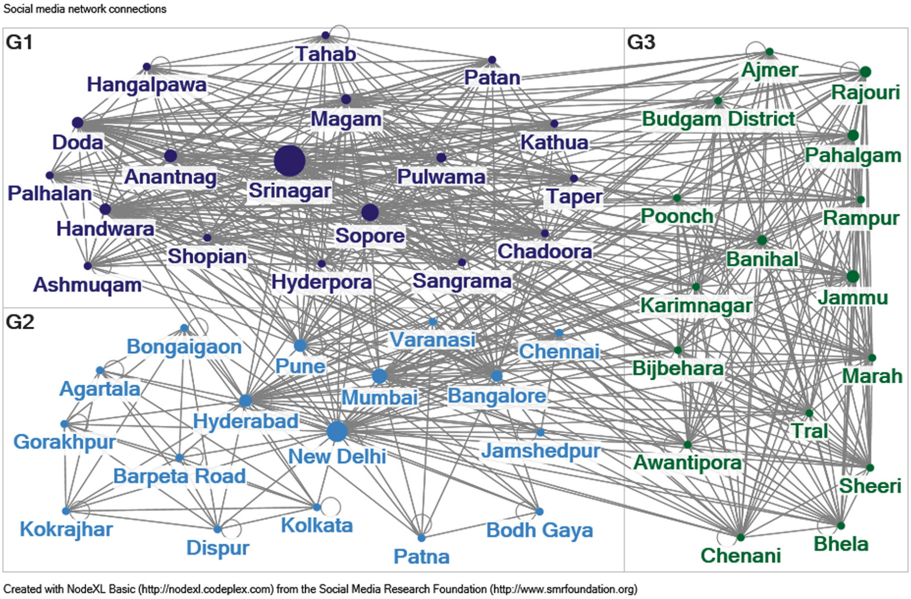


Fig. 2. Subgroups of locations (node size corresponds to the number of attacks experienced)

Subgroups in unexpected locations may indicate an external presence or additional support. For example, Karimnagar and Ajmer, shown in Fig. 2 are two cities geographically located outside of Jammu & Kashmir (J & K), but placed together as part of the two subgroups G1 and G3 comprising locations based in J & K. Sub-groups might actually be variants of the same group. As found in our network community analysis (Fig. 2), there is a strong likelihood that the two clusters G1 and G3 are "in-effect" the same, and thus represent an "implicit terrorist group." Counter-terrorist practices may benefit from effectively treating those in clusters as a single group, recognizing implicit networks.

5 Conclusion

Social network analysis (SNA) represents a tool for understanding terrorist networks. Krebs [20] was among the first to use SNA to analyze 9/11. Scholars argue that networked structures facilitate terrorists' use of violence [18] and networks influence terrorist acts' impact [21]. Several studies use SNA to examine network links within [22] or between extremist groups [1]. We extend extant research by treating the links between terrorist groups and the geographic location of their attacks as two-mode networks. To the best of our knowledge, this is the first work applying this approach to the study of terrorism and implicit terrorist networks.

Acknowledgment. This research was sponsored in part by the U.S. Army Research Laboratory and the U.K. Ministry of Defense under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defense or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copy-right notation hereon.

References

1. Asal, V.H., Park, H.H., Rethemeyer, R.K., Ackerman, G.: With friends like these... Why terrorist organizations ally. Int. Public Manag. J. **19**(1), 1–30 (2016)
2. Felmlee, D.H.: Interaction in social networks. In: Delamater, J. (ed.) Handbook of Social Psychology. Handbooks of Sociology and Social Research. Springer, Boston (2006). https://doi.org/10.1007/0-387-36921-X_16
3. Felmlee, D.H., Lungeanu, A., Kreager, D.: Online Dating Preferences: Two-Mode versus One-Mode ERGM Network Analysis. Presented at the meetings of the Population Association of America, Chicago, IL (2017)
4. Berrebi, C., Lakdawalla, D.: How does terrorism risk vary across space and time? An analysis based on the Israeli experience. Def. Peace Econ. **18**(2), 113–131 (2007)
5. McCauley, C.: Jujitsu politics: Terrorism and response to terrorism. In: Collateral Damage: The Psychological Consequences of America's War on Terrorism. Greenwood Publishing Group (2006)
6. Crenshaw, M.: The logic of terrorism. Terror. Perspect. **24**, 24–33 (2007)
7. Drake, C.J.: The role of ideology in terrorists' target selection. Terror. Polit. Violence **10**(2), 53–85 (1998)
8. National Consortium for the Study of Terrorism and Responses to Terrorism (START) Homepage. https://www.start.umd.edu/gtd. Accessed 26 Apr 2018
9. South Asian Terrorism Portal. India: Terrorist, Insurgent, and Extremist Groups. http://www.satp.org/satporgtp/countries/india/terroristoutfits/index.html. Accessed 26 Apr 2018
10. Basu, A.: Social network analysis of terrorist organizations in India. In: North American Association for Computational Social and Organizational Science (NAACSOS) Conference, pp. 26–28 (2005)
11. Saxena, S., Santhanam, K., Basu, A.: Application of social network analysis (SNA) to terrorist networks in Jammu & Kashmir. Strateg. Anal. **28**(1), 84–101 (2004)
12. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, 1st edn. Cambridge University Press, Cambridge (1994)
13. Breiger, R.L.: The duality of persons and groups. Soc. Forces **53**(2), 181–190 (1974)
14. McPherson, J.M.: Hypernetwork sampling: Duality and differentiation among voluntary organizations. Soc. Netw. **3**(4), 225–249 (1982)
15. Freeman, L.C.: Centrality in social networks: Conceptual clarification. Soc. Netw. **1**, 215–239 (1979)
16. Newman, M.E.: The mathematics of networks. New Palgrave Encycl. Econ. **2**, 1–12 (2008)

17. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. **70**(6), 1–6 (2004)
18. Arquilla, J., Ronfeldt, D.: Networks and Netwars: The Future of Terror, Crime, and Militancy. Rand Corporation, Santa Monica (2001)
19. Gartner, S.S.: An Introduction to Net Assessment 2.0: Special Issue on the Net Assessment of Violent Non-State Actors. CTX: Conflict Terrorism Exchange (2015)
20. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3), 43–52 (2002)
21. Gartner, S.S.: Ties to the dead: Connections to Iraq War and 9/11 casualties and disapproval of the president. Am. Sociol. Rev. **73**(4), 690–695 (2008)
22. Everton, S.F.: Disrupting Dark Networks. Cambridge University Press, Cambridge (2012)

Complex Networks for Terrorist Target Prediction

Gian Maria Campedelli[1,2](✉), Iain Cruickshank[2], and Kathleen M. Carley[2]

[1] Università Cattolica Del Sacro Cuore, Milan, Italy
gianmaria.campedelli@unicatt.it
[2] Institute for Software Research, Carnegie Mellon University, Pittsburgh, USA
icruicks@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract. Developments in statistics and computer science have influenced research on many social problems. This process also applies to the study of terrorism. In this context, network analysis is one of the most popular mathematical methods for analyzing terrorist organizations and dynamics. Nonetheless, few studies have applied network science to the analysis of terrorist events. Therefore, in this work we first introduce a novel method to analyze the heterogeneous dynamics of terrorist attacks through the creation of a dynamic meta-network of terror for the period 1997–2016. Second, we use our terrorist meta-network to test the power of Network-based Inference algorithm in predicting terrorist targets. Results are promising and show how this algorithm reaches high levels of precision, accuracy, and recall and indicate that network outcomes can be used in broader machine learning models.

Keywords: Complex networks · Terrorism · Prediction · Machine learning

1 Background

Terrorism represents one of the most challenging threats to the global community. Its heterogeneous and complex nature has fostered increasing interest in the scientific community, especially to inform policy-oriented measures.

In recent years the availability of large datasets, the diffusion of powerful machines and the advances in mathematical modelling techniques have contributed to the development of several computational approaches to the study of terrorism [1]. Among these approaches, network analysis has been employed in several specific subdomains. Indeed, a classic way to exploit network analysis is to understand and highlight internal dynamics and roles within terrorist groups [2, 3]. Following this analysis, scholars have tested and simulated the strength and resilience of terrorist networks, including dynamic and geographic processes [4, 5]. Another recent stream of research focuses on detecting terrorist and radical behaviors on social media platforms, such as Twitter [6]. Finally, researchers are trying to use event data to reconstruct multi-mode or multiplex networks in order to predict future terrorist attacks, locations and tactics [7, 8]. Our work lies in this latter field.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 348–353, 2018.
https://doi.org/10.1007/978-3-319-93372-6_38

2 Research Problem and Aims

The investigation of the evolutionary dynamics of terrorism can help in predicting where, when, and how next attacks will occur. Complex networks can play a relevant role in detecting patterns and underlying schemes. Specifically, researchers have tried to identify and analyze trends and motivations behind terrorist target selection, considering targets as relevant features for understanding possible consequences and predicting future attacks [9]. Within this line of research, our contribution seeks to (a) introduce a framework for analyzing terrorism through complex networks and (b) predict future targets using an algorithm from recent recommendation engine work.

3 Data and Methods

This work relied on data provided by the Global Terrorism Database (henceforth GTD) [10]. For our analysis, we used data of terrorist attacks that occurred in Western Europe and North America from 1997 to 2016[1] (Table 1). The choice of using only Western Europe and North America is motivated by the interest of focusing on wide areas which have been plagued by different types of terrorism, in terms of ideology, damages, targets and religious motivations. Regarding the time-span, relational information on events occurred prior to 1997 was not adequately complete and we therefore decided to rely only on robustly coded events.[2]

Table 1. Descriptive statistics of GTD data for Western Europe and North America (1997–2016)

	N	Avg.[a]	St. Dev.	Median	Max (year)	Min (year)
Attacks	2,032	101.6	45.37	94.5	219 (1997)	34 (2011)
Actors	284	29.5	11.12	30.5	50 (2010)	16 (2005, 2011)
Targets	22	13.4	2.60	13.5	46 (1997, 2001)	20 (2011)

[a]Average is referred to the mean number of plotted attacks, active actors and chosen targets in each year.

Employing these data, we built a meta-network using attacks, targets and perpetrators and actors for each time period. Given a temporal vector T = $(t_1, \ldots, t_n)$ where to each temporal element are associated $(k+m)$ networks that take the form of mathematical graphs G and each can be either monopartite in the form $G = (N, E)$ or

[1] Western Europe region data from 1997–2016 include Austria, Belgium, Cyprus, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Malta, Netherlands, Norway, Portugal, Spain, Sweden, Switzerland and United Kingdom. North America region data for the same period include Canada, Mexico and United States of America.

[2] We used the general target type information available in the dataset to prevent problems of over-specification and noise in the data, considering that a more general description reduces the risk of coding error, consequently preserving results reliability.

bipartite in the form $G = (U, N, E)$ where U and N are different sets of nodes and E are the edges that connect these nodes, we define a meta-network for the period t_i as:

$$\mathrm{M}_{t_i} = \bigcup \left[\bigcup_{j=1}^{k} G_j(N_j, E_j), \bigcup_{l=1}^{m} G_l(U_l, N_l, E_l) \right] \tag{1}$$

Equation (1) is simply the union of all networks associated to that specific period, both mono- and bipartite. Thus, for our work, we built twenty meta-networks in order to introduce a framework for analyzing the evolutionary dynamics of terror in Western Europe and North America from 1997 to 2016. Each meta-network was composed by 7 mono- and 3 bipartite networks. (Table 2).

Table 2. Description of meta-network composition

Network	Type	Edge type
Event × Event (combined event)	Monopartite	Binary
Event × Event (shared target)	Monopartite	Weighted
Event × Event (shared actor)	Monopartite	Weighted
Event × Actor	Bipartite	Binary
Event × Target	Bipartite	Binary
Actor × Actor (connected event)	Monopartite	Weighted
Actor × Actor (shared target)	Monopartite	Weighted
Actor × Target	Bipartite	Binary
Target × Target (shared actor)	Monopartite	Weighted
Target × Target (shared event)	Monopartite	Weighted

As a first exploratory experiment, we use the actor by target matrix to predict future targets using the Network-based Inference algorithm [11]. Particularly, this technique utilizes a resource sharing scheme for bipartite network projections. In our example, we consider the bipartite graph of Actors and Targets $G(A, T, E)$ where there are A is the set of actors and T of targets, and E represent the attacks that actors perform on targets. From this graph, we compute the resources moving from Actors to Targets by

$$w_{ij} = \frac{1}{k(a_j)} \sum_{l=1}^{m} \frac{g_{il} g_{jl}}{k(y_l)} \tag{2}$$

In Eq. (2), $k(a_j)$ is the degree of Actor i, $k(y_t)$ is the degree of Target l, and w_{ij} is weight for the n by n projected adjacency matrix. Given this weighted adjacency matrix and a current number of times target j is attacked by actors, $f(t_j)$, we compute a projected state of attacks by the n attackers in the next time step as

$$f^{'}(t_j) = \sum_{l=1}^{n} w_{jl} f(t_j) \tag{3}$$

So, using the previous history of targets and actors, we can compute a resource projection of the bipartite graph to a monopartite graph, which can then be used to predict likely future targets.

4 Target Prediction: Results

To analyze possible future targets, given attacks on previous targets, we used the bipartite graphs of actor by target. We choose two methods of doing the bipartite projection for Network-based Inference. First, we constructed a bipartite graph using all of the terrorist actors and the locations they attacked for the years from 1997 to 2015, where each terrorist actor was a different entity within each year. This produced a bipartite graph with integer-valued edge weights, where each entry corresponded to the number of times a terrorist actor attacked a particular target within a given year. So, as an example, Anarchists in 1997 is one actor and Anarchists in 1998 represents a different actor. Second, we constructed a bipartite graph using all the terrorist actors and the locations they attacked for the years from 1997 to 2015, where each terrorist actor only has one entry and their targets are summed across all years. Once again, this will be a bipartite graph with integer-valued edge weights. We then used these bipartite graphs to construct weighted, monopartite, projection graphs to predict likely targets for terrorist attacks in 2016 using the Network-based Inference (NBI) technique. It should be noted that the original NBI technique was used on binary graphs, and that we used a weighting function as opposed to a binary value in our projections. Additionally, as a baseline, we also used the targets from 2015 as those that would be attacked in 2016 (Table 3).

Table 3. Results of target selection predictive models using Network-based Inference (NBI) algorithm

Model	Accuracy	Precision	Recall
Baseline (repeat 2015 targets)	0.62	0.769	0.769
Network-based Inference predicted targets (each actor unique in each year)	0.762	1.0	0.62
Network-based inference predicted targets (only one actor across all years)	0.68	0.77	0.65
Network-based Inference predicted targets for only those actors existing in 2015	0.81	1.0	0.67

Using the NBI on a weighted bipartite graph of actors by targets improves the ability to predict targets in the next time frame, 2016, by about 10%. Interestingly, in 2016, several new actors had attacks on targets (such as anti-Trump actors, and Dissident Republican actors) that did not do any attacks in 2015. So, if we only take into account the attacks of those groups that existed in 2015, the accuracy of the method further improves to 81%[3]. Additionally, constructing the weighted projection where each actor only has one entry across all of the years shows less accuracy than constructing a weighted projection where each actor in each year gets an entry and only the 2015 actors are used to predict 2016 targets. Looking at the data, this result is likely a consequence of government-based targets being much more popular in the years before 2015, so that if all years are taken into account for 2016 attacks, NBI tends to over-estimate government-based targets.

This phenomenon of shifts in popular targets over the years makes up much of the loss in accuracy of NBI. For instance, much of the loss in accuracy comes from an unanticipated spike in attacks on religious-based targets in 2016. From 1997 to 2015, there were relatively few religiously-based targets that were attacked by terrorist actors in 2015, so the NBI technique struggles to account for a surge of religiously-based targets in 2016. That being said, some targets are habitually popular across years, such as vehicles and police based targets, and NBI is very accurate in predicting those attacks.

5 Discussion and Future Work

This work has introduced a novel method for analyzing terrorism complexity, exploiting the richness of open access data on terrorist events, and tested an algorithm developed in recommendation engine field for predicting future terrorist targets. Future work should expand the geographical range to other world regions to test what are the global connections of terrorism and to understand if vaster contexts highlight particular features. While enlarging the geographic scope may help on one side, on the other side it will be interesting to reduce the length of time slices. This may be a promising way to integrate outcomes in larger research designs where network analysis can enrich the processes of signal and anomalous pattern detection, exploiting existing time-series techniques for sparse data.

A different class of problems is related to the interest of finding clusters among terrorist groups. Maintaining wide time slices (e.g. yearly periods) would permit to collect enough data for understanding stable similarity relations among actors, events, and selected targets. Thus, further analyses should test algorithms designed for multimode networks in order to verify their performance and, eventually, develop tailored ones for this specific purpose.

This work feeds a rising process in terrorism studies. Research on terrorism is increasing and advances in multiple fields are fostering promising expectations, nevertheless our intuition is that scholars and analysts will only make relevant progresses if the scientific dialogue will integrate both the views of terrorist experts and computer scientists and mathematicians. As an example, the development of new network algorithms will require deep domain knowledge of the phenomenon to create algorithmic structures that are meaningful also in practical ways.

[3] It is expectable that taking into account attacks of those groups that existed at (t−1) improves accuracy of prediction. Though not empirically confirmed in this work, future work will test this hypothesis on all pairs of years included in the analysis.

Acknowledgments. This work was supported in part by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative Award N00014-17-1-2675 and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR or the U.S. government.

References

1. Subrahmanian, V.S. (ed.): Handbook of Computational Approaches to Counterterrorism. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-5311-6
2. Koschade, S.: A social network analysis of Jemaah Islamiyah: the applications to counterterrorism and intelligence. Stud. Confl. Terror. **29**, 559–575 (2006). https://doi.org/10.1080/10576100600798418
3. Belli, R., Freilich, J.D., Chermak, S.M., Boyd, K.A.: Exploring the crime-terror nexus in the United States: a social network analysis of a Hezbollah network involved in trade diversion. Dyn. Asymmetric Confl. **8**, 263–281 (2015). https://doi.org/10.1080/17467586.2015.1104420
4. Moon, I.-C., Carley, K.M.: Modeling and simulating terrorist networks in social and geospatial dimensions. IEEE Intell. Syst. **22**, 40–49 (2007). https://doi.org/10.1109/MIS.2007.4338493
5. Medina, R., Hepner, G.: Geospatial analysis of dynamic terrorist networks. In: Karawan, I. A., McCormack, W., Reynolds, S.E. (eds.) Values and Violence, pp. 151–167. Springer, Dordrecht (2009). https://doi.org/10.1007/978-1-4020-8660-1_10
6. Benigni, M.C., Joseph, K., Carley, K.M.: Online extremism and the communities that sustain it: detecting the ISIS supporting community on Twitter. PLOS ONE **12**, e0181405 (2017). https://doi.org/10.1371/journal.pone.0181405
7. Desmarais, B.A., Cranmer, S.J.: Forecasting the locational dynamics of transnational terrorism: a network analytic approach. Secur. Inform. **2**, 8 (2013). https://doi.org/10.1186/2190-8532-2-8
8. Tutun, S., Khasawneh, M.T., Zhuang, J.: New framework that uses patterns and relations to understand terrorist behaviors. Expert Syst. Appl. **78**, 358–375 (2017). https://doi.org/10.1016/j.eswa.2017.02.029
9. Brandt, P.T., Sandler, T.: What do transnational terrorists target? Has it changed? Are we safer? J. Confl. Resolut. **54**, 214–236 (2010). https://doi.org/10.1177/0022002709355437
10. National Consortium for the Study of Terrorism and Responses to Terrorism: Global Terrorism Database (Data file) (2016). https://www.start.umd.edu/gtd
11. Zhou, T., Ren, J., Medo, M., Zhang, Y.-C.: Bipartite network projection and personal recommendation. Phys. Rev. E **76** (2007). https://doi.org/10.1103/physreve.76.046115

Cybersecurity

Searching for Unknown Unknowns: Unsupervised Bot Detection to Defeat an Adaptive Adversary

Peter A. Chew(✉)

Galisteo Consulting Group, Inc., 4004 Carlisle Blvd NE Suite H, Albuquerque, NM 87107, USA
pachew@galisteoconsulting.com

Abstract. The use of bots in social media has become highly topical with the allegations that Russia actively uses such automated accounts to lend artificial amplification to desired messages, and as subterfuge in Western public debate. A key question analysts must answer, therefore, is which accounts are likely to be bots and how to separate these from 'real people'. In practice, current approaches to answering this question are time-consuming and heuristics-based, and even in many academic approaches, heuristics still play a key role. The reliance on heuristics, however, means that adversaries can easily adapt as soon as they find out what the heuristics are. In this paper, we propose a new, more robust, unsupervised approach. While still using as input the same basic elements that inform heuristics-based approaches – account names, timing and frequency of posts, words used, etc. – the approach abstracts away from specific heuristics and potentially allows 'unknown unknowns' to be searched for – patterns that would be unlikely to be human-generated. By virtue of this, our approach has the potential to reduce the time and costs for those seeking to make sense of the information space, while simultaneously making it harder for adversaries to 'game the system'. We demonstrate examples of the sort of output that our unsupervised approach to bot detection can yield.

Keywords: Bots · Social media · Influence operations

1 Geopolitical Background

Throughout 2017, the subject of Russian interference in Western democracies has been a hot topic in the West. For many among Western audiences the topic has come only recently to the fore – starting perhaps with the January 2017 report by the US Director of National Intelligence. But in fact, for many in Eastern Europe, geographically closer to Russia, this has been a prominent point of discussion since 2014 [2], when Russia annexed Crimea and pro-Russian forces intervened in Ukraine following the exile of former Ukrainian President Viktor Yanukovich, seen by many as pro-Russian. Discussion by Ukrainians that we have personally observed (albeit anecdotally) via our own participation in Russian-language social media, where suspicion of 'trolls' in political discussion has been rife since 2014, only tends to confirm that the Russian government *is* doing what Russian general Valery Gerasimov said in 2013 Russia *should* do:

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 357–366, 2018.
https://doi.org/10.1007/978-3-319-93372-6_39

> *Information conflict opens up significant opportunities for reducing the adversary's military potential in an asymmetric fashion. In North Africa we witnessed how technology was used to influence State structures and populations via information networks. It's imperative that we perfect methods of working in the information space, including in the defense of our own assets.* [3]

Russia's 'Internet Research Agency' (or AII, to use the Russian abbreviation), a 'troll farm' and propaganda factory, has been documented [4] as playing a role which, again, would be consistent with what Gerasimov called for. In fact, in a sense, such activities were not really new even in 2013; they are simply a reprise of the sort of propaganda efforts the Soviet Union engaged in during the Cold War, simply adapted to the age of the Internet.

One notable aspect of this 'information conflict' is the use of bots – automated accounts – which expand the reach of propaganda. Bots may exist for worthy and useful purposes – such as tweeting out weather conditions from a specified location at set intervals. However, bots may also be programmed, for example, to retweet any posts that contain certain combinations of keywords. It is not hard to imagine how this would make it easy for outfits such as the AII to artificially extend the reach of certain messaging, for example messaging that casts certain individuals or countries in either a positive or negative light. In fact, testimony by Facebook, Google and Twitter before the U.S. Senate on October 31, 2017 appears to confirm that this is exactly the kind of tactic that Russian-controlled bots *did* engage in during the 2016 U.S. Presidential election [5]. By programming bots to retweet certain content more than other content, it is possible for hostile foreign actors not only to spread destabilizing messaging at little cost, but also to purport that this messaging has more support than is actually the case, and, furthermore, fraudulently to make it appear that this support is domestic rather than foreign [6]. This would be consistent with how Gerasimov puts it:

> *Asymmetric actions – those which allow one to neutralize the enemy's advantage in armed conflict – have become widespread. These include the use of special forces and internal opposition forces [in the enemy's polity] to open up a front across the adversary state's whole territory. They also include influence operations, the forms and methods of which are constantly being perfected.* [3]

In light of all this, it is sobering to note that as many as 15% of active Twitter accounts may be bots [7].

As awareness has spread in the West of the threat, more thought has been devoted to what counter-measures that can be applied. These range from technical (for example, Twitter's deactivation of accounts identified as Russian bots [8]) to exposés of the techniques used by 'bot-herders', based on research and practitioners' experience [9].

However, this is an 'arms race', in the sense that the adversary is adaptive. When Twitter deactivates certain accounts, new ones can easily be created by an adversary. As soon as 'bot-spotters' publicize a heuristic like 'anything above 72 tweets a day is suspicious', 'bot-herders' can easily reprogram their bots to output 71 tweets a day, and create more bots to compensate for the loss in output – something that we have heard (in personal communication) does in fact happen. What is more, even just applying heuristics to comb through accounts is a time-consuming process, let alone the time required to develop reliable heuristics.

In view of all this, we strongly believe that any technological tool which is to remain useful in the long run must also be adaptive; it must not rely on previously observed patterns or heuristics, which may have to change as the adversary adapts. Rather, such a tool must abstract away as much as possible from the specific, and look to hard-to-fake 'signal' that differentiates the output of humans and machines, whether through the text they generate or through other aspects of their activity.

An approach like this will have a twofold benefit: broadly, it will tend to help those who seek to understand the 'information space' in an unbiased way, and at the same time it will make it harder for those who seek to obfuscate, fraudulently inflate the importance of certain messaging, and so on. It helps the former by reducing the amount of time and effort that must be put into developing and applying heuristics, because the approach becomes more data-driven than heuristic-driven; one is simply searching for things that 'stand out' in whatever way. And it hinders the latter, because it is very hard to be consistent in deception, especially when the indices by which deception is being measured are not overtly stated, and are in fact driven by the very data itself.

In the remainder of this article, we set out a method that we believe would fulfill these criteria. The key value of what we propose is in its ability to draw attention to patterns that would otherwise (owing to the scale of the data) be latent. This being the case, we think a demonstration of the sort of output such an approach can generate is most appropriate, rather than an empirical comparison against other methods, which would be almost impossible in any case given the dynamic nature of the problem. The structure of this paper is therefore as follows: in Sect. 2 we discuss alternative current approaches to bot identification and why we see them as theoretically unsuited to the task at hand. We also introduce unsupervised analytics and discuss how it has been used to great effect, including commercially. In Sect. 3, we show how unsupervised data analytics can be applied to Twitter data to reveal latent patterns distinguishing various types of account, including bots and humans. In Sect. 4, we demonstrate the results of applying this approach to a dataset of about 115,000 Russian and English language Twitter posts from October and November 2016, spanning the time of the US general election. Several non-obvious findings come to light, and possible botnets (coordinated groups of bots) are highlighted. In Sect. 5, we conclude on our research.

2 Typology of Approaches and Their Use in Practice

To set the context for discussion of different data-analytics approaches, we think it is helpful to start with a concrete discussion of how respected and authoritative practitioners of 'bot identification' actually approach the task in practice. While some (much?) of this work is likely hidden from general view because of government secrecy considerations, a clear and highly accessible account is published by the Atlantic Council's Digital Forensic Research Lab (DFRL) in [9]. This lists 12 heuristics that tend to set bots apart from humans in Twitter, and it is worth recapitulating them here. Nimmo suggests that the more of these that are true, the more likely the account is to be a bot.

1. More than 72 posts a day (more than one every 20 min).
2. Anonymity (no name, no profile information).

3. Amplification (few or no original posts, high rate of retweets).
4. Low posts, high results (few posts with high rate of likes/retweets indicates the account may be part of a botnet).
5. Content mirrors that of other accounts.
6. No avatar image.
7. Avatar image is stolen (e.g. from a celebrity).
8. Account name is a scramble of alphanumeric characters and/or does not match screen name.
9. Single account posts in multiple languages.
10. Commercial content (advertising mixed with occasional political tweets).
11. Bots often use the same URL shorteners (e.g. ift.tt).
12. Patterns of retweets and likes very similar to, or match, other accounts.

While applying these heuristics can be time-consuming, even when one can use a computer to calculate the counts needed, it is not hard to see how a **rules-based** program could be devised to implement most or all of these checks. The problem with this, as already alluded to, is that it is not hard to see how an adversary can and will adapt to defeat the checks, especially as the 12 heuristics are public knowledge.

Another way to automate bot detection, exemplified in [10] (developed in 2011), relies on **supervised learning**. It should be clear that the 12 heuristics above, at the most fundamental level, all reduce to building-blocks such as the specific words used in the posts, likes, retweets. In machine learning, these building-blocks can be treated as features. In [10], around 1,000 such features are used to train an algorithm, based on previously identified and labeled bots, to estimate an overall probability that a given account is a bot. As noted just three years after the algorithm was developed [11], the algorithm is trained on bots identified in 2011, so, as was pointed out, 'it's quite possible that there are now more advanced bots that are less easy to detect'. This is to say, in another way, that an adaptive adversary is perhaps as much a problem for supervised learning techniques as for rules-based implementations of bot detection techniques. The problem arises, in essence, because bots may evolve over time; the bots of today may not look like the bots of yesterday.

A third, more abstract, category of approach follows the principles of **unsupervised learning**. Approaches of this type have also been previously applied to identifying Twitter bots [12]. The key here is leveraging the insight that bots, unlike humans, tend to have account activities that are highly correlated with one another: this is part of what inescapably makes a bot, a bot. Note that the idea of looking for 'correlations' simply states differently the same insight apparent from many of the 12 heuristics above: bots tend to repost or copy content rather than create original content of their own. However, to state the concept as looking for things that are 'correlated' or 'similar' to one another is to state it in terms that align with the forte of unsupervised machine learning, which is to allow patterns – groups of things that are similar, or outliers that are somehow different from everything else – simply to emerge from data. This is the same basic principle that underlies search engines. It does not matter what the 'features' are or whether they evolve over time (clearly, the internet itself is evolving at a fast rate, yet a good search engine continues to function), what matters is that the basic elements of the search query in some sense match the content being searched for; the search engine does not even need to care what the specific elements are. The promise of an unsupervised approach, then, is that (again, perhaps, like a good search engine) it can open up content to those who want to know the 'truth' – to find what is out there – while making it harder to hide content in plain sight. Unlike heuristics-driven or supervised methods, unsupervised learning by its nature will adapt along with any adversary, skewing the playing-field towards those who seek to understand rather than to obfuscate.

3 Novel Application of Unsupervised Analytics to Bot-Spotting

3.1 SVD and Its Application to Text, LSA

In this section we propose a specific application of topic analysis (via Singular Value Decomposition or SVD [13], a signal processing technique), modified not only to help an analyst detect individual bots, but *at the same time* to understand how bots group into networks (botnets) and what topics social bots may coalesce around. In this respect, our approach is qualitatively novel. Our approach is unsupervised and will therefore reflect emergent patterns rather than relying on heuristics or previously observed patterns. By its nature, therefore, it can adapt along with any adversary countermeasures. Our technique takes a top-down approach to understanding the 'forest' before the 'trees', meaning that it automatically draws attention first to the most significant botnets (where significance is determined by how prolific the botnet might be, or how many bots it comprises). Bots by their nature have to be in the open (otherwise they would not influence users) and to amplify a message they also have in some way to duplicate content, whether through retweets, reposts, or likes. We use this basic insight about social bots in applying an algorithm to make their activity more apparent to a human without having to comb through thousands of individual posts, or laboriously apply heuristics. The purpose is to direct SVD towards the basic elements of Twitter posts that inevitably have to be used to influence humans and amplify a message to them.

While we think it would be possible ultimately to develop a method that would subsume all 12 of the heuristics listed in Sect. 2, here for simplicity we focus on those heuristics that lend themselves easily to text analysis. Text, after all, is the mainstay of Twitter posts. Of these 12 heuristics, at least 7 should be covered by the approach described here (1, 3, 5, 9, 10, 11, 12), and adapting the method to incorporate the other 5 could be done with varying degrees of difficulty.

The application of SVD to text has been well-known for some time and is commonly known as Latent Semantic Analysis (LSA) [14]. In LSA, a sparse matrix can be constructed such that each column represents a 'document' (in Twitter, this might equate to an individual post) and each row, a term (a word in the vocabulary used in any of the posts). Each cell (i,j) can then record the number of times term i appears in document j. In practice, a weighting scheme [15] is usually employed, an effect of which is to 'dampen' the weight of high-frequency but non-distinctive terms such as 'the', and to increase the weight of distinctive words. The matrix is sparse because each document tends to contain only a small fraction of the vocabulary. SVD is used to factorize the matrix, which we can call *X*, into three dense matrices according to the relationship

$X = USV^T$, where U is an orthonormal matrix of left singular vectors, S is a diagonal matrix of singular values, and V is an orthonormal matrix of right singular vectors [13].

Typically for LSA, a truncated SVD is computed such that equality above no longer holds and that the best rank-R least-squares approximation to matrix X is formed by keeping the R largest singular values in S and discarding the rest. This also means that the first R vectors of U and V are retained, where R indicates the number of concept dimensions in LSA. Each column vector in U maps the terms to a single arbitrary concept, such that terms which are semantically related (as determined by patterns of co-occurrence) will tend to be grouped together with large values in columns of U.

Translating this mathematical explanation of LSA into text analysis terms, we can say that U groups terms (words) into natural topics, while V does the same for documents. S, on the other hand, encodes the order of prominence of each of the topics in the source data. In this sense, the topics just 'emerge' from the data based on the co-occurrence of terms within documents; no information is needed a priori.

Furthermore, the output of LSA can be used to estimate the similarity of any term to any other term, any document to any other document, or any term to any document. Because U and V are orthonormal matrices, this is done simply by calculating the cosine between the respective term or document vectors from U and V.

3.2 Reformulating LSA for Bot Detection

In detecting Twitter bots, and the patterns of similarities between bot activity that imply networks of bots (botnets), we are interested not so much in individual Twitter posts as we are in whole accounts (most of which post multiple times) and the features that define those accounts. In this, our input is no different from that of the rules-based and supervised approaches already described in Sect. 2.

However, the input is different from that of 'plain-vanilla' LSA, where the input is a term-by-document matrix. Here, because our focus is on Twitter *accounts*, our starting-point is a term-by-*account* matrix, although the construction of this matrix and its weighting can proceed as if it were a term-by-document matrix. Entries in this matrix will now simply record the (weighted) number of times a 'term' is used by an account.

It is worth noting that to determine what qualifies as a 'term', we need a tokenizer. For this, we use some simple Regular Expressions code which splits the text at every 'nonword' character (punctuation, white space, etc.):

```
Shared Function Tokenize(ByVal s As String) As
IEnumerable

  Return Regex.Split(s, "\W", RegexOptions.Compiled)

End Function
```

This approach works for virtually all languages, including those in other scripts such as Russian and Arabic. It is also worth noting that URLs (mentioned above in the context of URL shorteners bots tend to employ) get chunked into their constituent parts by this approach – so, http://ift.tt yields the ‘words’ http, ift and tt. This is useful, because patterns of similarity can then emerge between different accounts that use the ift.tt shortener, even if the specific shortened URLs are different.

Moving from a term-by-Twitter-post to a term-by-account matrix, however, loses one key piece of information which is perhaps the most important in bot identification: the frequency of posts. Nowhere in the term-by-account matrix do we see how many times an account posted, because all the posts are now combined together. It is relatively easy to compensate for this, however, and we can do so by adding in a ‘vocabulary’ item for each period of time (for example, date) in which an account posted. For example, if an account posted 100 times on 1/1/2017, the ‘term’ 1/1/2017 is included and its unweighted frequency is 100. Weighting can then be applied in the normal way. Including this piece of information is key in allowing some interesting insights to emerge from the data (see Results below).

By constructing the matrix in this way and applying SVD, the ‘topics’ that emerge show patterns of words that co-occur frequently across accounts, and patterns of accounts that tend to use very similar combinations of ‘words’ (understood broadly to include elements like URLs). For bots that tend to mimic one another and other accounts, whose identities might be hidden to a casual Twitter user looking at the ‘trees’ rather than the ‘forest’, significant patterns will become quickly apparent. The technique potentially allows not just individual bots to be identified, but whole groups of bots en masse. We also have a way (through calculation of mutual account similarity, using the cosines between vectors from the V matrix as described above) to see which accounts behave very similarly to which other accounts. In short, SVD gives a single unified framework to explore the data and find patterns in many different ways.

4 Demonstration and Results

4.1 Demonstration

We applied our technique to a dataset of 115,169 Twitter posts created between October 13, 2016 and November 28, 2016 (spanning the period from about 3-4 weeks prior to the November 8 US general election, to about the same amount of time after it). These posts originated from 57,888 unique Twitter accounts. The posts were collected by sampling from the Twitter firehose.

As described in 3.2, ‘terms’ were included in the term-by-account matrix for the date of each post, allowing the frequency of posts to be indirectly captured for each account.

Because of how the data was collected, we do not have a ‘ground truth’ for this dataset as to which accounts are bots and which are not. We also did not know for sure in advance that any of the 57,888 accounts *were* bots. Note that is part of the purpose of this paper: to show that useful results can be derived using an unsupervised technique from a previously unseen dataset (as opposed to showing that X% of previously identified bots are identified by the current technique).

We ran a truncated SVD to yield 90 ‘concepts’, and indicate below some of the non-obvious insights that stood out to us.

4.2 Results

4.2.1 Top Topical Pattern

The top pattern to emerge is a group of accounts all posting on a Russia/Syria/naval topic. A large number of distinct accounts (at least 100) closely align with this pattern, and the identical post ‘RT @PrisonPlanet: Russia Is Deploying The Largest Naval Force Since The Cold War For Syria: NATO Diplomat. https://t.co/Rj3eo4PelB’ was repeated by every one these accounts. Following are some of the 100 accounts we identified that retweeted this, a large number of which seem to have ‘impersonal’ names:

> buttonlol, __hunts, 2ndfor1st, 411mssip, 4n0rdc, 61rinaldi, 8reallyisenough, aaronmiller7064, alfolart, angieevanhorn, antjgrant, avivaldi1, becerra_erika, betapatersi, bhattnaturally1, bhopalibhopali, billyjackyl, blckgirlsmatter, bouncermaniac, brandonbast, bumpsethurdle, bundrenstanley, bydeeps, cajuzin, mazurekpaul, mia_liner, motherlandhome, murphy13272, mustermann2557, native4trump

Very quickly, therefore, we have identified a large group that has retweeted the same post and the topic of that post, without having to know beforehand what topic it was, let alone anything about the dataset.

4.2.2 Second Topical Pattern

The top accounts most representative of the second ‘topic’ each posted just two posts, and what is interesting is that each of these accounts posted the *same* identical two posts. Furthermore, the two posts are both on the subject of Gaddafi, and if that were not enough, there is a strong correspondence between the *dates* on which those two posts were posted. For example, globalspectator, mido1974, nasush_ (and others) posted according to the following pattern:

	RT @crimesofbrits: On this day in 2011, Muammar al-Gaddafi was lynched by Brit/French/US backed death squads. NATO destruction of Liby…	RT @mehdirhasan: Did NATO play a role in rape &amp; murder of Colonel Gaddafi, 5 years ago this week? On @ajupfront I ask NATO's ex-boss:\nhttps…
globalspectator	10/20/2016 8:44:27 AM	10/21/2016 2:46:05 AM
mido1974	10/20/2016 10:44:14 AM	10/21/2016 7:41:51 AM
nasush_	10/20/2016 11:02:59 AM	10/21/2016 7:06:07 PM

It seems implausible that this could have occurred by chance as a result of human activity. This topical pattern, therefore, even more than the first, seems to reveal a botnet. Inspection of further results in the second topic appears to indicate that at least 15 distinct accounts are in this botnet (all displaying the same temporal/topical pattern).

4.3 Topical Pattern 44

A similar and even more striking pattern occurs in topic 44, this time among accounts posting in Russian. While just 4 accounts are shown, 5 others display a similar pattern.

Note that these accounts post almost identical tweets consistently within a couple of minutes of one another. Again, it seems almost impossible that this could happen unless all the accounts in question were bots, and part of a coordinated botnet.

	НАТО опасается использования «Адмирала Кузнецова» для атак на Алеппо[1]	Bloomberg: Путин заявил о недопустимости провокаций против НАТО[2]	НАТО отказалось обсуждать в Москве вопросы безопасности на Балтике[3]
molodost__bz	10/25/2016 11:20:00	10/28/2016 03:35:54	10/31/2016 18:53:22
_deti_zhdut_	10/25/2016 11:19:58	10/28/2016 03:35:49	10/31/2016 18:53:23
korabliks	10/25/2016 11:19:55	10/28/2016 03:35:42	10/31/2016 18:53:15
nina55055	10/25/2016 11:21:07	10/28/2016 03:36:06	10/31/2016 18:53:38

It should be noted that it made no difference to our unsupervised bot detection algorithm that these accounts posted in Russian versus English; as noted above, by using SVD/LSA, we are able to cast (and solve) the problem simply as one of looking for accounts where the 'signal' closely matches.

Also of interest in this case is that the posts were not quite identical, yet our approach still found the pattern. The consistent differences between the accounts are yet further strong indication that they are bots. Note that molodost__bz consistently prefixes each post with '#news', while _deti_zhdut_ postscripts each post with the character ˘. Is this an attempt to evade detection by deliberately making the posts different from one another? Obviously, if it was, with our approach to bot detection, it failed, because the signal was stronger than the purposely introduced 'noise'.

molodost__bz #news НАТО опасается использования «Адмирала Кузнецова» для атак на Алеппо #news Bloomberg: Путин заявил о недопустимости провокаций против НАТО
_deti_zhdut_ НАТО опасается использования «Адмирала Кузнецова» для атак на Алеппо ˘ Bloomberg: Путин заявил о недопустимости провокаций против НАТО ˘

We think this illustrates well why LSA is so well suited to this task: too much 'noise' of this sort, and it might well be possible to evade detection, but then such countermeasures by adversaries might end up being self-defeating: the output of the bots needs to have enough signal to be intelligible to humans. LSA is a well-established method for finding the signal in text, precisely what we need to do here.

5 Conclusion

In this paper we have presented a highly flexible, adaptive method for bot identification, and showed that it yields very promising results. The method works regardless of language and the results show that it is able to overcome 'noise' (whether purposely introduced to obfuscate, or otherwise) and show emergent patterns of coordinated messaging 'amplification' of the sort Russia is alleged to have used recently against

Western democracies. The method focuses in on the largest-scale patterns first, enabling an analyst to target his or her energies and time to best use. Given an adaptive adversary, whose track record shows a desire to manipulate information deceptively, we believe that this type of unsupervised approach is the best way to skew the playing-field to the advantage of those who seek to understand rather than to deceive.

References

1. ODNI: Assessing Russian activities and intentions in recent US elections (2017). https://www.dni.gov/files/documents/ICA_2017_01.pdf
2. Darczweska, J.: The anatomy of Russian information warfare: the Crimean operation, a case study. Ośrodek Studiów Wschodnich im. Marka Karpia (2014). https://www.osw.waw.pl/sites/default/files/the_anatomy_of_russian_information_warfare.pdf
3. Gerasimov, V.: Ценность науки в предвидении (The value of science in foresight). Военно-промышленный курьер. VPK News, 27 February 2013. http://www.vpk-news.ru/articles/14632. (quotations above are English translations from Russian by the current author)
4. Chen, A.: The Agency. New York Times, 2 June 2015. https://www.nytimes.com/2015/06/07/magazine/the-agency.html
5. Dwoskin, E., Shaban, H., Timberg, C.: Facebook, Google and Twitter testified on Capitol Hill. Here's what they said. Washington Post, 31 October 2017. https://www.washingtonpost.com/news/the-switch/wp/2017/10/31/facebook-google-and-twitter-are-set-to-testify-on-capitol-hill-heres-what-to-expect/
6. Nimmo, B.: #BotSpot: How Bot-Makers Decorate Bots. Atlantic Council, Digital Forensic Research Lab (2017). https://medium.com/dfrlab/botspot-how-bot-makers-decorate-bots-4d2ae35bdf26
7. Varol, O., Ferrara, E., Davis, C., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of International Conference on Web and Social Media (ICWSM). AAAI (2017)
8. Sharkov, D.: Full list of Russian Twitter bots banned in election meddling probe. Newsweek, 3 November 2017. http://www.newsweek.com/full-list-russian-twitter-bots-banned-election-meddling-probe-700703
9. Nimmo, B.: #BotSpot: Twelve ways to spot a bot. Atlantic Council, Digital Forensic Research Lab (2017). https://medium.com/dfrlab/botspot-twelve-ways-to-spot-a-bot-aedc7d9c110c
10. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. arXiv preprint 2014. arXiv:1407.5225
11. MIT Technology Review: How to spot a social bot on Twitter, 28 July 2014. https://www.technologyreview.com/s/529461/how-to-spot-a-social-bot-on-twitter/
12. Chavoshi, N., Hamooni, H., Mueen, A.: Identifying correlated bots in Twitter. In: Spiro, E., Ahn, Y.-Y. (eds.) SocInfo 2016. LNCS, vol. 10047, pp. 14–21. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-47874-6_2
13. Golub, G., Van Loan, C.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
14. Landauer, T., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Process. **25**, 259–284 (1998)
15. Dumais, S.: Improving the retrieval of information from external sources. Behav. Res. Methods Instrum. Comput. **23**, 229–236 (1991)

Using Random String Classification to Filter and Annotate Automated Accounts

David M. Beskow(✉) and Kathleen M. Carley

School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue,
Pittsburgh, PA 15213, USA
dbeskow@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract. Automated social media *bots* have existed almost as long as the social media platforms they inhabit. Their emergence has triggered numerous research efforts to develop increasingly sophisticated means to detect these accounts. These efforts have resulted in a *cat and mouse* cycle in which detection algorithms evolve trying to keep up with ever evolving *bots*. As part of this continued evolution, our research proposes using random string detection applied to user names to filter twitter streams for potential bot accounts and thereby generating annotated data.

1 Introduction

Automated social media accounts, often called "bots", are increasingly used on many social media sites. Ever since social media sites built Application Programming Interfaces (API's) that allow their platforms to integrate with other platforms and applications, various actors have developed computer routines that conduct a variety of automated tasks on the respective social media ecosystems. While some bots are designed for positive purposes [9], many others range from nuisance (i.e. a spam bot) to propaganda [14], suppression of dissent [20], and network infiltration/manipulation [1,7]. They have recently gained widespread notoriety due to their use in several major international events, including the British Referendum known as "Brexit" [10], the American 2016 Presidential Elections [2], the aftermath of the 2017 Charlottesville protests [8], and most recently with less publicized reporting regarding the conflict in Yemen [13].

As these bots have proliferated and their use is being discussed broadly in the media and political bodies, researchers have increasingly developed methods to detect these accounts. The same openness and ease of use of the social media API's that facilitates the creation and use of automated accounts also facilitates the collection of data used to detect them. As detection efforts proliferate, bot engineers change and adapt in order to survive and succeed in a dynamic environment. The requirement for higher accuracy in the midst of a changing *signal* motivates our efforts to improve not only the models that detect bots, but the labeled data that is used to train them.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 367–376, 2018.
https://doi.org/10.1007/978-3-319-93372-6_40

Our work makes two primary contributions to the literature. First, we propose a novel random string detection model that is specifically designed to detect 15 character randomly generated strings. When applied to the *screen name* field of Twitter data, this technique is able to easily filter accounts that are likely bot accounts. Second, by applying this filtering technique to a large sample collected from the Twitter Streaming API, we have produced a large and diverse annotated data set for use in training more robust specialized and general purpose bot detection models.

This paper begins with a brief description of the background of general bot detection, as well as past efforts perform random string detection. We will then describe the models and algorithms that we developed for random string classification, as well as methods that we used to evaluate them on the narrow tasks that they were created for. Finally, we describe how we've applied this algorithm to create a large and diverse annotated Twitter bot data set for use by the research community.

2 Related Work

2.1 Twitter Bot Detection

Although early work on classifying Twitter accounts dates back to as early as 2008 [11], the deliberate detection of automated accounts on the Twitter Platform began in earnest in 2010 when [3] conducted three-class classification (human, bot, cyborg) using an ensemble model. In 2011, a team from Texas A&M became the first to use *honey pots* to detect thousands of bots [12]. These *honey pots* used bots that generate nonsensical content, designed only to attract other bots. The Texas A&M bots attracted thousands of bots, and generated a labeled data set that has been used on many later research efforts. This *honey pot* method was repeated by others to create similar data sets in other parts of the world [19].

In 2014, Indiana University and the University of Southern California launched the *Bot or Not* online API service [4]. This used traditional classification models trained on the Texas A&M dataset to help users evaluate whether or not an account is a bot. *Bot or Not* leverages network, user, friend, temporal, content, and sentiment features with Random Forest classification.

In 2015 the Defense Advanced Research Projects Agency (DARPA) sponsored a Twitter bot detection competition that was titled "The Twitter Bot Challenge" [19]. This four week competition pitted four teams against each other as they sought to identify automated accounts that had infiltrated the informal Anti-Vaccine network on Twitter. Most teams in the competition tried to use previously collected data (mostly collected and tagged with *honey pots*) to train detection algorithms, and then leverage tweet semantics (sentiment, topic analysis, punctuation analysis, URL analysis), temporal features, profile features, and some network features to create a feature space for classification. All teams used various techniques to identify initial bots, and then used traditional classification models (SVM and others) to find the rest of the bots in the data set.

Most recently, the team from Indiana University have re-branded *Bot-or-Not* to *Botometer*, increasing the set of features to 1,150 account related features [5]. Their team compared Random Forests, AdaBoost, Logistic Regression and Decision Tree classifiers and still found that Random Forests performed best. They also attempted to update their training data by manually annotating tweet accounts, and merging this with the original Texas A&M Dataset (collected in 2011).

The continued use of the 2011 Texas A&M data highlights the difficulty that researchers have in creating and/or updating the labeled data that is used train algorithms to find these automated accounts. The use of aging training data for bot classification also ensures that emerging bots are likely to avoid detection. Additionally, since bots have a variety of purposes as well as a spectrum of actors that create/use them, the collection technique used for labeled data will bias the detection toward that family of bots. For example, the *honey pot* collection technique will bias toward bots that randomly follow accounts, but may not detect intimidation bots that conduct targeted following and messaging.

2.2 Classifying Algorithmic Character Strings

Classifying strings as *random* or *not random* in order to filter or flag anomalous events has a limited background.

Several methods have been proposed for identifying or highlighting the randomness of character strings. Some have proposed leveraging Shannon's Entropy calculation [18] as a method for sorting strings by a measure of randomness. Some cyber security research teams have proposed a similar detection methods in order to detect domain names that are generated by Domain Generation Algorithms (DGA). These teams have separately used Kullback-Leibler Divergence [21], a dictionary approach [15] and Markov modeling [17].

The past research most closely connected to our effort was conducted by LinkedIn in 2013. At that time [6] presented the application of the Naive Bayes model on Character N-grams features of LinkedIn account names in order to identify *spammy* accounts (first and last name as provided by the account owner). This effort was very effective, and replaced the legacy spam detection models that LinkedIn was using on their OSN. To date, our team has not found any team that has replicated a similar approach to Twitter screen names.

2.3 Project Background

Our team has focused on detecting, characterizing, and modeling the behavior of bots, bot networks and their creators. In doing this we've studied several recorded bot events. Recently we focused on a known and publicized bot attack against the Atlantic Council Digital Forensic Labs (DFR Lab), and tangentially against the NATO Public Affairs Office. This attack primarily occurred between August 28 and August 30, 2017. We also focused on a recorded bot harassment event against journalists in Yemen [13]. In both events we observed numerous bot accounts that used 15 character randomly generated alpha-numeric strings for the screen name.

Examples of this include **Wy3wU4HegLlvHgC**, **5JSQavWW3tvQwA7**, and **gG6RKc6QBqOLKyU** (these are not real Twitter accounts). Note that these randomly generated strings always sample from upper and lower case alpha-numeric characters. Observing this phenomenon motivated the construction of this algorithm and its application on Twitter at large in order to observe other bots and bot actors that are using these same type of bot screen names. More importantly, we hope this dataset can be used as a large and diverse annotated bot training data for larger and more comprehensive machine learning models.

3 Modeling

3.1 Feature Engineering

In order to develop a random string detection model for this unique case, we constructed training data consisting of 4,000 non-random Twitter screen names (randomly sampled from Twitter and manually verified as non-random) and 4,000 randomly generated 15 digit strings. We then developed a combination of heuristic filtering and traditional machine learning models to label the string as *random* or *not random*. This development is described below.

For feature engineering, the primary feature that we extracted from the strings was character n-gram. For string s with length m, a character n-gram is the $(m - n + 1)$ sequential substrings of length n found in string s. In our case, we explored several settings for n, to include using multiple values in the same feature set (i.e. using both diagrams and trigrams).

We then transformed the resulting sparse character n-gram matrix using term frequency-inverse document frequency (TF-IDF). TF-IDF is defined in Eqs. 1 and 2 below, and is used to scale the characters by the information that they provide. In our case, frequent characters in a string provide information, but not if they're frequent in all of the strings. To calculate the IDF for character c in strings s, we take the logarithm of the ratio of the total number of strings in corpus S by the number strings that contain c, as shown in Eq. 1.

$$idf(c, S) = log \frac{N}{|\{c \in S : c \in s\}|} \tag{1}$$

We then calculate the TF-IDF for character c in string s found in corpus S as follows

$$tfidf(c, s, S) = tf(c, s)\dot{i}df(c, S) \tag{2}$$

This therefore weights characters that have a high local frequency but a lower global frequency. At first it may seem that TF-IDF is unnecessary since each character n-gram is equally likely in random strings, given a strong pseudo-random number generator. n-grams are not equally likely for human generated strings, however. Given this fact we felt it appropriate to transform the data with TF-IDF.

These features were merged with several other features. We started by merging the normalized count of upper case, lower case, and numeric characters. n-gram generation by default converts all text to lower case. We maintained this default behavior, but saw that the number of upper and lower case in letters in particular provided a strong signal. Since our training data contained some human generated strings that were not 15 characters in length, we normalized these counts.

Additionally, we included the Shannon string entropy in our feature set. Shannon string entropy, while not strong enough to use by itself in our case, still provides a strong signal that we felt would be useful. We will test this assumption below. Shannon entropy is defined in 3, where p_i is the normalized count for each character found in the string.

$$H\left(A\right) = -\sum_{i=1}^{n} p_i log_2 p_i \tag{3}$$

The A full table of features is given in Table 1.

Table 1. Features for random string detection

Feature	Type	Description
2–3 character N-gram	Numeric	Term frequency inverse document frequency of n-gram
No. lower case	Numeric	Normalized count of lower case letters
No. upper case	Numeric	Normalized count of upper case letters
String entropy	Numeric	Shannon String entropy

We used the $scikit - learn$ package [16] to explore and build the machine learning classification model for Random Strings. We evaluated Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) with 10 fold cross-validation. The results are presented in Table 2. We conducted model comparisons between these models, and found that the SVM models provided *Highly Significant Improvement* in all cases ($p.value < 0.001$). Given these results, we used SVM for our production model.

Table 2. Model performance in classifying randomly generated strings for screen-names

Model	% Correct	Kappa
Naive Bayes	93.3%	0.8659
Logistic regression	94.25%	0.948
SVM	97.4%	0.975

Before predicting whether or not a string was random, we first applied several heuristic filters. These verified that (1) the string was 15 characters in length, and (2) contained at least one capital letter, lower case letter, and numeric digit. This final filter was applied given that 15 character strings have a 0.02% chance of not containing a capital or lower case letter and a 7% chance of not containing a numeric digit. This heuristic was applied given that precision was a higher priority than recall.

In Fig. 1 we evaluate the best value of n (number of characters for n-gram) as well as whether or not using Shannon's Entropy as a column feature provides leverage in prediction. In this visualization we see that digrams with Shannon's entropy provides the best leverage in predicting random strings.

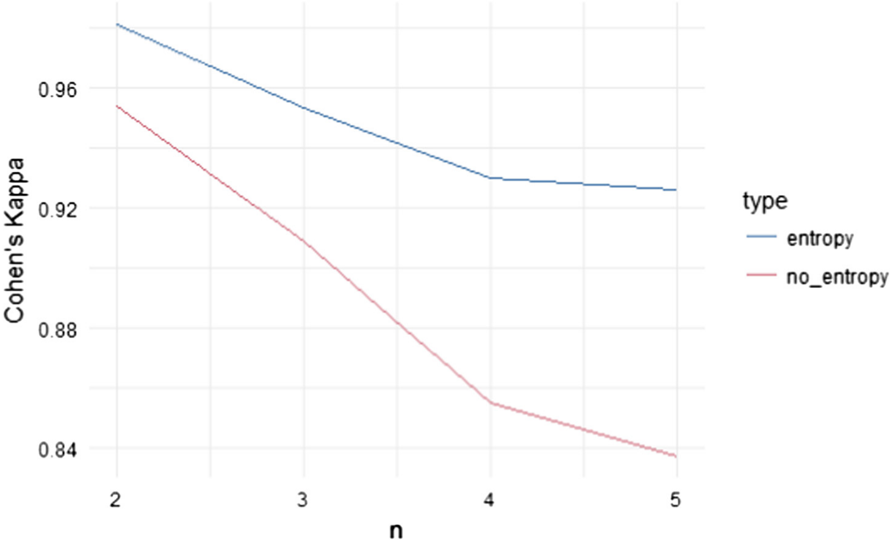


Fig. 1. Evaluating n (number of characters in n-gram) and use of Shannon's entropy as a feature

In addition to exploring the feature based machine learning models discussed above, we also explored the use of Markov model of character sequencing, but found during initial exploration that this did not have sufficient power to classify the strings given the inherent random nature of human generated screen names. Additionally, we explored using Shannon entropy as the only measure for filtering these strings. Once again, while helpful, this method did not demonstrate sufficient power for our purposes.

3.2 Model Deployment

Our primary use for the algorithm was to filter accounts with 15 character random strings from a Twitter data stream. To do this we ran a random sample from the Twitter Streaming API from 14 December 2017 to 10 January 2018. During this time the stream collected approximately 33 million tweets. This collection was done without any text or geographic filters, and stored the raw JSON files that are returned by the Twitter API.

Having performed the collection, we next applied our algorithm to all 33 million tweets, filtering out all accounts that were labeled as having 15 digit randomly generated screen name. This produced a collection of 487,000 tweets from 235,000 unique accounts.

4 Model Evaluation

Given the desired use case of annotating diverse bot accounts, we conducted two evaluations on our results. First, we wanted to estimate the false positive rate on our random string detection, since false positives have a high likelihood of not being an autonomous account. To accomplish this we randomly selected 1,000 of the screen names that were labeled as random, and manually identified those that contained clear words or acronyms. Given this method, we estimate that our false positive rate is 5.6%.

Additionally, we wanted to estimate the percentage of random character screen name accounts that are autonomous, or appear autonomous. In other words, how many of our true positive random string accounts are truly autonomous. To estimate this, we randomly sampled 100 accounts, verified that the user name appeared random, and inspected the account in the Twitter web client. Of the 100 that we manually inspected, five were suspended, eight provided no results (most likely the account was closed by the user), and all others exhibited autonomous behavior. After thoroughly evaluating these 100 randomly sampled accounts we were satisfied that this methodology provides annotated bot data that is at least as accurate as honey pot data, and likely has a wider range of bot types.

4.1 Data Characterization

One of our first tasks in exploring the data was to check on those features that are highly indicative of an autonomous account to see if our data exhibited these tell-tale characteristics. In general, autonomous accounts produce tweets at a much higher volume and rate than human actors. The mean number of tweets for our accounts was 7,918 (median = 1,125). In general bots have a low number of followers (most people don't follow bots), but they tend to follow many accounts, trying to build influence. Following this pattern, the median number of followers is in our data set is 55, but the median number of accounts they follow is 130.

94% of the roughly 500,000 tweets in this dataset are associated with seven languages. Somewhat surprisingly, the volume associated with Japanese and Arabic accounts is greater than those associated with English speaking accounts.

A full breakdown of the languages and a short general description of our observations are provided in Table 3. Only 674 tweets contained coordinate locations, and these locations are strongly correlated to the languages mentioned below.

Table 3. Characterization by language

Language	% Total	General description
Japanese	184, 385	High concentration of anime media sharing
Arabic	111, 523	High percentage of young accounts, some automated Koran passage sharing
English	94, 804	Contains a high number of non English hash tags
Korean	41, 870	Varied
Thai	14, 195	High concentration of adult content
Russian	13, 461	Varied

The mean age of the accounts is 274 days, with 50% of the accounts created in the last 150 days and 75% of the accounts younger than 1 year old. The relative young age of these accounts is highly indicative of their automated behavior. The oldest accounts are associated with English account settings, and date back to 2008.

Given the fact that our data set contains primarily bot accounts, we observed a number of account suspensions during the course of our study. Between mid December 2017 and mid March 2018, 23,532 accounts (~10%) were suspended by Twitter, while 2,201 accounts (~1%) were removed by the user. As the media and politicians put pressure on Social Media companies, the natural response is to increase their policing of this autonomous behavior on their platforms.

5 Conclusion

Research in this area is limited by a rich enough data set that supports identification of the wide range of types of bots, and that is sufficient to support studies of bot-evolution. While the data used herein begins to address this issue, it is by no means comprehensive and needs further expansion. We are working on such expansion. However, restrictions on data sharing make it difficult to share this data. Consequently, we are also working on data format that can be shared.

Bots are part of the conversation in social media. But not all bots are the same. They vary in what they do, how they do it, and intent. While some bots act independently others work in concert and still others are part of a cyborg - a human-bot partnership. Research is needed to characterize types of bots and their evolution. Research is also needed to identify the mapping between types of bots in use and types of information maneuver or social-group creation that, that type of bot supports or thwarts.

6 Future Work

Our future effort begins with the exploration of this dataset so that we can cluster these accounts by type and function. We then intend to develop and train several specialized as well as a general purpose bot detection algorithms for use in detecting and classifying bots. Once complete, our effort will shift to the detection and characterization of bot networks and the actors behind them.

Acknowledgment. This work was supported in part by the Office of Naval Research (ONR) Multidisciplinary University Research Initiative Award N000140811186 and Award N000141812108, the Army Research Laboratory Award W911NF1610049, Defense Threat Reductions Agency Award HDTRA11010102, and the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the ONR, ARL, DTRA, or the U.S. government.

References

1. Benigni, M., Carley, K.M.: From tweets to intelligence: understanding the Islamic jihad supporting community on Twitter. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) SBP-BRiMS 2016. LNCS, pp. 346–355. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39931-7_33
2. Bessi, A., Ferrara, E.: Social Bots Distort the 2016 US Presidential Election Online Discussion (2016)
3. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on Twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 21–30. ACM (2010)
4. Davis, C.A., Varol, O., Ferrara, E., Flammini, A., Menczer, F.: Botornot: a system to evaluate social bots. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274. International World Wide Web Conferences Steering Committee (2016)
5. Ferrara, E.: Measuring social spam and the effect of bots on information diffusion in social media. arXiv preprint arXiv:1708.08134 (2017)
6. Freeman, D.M.: Using Naive Bayes to detect spammy names in social networks. In: Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security, pp. 3–12. ACM (2013)
7. Freitas, C., Benevenuto, F., Ghosh, S., Veloso, A.: Reverse engineering socialbot infiltration strategies in Twitter. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 25–32. ACM (2015)
8. Glaser, A.: Russian bots are trying to sow discord on Twitter after charlottesville (2017)
9. Graham, T., Ackland, R.: Do socialbots dream of popping the filter bubble? In: Socialbots and Their Friends: Digital Media and the Automation of Sociality, p. 187 (2016)
10. Howard, P.N., Kollanyi, B.: Bots,# strongerin, and# brexit: computational propaganda during the uk-eu referendum. Browser Download This Paper (2016)

11. Krishnamurthy, B., Gill, P., Arlitt, M.: A few chirps about Twitter. In: Proceedings of the First Workshop on Online Social Networks, pp. 19–24. ACM (2008)
12. Lee, K., Eoff, B.D., Caverlee, J.: Seven months with the devils: a long-term study of content polluters on Twitter. In: ICWSM (2011)
13. Al Bawaba The Loop. Thousands of Twitter bots are attempting to silence reporting on Yemen (2017)
14. Lumezanu, C., Feamster, N., Klein, H.: # bias: measuring the tweeting behavior of propagandists. In: Sixth International AAAI Conference on Weblogs and Social Media (2012)
15. Namazifar, M.: Detecting randomly generated strings, December 2015. Accessed 25 Dec 2015
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
17. Raghuram, J., Miller, D.J., Kesidis, G.: Unsupervised, low latency anomaly detection of algorithmically generated domain names by generative probabilistic modeling. J. Adv. Res. **5**(4), 423–433 (2014)
18. Shannon, C.E.: The bell system technical journal. In: A Mathematical Theory of Communication, vol. 27, pp. 379–423 (1948)
19. Subrahmanian, V.S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F.: The DARPA Twitter bot challenge. Computer **49**(6), 38–46 (2016)
20. Verkamp, J.-P., Gupta, M.: Five incidents, one theme: Twitter spam as a weapon to drown voices of protest. In: FOCI (2013)
21. Yadav, S., Reddy, A.K.K., Reddy, A.L., Ranjan, S.: Detecting algorithmically generated malicious domain names. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, pp. 48–61. ACM (2010)

Understanding Cyber Attack Behaviors with Sentiment Information on Social Media

Kai Shu[1](✉), Amy Sliva[2], Justin Sampson[1], and Huan Liu[1]

[1] Computer Science and Engineering, Arizona State University, Tempe, AZ, USA
{kai.shu,jsampso1,huan.liu}@asu.edu
[2] Charles River Analytics, Cambridge, MA, USA
asliva@cra.com

Abstract. In today's increasingly connected world, cyber attacks have become a serious threat with detrimental effects on individuals, businesses, and broader society. Truly mitigating the negative impacts of these attacks requires a deeper understanding of malicious cyber activities and the capability of predicting these attacks before they occur. However, detecting the occurrence of cyber attacks is non-trivial due to the anonymity of cyber attacks and the ambiguity or unavailability of network data collected within organizations. Thus, we need to explore more nuanced auxiliary information that can provide improved predictive power and insight into the behavioral factors involved in planning and executing a cyber attack. Evidence suggests that public discourse in online sources, such as social media, is strongly correlated with the occurrence of real-world behavior; we believe this same premise can provide predictive indicators of cyber attacks. For example, extreme negative sentiments towards an organization may indicate a higher probability that it will be the target of a cyber attack. In this paper, we propose to use sentiment in social media as a sensor to better understand, detect, and predict cyber attacks. We develop an effective unsupervised sentiment predictor model utilizing emotional signals, such as emoticons or punctuation, common in social media communications, and a method for using this model as part of a logistic regression predictor to correlate changes in sentiment to the probability of an attack. Experiments on real-world social media data around well-known hacktivist attacks demonstrate the efficacy of the proposed sentiment model for cyber attack understanding and prediction.

Keywords: Cyber attack · Sentiment analysis · Social media

1 Introduction

As networked and computer technologies continue to pervade all aspects of our lives, the threat from cyber attacks has also increased. The broad range of increasingly common cyber-attacks, such as DDoS attacks, data breaches,

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 377–388, 2018.
https://doi.org/10.1007/978-3-319-93372-6_41

and account hijacking, can have an extremely detrimental impact on individuals, businesses, and broader society. Thus, understanding these attacks and predicting them before they occur is an emerging research area with widespread applications. However, detecting attacks, much less predicting them in advance, is a non-trivial task due to the anonymity of cyber attackers and the ambiguity of network data collected within an organization; often, by the time an attack pattern is recognized, the damage has already been done. Evidence suggests that the public discourse in external sources, such as news and social media, is often correlated with the occurrence of larger phenomena, such as election results or violent attacks. Social media, in particular, turns users into "social sensors" empowering them to participate in an online ecosystem that interacts with behavior in the physical world. We believe the same principle can apply to cyber attacks, where open source data may provide indicators to help understand the social and behavioral phenomena leading up to an attack.

In this paper, we propose an approach that uses sentiment polarity as a sensor to analyze the social behavior of users on social media as an indicator of cyber attack behavior. For example, extreme negative sentiment towards an organization may indicate a higher probability of it being the target of a cyber attack. However, measuring sentiment itself in social media is a challenging task due to: (1) the data challenge, where ground truth datasets with sentiment labels are often unavailable; and (2) the feature challenge, where effective and robust features must be extracted from short and noisy social media posts. Both challenges make standard supervised sentiment analysis methods inapplicable. Instead, we developed an unsupervised sentiment prediction method that utilizes emotional signals to enhance the sentiment signal from sparse textual indicators. In this model, we incorporate both emotion words and emoticons separately, as well as modeling the correlations among them in an unsupervised manner.

To explore the efficacy of sentiment polarity as an indicator of cyber-attacks, we performed experiments using real-world data from Twitter that corresponds to known attacks by a well-known hacker group. The experimental results show that the proposed sentiment prediction framework can recognize distinct behavioral patterns associated with these attacks. We also performed a temporal analysis on the sentiment for these attacks, which provides deeper understanding of the progression of ongoing cyber attack behaviors over time. Our contributions are summarized as follows:

(a) We propose to utilize sentiment polarity in social media as a sensor to understand and predict the social behaviors related to cyber attacks;
(b) We develop an unsupervised sentiment analysis using emotional signals, which models emotion indications without requiring labeled sentiment data beforehand; and
(c) We conduct experiments on real-world Tweet data related to several cyber-attacks by a well-known hacker group to demonstrate the effectiveness of the proposed sentiment prediction framework.

2 Related Work

Our related work mainly falls into the following two categories: (1) sentiment analysis on social media; and (2) cyber attack analysis on social media.

Sentiment analysis on social media. Sentiment analysis has been an important task for natural language processing, and has been widely used in various social media applications, such as poll rating prediction [15], stock market prediction [2], fake news [21], emoji analysis [14] and so on. Existing methods can be categorized as either supervised [5,8], meaning they are trained on labeled ground-truth data, and unsupervised [7,9,15], which are not trained on labeled data, but rather find patterns or groups in the existing datasets. Due to the lack of label information and the large-scale data produced by social media, unsupervised learning becomes more and more important in real-world social media sentiment analysis applications. Unsupervised methods often rely on a pre-defined sentiment lexicon to determine the sentiment score. The lexicon words are collected from (1) human annotators, such as in the General Inquirer [22] and Multi-Perspective Question Answering (MPQA) corpus [24] work; (2) a dictionary that contains semantic/linguistically related words, such as WordNet [16]; or (3) a corpus that can be used to infer sentiment polarity of words by exploring the relation between the words and some observed seed sentiment words in the corpus [15]. Recently, Hu *et al.* proposed a new state-of-the-art unsupervised sentiment analysis method that specifically leverages the way people communicate on social media, utilizing emoticon information, punctuation, and other sources of emotional signals to better predict sentiment on social media posts [7]. In this paper, we build on the success of this method to develop our sentiment predicting approach.

Cyber attack analysis on social media. In recent years, online social media has been a promising source of cyber attack analysis and understanding, such as threat intelligence fusion [13], malicious cyber discussion detection [12], etc. One line of research is to utilize social media platforms in specific domains to extract expert information as indication features [11,18,23]. In [11], Liao *et al.* utilize technology blog posts to extract key attack identifiers, such as source IP and MD5 hashes. Sabottke *et al.* estimate the level of interest in existing common vulnerabilities and exposures (CVE) and further predict the indication probability for real attacks [18]. Ritter *et al.* extract relevant Tweets of specific event with only a small set of supervised information to better collecting useful data for cyber decision making [17]. Recently, Khandpur *et al.* used social media as a crowd-sourced sensor to gain insights into ongoing cyber attacks by adapting queries for searching Tweets, and better predict attacks, such as DDOS attacks [10].

3 Sentiment Sensor Modeling

In this section, we introduce our framework of predicting sentiment polarity in an unsupervised way. Then, we discuss how we build temporal sentiment analysis over time, which enables correlation of sentiment trends with other time series of real-world events, such as cyber attacks.

3.1 Unsupervised Sentiment Extraction

The proposed model is motivated by the observation that social media communication, such as Twitter, includes emotional signals (e.g., emoticons, specialized punctuation.) that could be strongly correlated with the sentiment in a social media post or the words in it. Our goal is to use the emoticons in social media posts to indicate the sentiment score of entire posts. Specifically, we aim to model the following emoticon information:

Table 1. List of emoticons with sentiment polarity

Positive	**:-), (-:, =), (=, (:, :), :D, :d, d:, :), (:, 8), (8, 8), ;), ;), ;), (;, ;-), (-;, (;, ^ _ ^**
Negative	**:-(, :-(,)-:, =(,)=, :(,):, 8(,)8**

Post-level Emoticon Indication. Based on sentiment consistency theory [18], post level emotion indication assumes the strong correlation of sentiment polarity of a post and the corresponding emotion signals. The key idea of modeling post-level emotion indication is to make the sentiment polarity of a post as close as possible to the emotion indication.

Word-level Emoticon Indication: The overall sentiment of a post is also positively correlated with the sentiments of the words in that post. By modeling word-level emotional signals, we can utilize the valuable information in the sentiment analysis framework to infer sentiment polarity of a post.

To model post-level emoticon indication, we can build a classifier $y = f(\mathbf{x})$, where $y \in [0, 1]$ indicates the sentiment indicated by the emoticons themselves in the posts, and $\mathbf{x}$ represents the list of features can be extracted from social media posts. As shown in Table 1, there are commonly used emoticons that correspond to positive and negative sentiments. In addition, to model the word-level indication, we extract different types of features from post text, including platform-independent and platform-specific features, on social media.

For platform-independent features $\mathbf{x}^{(1)}$, we adopt the widely used n-gram features [4] with TF-IDF adaption to capture word-level patterns. We consider both the term frequency (TF) and inverse document frequency (IDF) to compute $\mathbf{x}^{(1)}$ in a post. Let $\mathcal{V} = \{w_1, w_2, \ldots, w_n\}$ denotes the vocabulary of the entire corpus, and $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$ denotes the set of all social media posts. Then the TF score for for word w_i in document d_j is computed as: $tf_{ij} = 1$ if $w_i \in d_j$, otherwise $tf_{ij} = 0$. The IDF measures whether a specific word is common or rare in the corpus, and it is computed as $idf_{(ij)} = \log \frac{m}{1+|d_j \in \mathcal{D}: w_i \in d_j|}$, where m is the total number of posts, $|d_j \in \mathcal{D} : w_i \in d_j|$ is the number of posts that word w_i appears in post d_j. Thus, we have the platform-independent feature vector computed as $\mathbf{x}_{ij}^{(1)} = tf_{ij} \times idf_{ij}$. The feature vector for each post is $\mathbf{x}_j^{(1)} = \mathbf{x}_{1j}^{(1)} \oplus \mathbf{x}_{2j}^{(1)} \cdots \oplus \mathbf{x}_{nj}^{(1)}$, and $\oplus$ is the concatenation operation.

For platform-dependent features $\mathbf{x}^{(2)}$, we aim to capture the specific linguistic patterns in the particular social media platform. In Twitter, for example, we apply the following heuristics to obtain additional features as shown in Table 2. We selected these features because they provide useful indications of sentiment. For example, the question mark and exclamation marks can usually indicate a stronger sentiment strength. By concatenating all these features, we can obtain platform-dependent feature vectors $\mathbf{x}_j^{(2)}$ for post d_j. Finally, we combine platform-independent and platform-dependent features together, and get $\mathbf{x}_j = \mathbf{x}_j^{(1)} \oplus \mathbf{x}_j^{(2)}$.

We can next apply existing widely applied classifiers f to build a sentiment prediction model using these extracted features $\mathbf{x}$, such as Naïve Bayes, decision trees, logistic regression, K-nearest neighbors (KNN) clustering, and support vector machines (SVM). In this paper, we empirically adopt logistic regression as the classifier due to the fact that it is simple to train and understanding, but also very effective as a classifier. Note that even though the proposed model requires posts with emoticons to learn model parameters, it can predict the sentiment of posts without emoticons. The predicted sentiment score is represented by $\hat{y} \in [0, 1]$; the large the predicted value, the more positive the sentiment.

Table 2. Platform-specific features in social media posts

Feature	Description
HASHTAG	The number of hashtag in the post, e.g., #cybersecurity
QUESTIONMARK	The number of question marks (**?**) in the post
EXCLAMATION	The number of exclamation marks (**!**) in the post
NEGATION	The number of negative words in the post, e.g., **not**
TEXT_LEN	The length of the post by removing irrelevant mentions and URLs

3.2 Temporal Sentiment Analysis

Sentiment time series have been widely used for event prediction, such as political election prediction [1], and disaster event detection [19]. Similar, in cyber attack scenario, we are not only interested in predicting the sentiment score of individual post, but also the temporal variation of sentiments over time. We aim to provide insights on: (1) characterizing how social media users change the sentiment polarity towards public events; (2) understanding how sentiments can indicate upcoming attack events; and (3) assessing the attack effects after the attack.

To tackle these questions, we collect the related social media posts $\mathcal{D}$ by querying relevant keywords (detailed in Sect. 4.1) in social media within a specific time range $\mathcal{T}$ that covers the time interval before, during, and after the attack event occurs. We build the sentiment time series $S = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ chronologically by using the pre-learned sentiment predictor. We can also group and averaging the sentiments in different time granularities, such as per day, to explore the sentiment variations.

4 Experiments

In this section, we describe experiments we conducted using real-world datasets to demonstrate the effectiveness of the sentiment prediction approach described above and the ability to use this as a sensor for identifying cyber attack behavior over time and predicting future attacks.

4.1 Datasets

For our experiments, we used several different datasets. First, to empirically test the efficacy of using sentiment in social media as a sensor for indicating future cyber attacks, we first collected historical Twitter data using Gnip[1]. Gnip allows for historical queries against Twitter to be grouped by tags (i.e., topics) of interest. For looking at cyber attack behavior, we grouped our keywords into the following tags: (1) attack sources (i.e., tweets from or about known hacking organizations); (2) DDOS; (3) phishing; (4) exploits; (5) cyber security; (6) vulnerability announcements; (7) vulnerabilities; (8) CVEs; and (9) specific attack targets of interest. For each tweet, we used the sentiment analysis approach defined above to determine the sentiment value. We then aggregated over this data to create a time series with the mean sentiment scores per tag per day. It is used in combination with cyber incident reports provided by a financial company, C1, between April 2016 and September 2016, and a defense company, C2, between November 2016 and February 2017[2] to develop a predictive model of cyber attacks (see Sect. 4.2). The reports provide details on three kinds of attacks: malicious email, malicious URL, and malware on endpoint.

In addition, to analyze the behavioral patterns associated with cyber attacks, we also collected another dataset from Twitter related to a well-known hacktivist group. We identified three attack events perpetrated by this group from 2016–2017, designated A_1, A_2, and A_3[3]; these are different from the cyber attack incident data described above because these focus on hacktivist attacks that are known to be reactions to certain societal events, rather than typical cyber

[1] http://support.gnip.com/.

[2] The names of the companies have been anonymized.

[3] The attack events are anonymized here.

behavior targeting individual companies. Note that these attacks are sometimes benign behaviors, which means they are performed not for malicious intent (e.g., stealing credentials.) but more as a form of online protest or demonstration by a group seeking to influence societal events. Based on the three selected attacks, we developed specific keywords that were use to query GNIP, gathering tweets that occurred up to 3 weeks before and 1 week after the attack.

4.2 Experimental Results

Sentiment Clustering. Our first experiment seeks to evaluate the effectiveness of extracted features $\mathbf{x}$ for sentiment prediction, using cluster analysis to assess the performance of our unsupervised sentiment model. Because we do not have access to ground truth, we cannot compute standard accuracy measures. However, a cluster analysis will enable us to measure the quality of the patterns discovered by the unsupervised sentiment analysis. We try to answer the following questions: (1) Are the proposed sentiment features able to cluster all the tweets into distinct clusters that match intuitive understanding of sentiment? and (2) What is the proper cluster size we should use decide to sentiment degree?

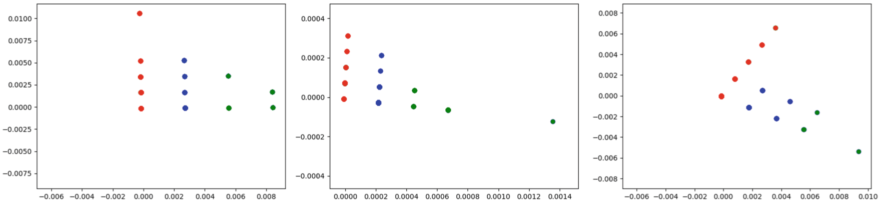
To answer these questions, we use k-means clustering [6] based on an extracted feature vector $\mathbf{x}_j$ for each post d_j, including the label assigned by our emotion-based sentiment analysis technique. The clustering performance is evaluated using the standard concepts of separation (i.e., the difference between elements in different clusters) and cohesion (i.e., the similarity of elements in the same cluster) captured in the widely used silhouette score metric. The Silhouette Score s is defined as $s = \frac{1}{m}\sum_{j=1}^{m}\frac{b(d_j) - a(d_j)}{\max(b(d_j),a(d_j))}$, where $a(d_j) = \frac{1}{|C|}\sum_{d_k\in\mathcal{D},d_k\neq d_j}\|\mathbf{x}_j - \mathbf{x}_k\|^2$ indicates the within-cluster average distance (cohesion) in cluster C, and $b(d_j) = \min\frac{1}{|G|}\sum_{d_k\in G}\|\mathbf{x}_k - \mathbf{x}_j\|^2$ indicates the distance of d_j with posts in other clusters (separation). Note that $s \in [-1, 1]$, and the higher the score, the clusters show better separation from each other and a greater degree of internal consistency. If our unsupervised sentiment analysis approach is successful, it will produce results that have a high silhouette score across clusters that seem consistent with different sentiment categories. We also applied Priciple Component Analysis (PCA) for feature dimension reduction to better visualize the results of our cluster analysis. The results for A1, A2, and A3 are shown as in Table 3.

We can see better silhouette scores when we use three clusters over the sentiment feature space (i.e., positive, negative, and neutral) in all three cases. The high scores also indicate that the clusters are well separated and internally cohesive, indicating that the sentiment prediction model is able to use the emotional signal features to classify sentiment with a high degree of discrimination.

We apply PCA to project the original feature space to low dimensions for easy visualization. As shown in Fig. 1, we can see that the cluster analysis results for the A_1, A_2, and A_3 attacks. In all cases, we observe three very distinct clusters for positive, negative, and neutral sentiment, which is consistent with the high silhouette scores.

Table 3. Results of post clustering on sentiment features

Dataset	PCA	2		3		No	
	Cluster size	2	3	2	3	2	3
A1	Silhouette score	0.914	0.976	0.865	0.908	0.865	0.925
A2	Silhouette score	0.921	0.981	0.803	0.914	0.853	0.902
A3	Silhouette score	0.908	0.986	0.917	0.942	0.903	0.914



(a) A1 (b) A2 (c) A3

Fig. 1. Cluster analysis visualizations case studies on hacker event dataset

Temporal Variation Analysis. We conducted temporal sentiment analysis on attack events A_1, A_2, and A_3. The motivation to analyze sentiment variation is that it could provide informative behavioral indicators to predict attack events based on trends in public discourse on social media.

As shown in Figs. 2 and 3, the average sentiment scores are strongly correlated with the public event that preceded the attack. We have the following observations; (1) before the attack happens, the sentiment scores tend to be relatively stable, which may indicate the normal public discussion among users about a particular event; (2) while several days before the attack happens, the sentiment scores are very strongly negative, which may reflect the general public's unsatisfied attitude towards the event and indicate the potential for an upcoming attack; and (3) after the attack happens, the sentiment tends to increase again, which may indicate the positive response of social media users to the attacks (or the changes in the discussion precipitated by the attacks). Thus, there are distinct behavioral patterns in sentiment over time for indicating cyber attacks.

Cyber Attack Prediction Using Sentiment. Finally, we evaluate the sentiment trends for actually predicting cyber attacks before they occur. We use our historical data from Twitter and the cyber incident reports from C1 and C2 to develop a logistic regression classifier for each event type (i.e., malicious email, malicious URL, and endpoint malware). The features are the aggregate sentiment scores per tag/topic of interest, and the classes are the probability of an attack occurring. We varied the time lag between the sentiment scores and an attack between 0 and 10 days. We divide the data into a training set and a testing set in which the training set includes the first 80% of the time period the

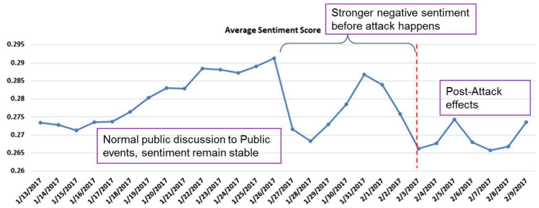


Fig. 2. Sentiment temporal variations on attack A1.

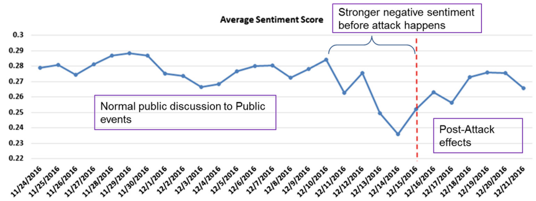


Fig. 3. Sentiment temporal variations on attack A2.

data covers, and the test set includes the other 20%. In each test, the logistic regression models predict whether or not an attack of each type occurs.

The results are summarized in Table 4. We found that the model performs much better in predicting either malicious-email or endpoint-malware attacks as

Table 4. Results for predicting endpoint malware attacks using sentiment sensor

Attack type	Warning threshold	Accuracy	Precision	Recall	F1
Malicious email	*0.1*	0.5	0.905	0.5	0.644
	0.2	0.5	0.905	0.5	0.644
	0.3	0.905	0.905	0.5	0.580
	0.4	0.905	0.905	1	0.856
	0.5	0.905	0.905	0.905	1
Malicious URL	*0.1*	0.5	0.170	0.5	0.253
	0.2	0.5	0.170	0.5	0.253
	0.3	0.5	0.170	0.5	0.253
	0.4	0.5	0.170	0.5	0.253
	0.5	0.170	0.170	0.170	1
Endpoint malware	*0.1*	0.5	0.801	0.5	0.615
	0.2	0.5	0.801	0.5	0.615
	0.3	0.5	0.801	0.5	0.555
	0.4	0.801	0.801	1	0.802
	0.5	0.801	0.801	0.801	1

opposed to the malicious-URL attack type, with very high precision and recall scores for both of these. Our data also showed that the time lag for Twitter events is rather small, with more successful prediction occurring with a time lag of between 1 and 3 days. In addition, more variation in the false positive and true positive rate is seen at even higher thresholds between 0.6 and 0.7, and using these we are able to generate the ROC curve as shown in Fig. 4 for the malicious email event. This indicates that while sentiment shows promise as a predictor of cyber attacks, it is still only a weak signal and may need to be combined with other evidence or further amplified.

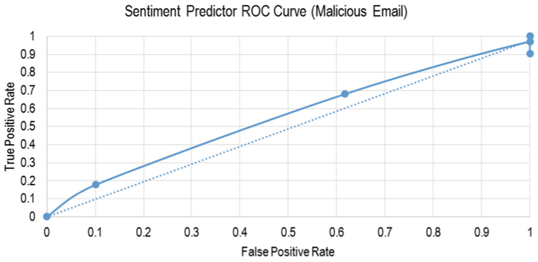


Fig. 4. The ROC curve for sentiment prediction for C2 malicious email attack

5 Conclusion and Future Work

In this paper, we use sentiment in social media as a sensor for understanding and predicting cyber attacks. The proposed sentiment extractor works in an unsupervised way utilizing emoticon signals for model learning. Experiments on real world datasets demonstrate the ability of sentiment score to (1) capture temporal correlations between attack events and inherent factors with ongoing public discourse; and (2) predict real-world cyber attacks, such as malicious email, malicious URLs, and endpoint malware against particular targets.

There are several interesting future directions. First, we can explore temporal process models, such as hawk process [3] for better modeling the sentiment variations over time for cyber attack prediction. Second, we can build temporal correlation networks [20] among general and specific attack Tweets to better predict the intensity of ongoing attacks. Third, we can explore other social features, such as credibility and veracity, to better understand the underlying social and behavioral patterns to help improve our cyber attack predictions.

Acknowledgements. This material is based upon work supported by, or in part by, the ONR grant N00014-16-1-2257 and N00014-17-1-2605, and the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0108. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, ONR, or the U.S. Government.

References

1. Bermingham, A., Smeaton, A.: On using twitter to monitor political sentiment and predict election results. In: SAAIP 2011 (2011)
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011)
3. Da Fonseca, J., Zaatour, R.: Hawkes process: fast calibration, application to trade clustering, and diffusive limit. J. Futures Markets **34**(6), 548–579 (2014)
4. Fürnkranz, J.: A study using n-gram features for text categorization. Austrian Res. Inst. Artif. Intell. **3**(1998), 1–10 (1998)
5. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford. vol. 1, 12 (2009)
6. Hartigan, J.A., Wong, M.A.: Algorithm AS 136: A k-means clustering algorithm. J. Roy. Stat. Soc. C (Appl. Stat.) **28**(1), 100–108 (1979)
7. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: WWW 2013 (2013)
8. Hu, X., Tang, L., Tang, J., Liu, H.: Exploiting social relations for sentiment analysis in microblogging. In: WSDM 2013 (2013)
9. IU, J.B., IU, H.M.: Twitter mood predicts the stock market (2011)
10. Khandpur, R.P., Ji, T., Jan, S., Wang, G., Lu, C.T., Ramakrishnan, N.: Crowdsourcing cybersecurity: Cyber attack detection using social media. arXiv preprint arXiv:1702.07745 (2017)
11. Liao, X., Yuan, K., Wang, X., Li, Z., Xing, L., Beyah, R.: Acing the IOC game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In: SIGSAC 2016, (2016)
12. Lippmann, R.P., Campbell, J.P., Weller-Fahy, D.J., Mensch, A.C., Campbell, W.M.: Finding malicious cyber discussions in social media. Technical report, MIT Lincoln Laboratory Lexington United States (2016)
13. Modi, A., Sun, Z., Panwar, A., Khairnar, T., Zhao, Z., Doupé, A., Ahn, G.J., Black, P.: Towards automated threat intelligence fusion. In: ICIC 2016 (2016)
14. Morstatter, F., Shu, K., Wang, S., Liu, H.: Cross-platform emoji interpretation: analysis, a solution, and applications. arXiv preprint arXiv:1709.04969 (2017)
15. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: ICWSM 2010 (2010)
16. Peng, W., Park, D.H.: Generate adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. Urbana **51**, 61801 (2004)
17. Ritter, A., Wright, E., Casey, W., Mitchell, T.: Weakly supervised extraction of computer security events from twitter. In: WWW 2015 (2015)
18. Sabottke, C., Suciu, O., Dumitras, T.: Vulnerability disclosure in the age of social media: exploiting twitter for predicting real-world exploits. In: USENIX 2015 (2015)
19. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: WWW 2010 (2010)

20. Shu, K., Luo, P., Li, W., Yin, P., Tang, L.: Deal or deceit: detecting cheating in distribution channels. In: CIKM 2014 (2014)
21. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: a data mining perspective. ACM SIGKDD Explor. Newslett. **19**(1), 22–36 (2017)
22. Stone, P.J., Dunphy, D.C., Smith, M.S.: The General Inquirer: A Computer Approach To Content Analysis (1966)
23. Tsai, F.S., Chan, K.L.: Detecting cyber security threats in weblogs using probabilistic models. In: Yang, C.C., Zeng, D., Chau, M., Chang, K., Yang, Q., Cheng, X., Wang, J., Wang, F.-Y., Chen, H. (eds.) PAISI 2007. LNCS, vol. 4430, pp. 46–57. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-71549-8_4
24. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: ACL 2005 (2015)

Social Cyber-Security

Kathleen M. Carley[1](✉), Guido Cervone[2], Nitin Agarwal[3], and Huan Liu[4]

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA
kathleen.carley@cs.cmu.edu
[2] Pennsylvania State University, University Park, State College, PA 16801, USA
[3] University of Arkansas Little Rock, Little Rock, AR 72204, USA
[4] Arizona State University, Tempe, AZ 85281, USA

Abstract. Social Cyber-Security is an emerging scientific discipline. Its methodological and scientific foundation, key challenges, and scientific direction are described. The multi-disciplinary nature of this field and its emphasis on dynamic information strategies is considered.

Keywords: Social cyber-security · Network science · Social media analytics

1 The Social Cyber-Security Perspective

Social Cyber-security is an emerging scientific area focused on the science to characterize, understand, and forecast **cyber-mediated** changes in human behavior, social, cultural and political outcomes, and to build the cyber-infrastructure needed for **society** to persist in its essential character in a **cyber-mediated** information environment under changing conditions, actual or imminent social cyber-threats. An example is the technology and theory needed to assess, predict and mitigate instances of influence and community manipulation through alterations in, or control of, the cyber-mediated information environment via bots, cyborgs (combination of bot and human) and humans.

Fundamental to this area is the perspective that we need to maintain and preserve a free and open information environment in which ideas can be exchanged freely, the information source is known, disinformation and false data are identifiable and minimized, and technology is not used to distort public opinion. This relies on the notion that movement of information should not compromise the infrastructure, and that actors should not be able to compromise the cyber-environment so as to unduly influence or manipulate individuals, groups and communities. Types of events to be prevented include viral retweeting of messages containing images which if downloaded release malware, or the use of bots to manipulate groups into accepting fake news as real.

In cyber-security much of the emphasis has been on attacks on and through the cyber-infrastructure aimed at impacting technology, stealing or destroying information, and stealing money or identities [1]. In contrast, in social cyber-security the emphasis is influencing or manipulating individuals, groups or communities and so affecting their behaviors with an emphasis on socio-political-cultural consequences. An example is Russian interference in US elections and spread of fake news after Black Panther movie. While some issues overlap both cyber-security and social cyber-security, the emphasis

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 389–394, 2018.
https://doi.org/10.1007/978-3-319-93372-6_42

is different. Cyber-security focuses on technology and social cyber-security on social context and policy. The research in social cyber-security is not focused on maintaining individual privacy, but at how groups are manipulated and opinions shaped. While phishing is in both areas, for those interested in privacy the goal is to avoid individual data being compromised, whereas the goal for those in social cyber-security is the use of phishing as part of a group-level social influence campaign.

2 Social Cyber-Security as Computational Social Science

Social cyber-security is an inherently multi-disciplinary multi-methodological multi-level computational social science. Emerging theories blend political science, sociology, communication science, organization science, marketing, linguistics, anthropology, forensics, decision science, and social psychology. Key relevant theories are related to persuasion [2], social influence [3], participatory democracy [4], individualized collective action [5], information diffusion [6], manipulation [7], group formation and dissolution [8], identity creation [9], strategic messaging [10], information warfare [11], digital forensics [12] and power [13]. Researchers in this area employ multi-technology computational social science tool chains [14] employing network analysis and visualization [15], language technologies [16], data-mining and statistics [17], spatial analytics [18], and machine learning [19]. Finally, the theoretical results and analytics are often multi-level focusing simultaneously on change at the community and conversation level, change at the individual and group level, and so forth.

Social cyber-security is a computational social science and as such, the approach is noticeably distinct from a pure computer science approach or a pure social science approach. The methods and theories being developed: (a) take the socio-political context into account methodologically and empirically; (b) are predicated on issues of influence, persuasion, manipulation, and theories that link human behavior to behavior in the cyber-mediated environment; and (c) are focused on operational utility rather than just improving scores for machine learning algorithms or theory testing. To illustrate the difference, we consider the issue of disinformation and fake news in Twitter.

A purely computer science machine learning approach would start with a training set containing a set of tweets which had been labelled whether containing fake news or not. This set might be split in two groups, one used to train new algorithms and one used to assess their efficacy. Algorithms would then be devised to empirically categorize tweets as to whether or not, and with what certainty, they contained fake news. The precision and recall of the algorithm would be measured and compared against older algorithms to determine their utility. The goal is prediction; however, the algorithms would have limited utility in context other than that in which they were trained. Data sets are widely shared and reused; but, few relevant social cyber-security data sets exist.

In contrast, a pure social science approach might take a set of tweets in some context, identify through secondary sources which were fake, and then statistically assess differences in the number, content, users etc., using the analysis to test a theory about fake news that is predicated on human social behavior but ignores the role of the technology. Data reuse is often confined to the research group and rare for qualitative data.

Qualitative or quantitative support for theories determines their utility. The goal is explanation; however, those explanations are often nuanced to specific socio-cultural settings.

A social cyber-security approach considers both how the technology can be employed to impact: (1) messaging – i.e., who gets what messages when, presentation and access; and (2) group formation – i.e., who communicates with whom when, influence, and group and actor identification. Complex network analytics, visualization, statistics and text mining are used to create empirical profiles of messages that do and don't contain fake news, users that do and don't send the messages, and users who are or are not receptive. New methods are often tested on both new and old data. Method and theory are co-developed, reusable, and extensible to new domains. Their utility resides their ability to support explanation, and prediction in the wild (Table 1).

Table 1. Contrasting approaches to fake news.

Characteristics	Computer science	Social science	Social cyber-security
Operationally focused	No	No	Yes
Data reuse	High	Low	Medium
Utility based on	Precision and recall	Theory development and validation	Operational value assessment and prediction value
Tests theory about human behavior	No	Yes	Yes
Empirically driven	Yes	Sometimes	Yes
Considers:			
Socio-political context	No	Yes	Yes
Media's features	Minimally	Minimally	Strongly
Adversarial actions	No	Sometimes	Yes
Social influence	No	Yes	Yes
Individuals & groups	No	Sometimes	Yes
Classes of users	Sometimes	Sometimes	Sometimes

3 What Are Key Challenges to Doing Research in Social Cyber-Security?

The rapid rate of change in cyber-technologies, evolving legal and policy constraints, and rapid global information flow are creating an environment in which technical, policy and economic issues are strongly impacting what science can be done, what science needs to be done, how that science can be done, and what is required for those who can do that science.

A key challenge is data control. Data are held by and controlled by a few providers who restrict who, how, when and what can be accessed, as well as how, or if, the data are maintained. While data access is always problematic, the degree of external management, volume of data and pervasiveness of controlled data is unprecedented. While

Twitter is only a small portion of the digital landscape it is like a canary in a mine, the early indicator of evolving trends in cyber-space. Unlike other platforms, Twitter is more science friendly due to public tweets falling under the creative commons license and therefore being open and free data that can be harvested for automated analysis. Many scientific papers in the social cyber-security area have focused on Twitter.

Twitter data are not, however, as open as it might seem. There are three dominant ways to access this data: (1) use one of the two Twitter APIs, (2) gain access from Twitter to the 10% feed, and (3) buy Tweets from one of the intermediaries who have access to the 100% feed and historical data. The Twitter APIs provide access to only some of the meta-data around the Tweet, focus on more recent Tweets, and the quality of the sample depends on whether bounding boxes or search terms are used [20]. Further, the samples are biased [21]. Gaining access to the 10% feed typically requires getting one of the few Twitter grants or buying data. The 1% API and the 10% feed are not a random sample of all Twitter data given the search criteria; however, the biases are not well known. Buying the data is extremely costly, but can give you some historical data. Intermediates who provide Twitter data are expected to continuously clean the data and remove recalled Tweets and those by suspended users. They also cannot provide the full meta-data which can reduce the ability to link data sets. Further, these companies may "enhance" the data by adding their determination of language, location or whether the Tweeter is a bot – without explaining how this was determined. Consequently, basic research is needed on bias estimation, impact of missing data, and learning from irreproducible results as the data needed for reproduction may have been deleted.

On the policy side, policies and laws are out-of-sync with the new technologies. Importantly, the rate of change in the technology is such that new forms of illicit activity are emerging at an unprecedented rate. Policies designed to impede, punish or otherwise curtail such activities lag behind the technology. Many policy and law makers have minimal understanding of the technology and so design policy and law that are often irrelevant, or unenforceable, or so restrictive that they prevent the science from being done that would inhibit or detect early social cyber-attacks. Illustrative areas are organizational security, privacy versus detection, and global policies.

Organizations are at risk from social cyber-security attacks. Phony Facebook updates, malware embedded in tweeted image, phishing etc., create organizational insecurities ranging from brand manipulation to compromising personnel to get access to intelligence to destruction of data or machines from social media delivered malware[1]. A 2016 report argued that one in five organizations suffers from a malware attack via social media[2]. The cyber-environment creates yet another risk, in that data-mining coupled with massive on-line data opens the door to corporate secrets being discovered simply by assessing corporate activity including purchasing, personnel hiring, changes in board of directors and so on. Organizations are responding by creating various social

[1] https://www.infosecurity-magazine.com/blogs/top-10-worst-social-media-cyber/.

[2] https://www.pandasecurity.com/mediacenter/social-media/uh-oh-one-out-of-five-businesses-are-infected-by-malware-through-social-media/.

cyber-security policies such as restricting access to the internet from work, using institutional settings on platforms such as email and dropbox, and increased social cyber-security training general cyber-security training. Drawing from the lessons learned in the nuclear industry, effective organizational policies need to be concerned with heedful interaction, and creating a social cyber-security awareness.

4 Summary

Social cyber-security is an emerging scientific area concerned with social influence and group manipulation. An estimate of the number of articles based on a snowball from key words in the area and removing those focused exclusively on machine learning algorithms, privacy, or using only a social science approach reveals an exponential growth - see Fig. 1. New research is needed in many areas including bias estimation and reduction in data; movement of actors and ideas within and between media; semi-automated identification, assessment of impact of, and effectiveness of counter-messaging for different forms of information strategies; approaches to inoculate individuals and groups against disinformation and effectiveness of those strategies. Future research in this new scientific area is needed to shape the social cyber-environment and promote social cyber-security.

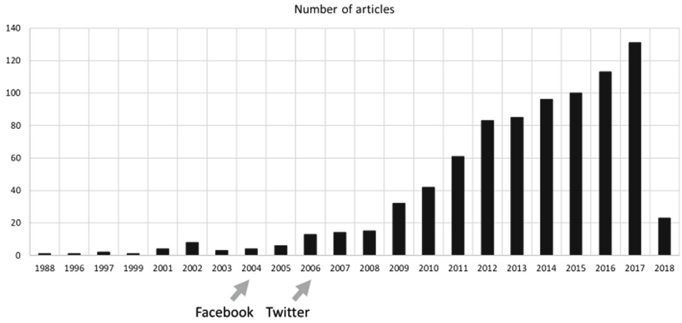


Fig. 1. Number of articles in social cyber-security by year

References

1. Reveron, D.S.: Cyberspace and National Security: Threats, Opportunities, and Power in A Virtual World. Georgetown University Press, Washington D.C. (2012)
2. Gass, R.H., Seiter, J.S.: Persuasion: Social Influence and Compliance Gaining. Routledge, UK (2015)
3. Benigni, M., Joseph, K., Carley, K.M.: Online extremism and the communities that sustain it: detecting the ISIS supporting community on Twitter. PLoS ONE **12**(12), e0181405 (2017)
4. Sunstein, C.R.: #Republic: Divided democracy in the age of social media. Princeton University Press, Princeton (2018)
5. Bennett, W.L.: The personalization of politics: political identity, social media, and changing patterns of participation. Ann. Am. Acad. Polit. Soc. Sci. **644**(1), 20–39 (2012)

6. Wu, L., Liu, H.: Tracing fake-news footprints: characterizing social media messages by how they propagate. In: The Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM2018), pp. 637–645. ACM, NY (2018)
7. Colliander, J., Dahlén, M.: Following the fashionable friend: the power of social media: weighing publicity effectiveness of blogs versus online magazines. J. Advert. Res. **51**(1), 313–320 (2011)
8. Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X.: Group formation in large social networks: membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 44–54. ACM, NY (2006)
9. Joseph, K., Wei, W., Benigni, M., Carley, K.M.: A social-event based approach to sentiment analysis of identities and behaviors in text. J. Math. Sociol. **40**(3), 137–166 (2016)
10. Benigni, M., Joseph, K., Carley, K.M.: Mining online communities to inform strategic messaging: practical methods to identify community-level insights. Comput. Math. Organ. Theor. **24**, 224–242 (2017)
11. Cordesman, A.H., Cordesman, J.G.: Cyber-Threats, Information Warfare, and Critical Infrastructure Protection: Defending The US Homeland. Greenwood Publishing Group, Westport (2002)
12. Al-khateeb, S., Hussain, M.N., Agarwal, N.: Social cyber forensics approach to study twitter's and blogs' influence on propaganda campaigns. In: Lee, D., Lin, Y.-R., Osgood, N., Thomson, R. (eds.) SBP-BRiMS 2017. LNCS, vol. 10354, pp. 108–113. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60240-0_13
13. Entman, R.M.: Framing bias: media in the distribution of power. J. Commun. **57**(1), 163–173 (2007)
14. Benigni, M., Carley, K.M.: From tweets to intelligence: understanding the islamic jihad supporting community on Twitter. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) SBP-BRiMS 2016. Lecture Notes in Computer Science, vol. 9708, pp. 346–355. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-39931-7_33
15. Carley, K.M., Wei, W., Joseph, K.: High dimensional network analytics: mapping topic networks in twitter data during the Arab spring. In: Cui, S., Hero, A., Luo, Z.-Q., Moura, J. (eds.) Big Data Over Networks. Cambridge University Press, Boston (2016)
16. Hu, X., Liu, H.: Text Analytics in Social Media. In: Aggarwal, C., Zhai, C. (eds.) Mining text data, pp. 385–414. Springer, Boston (2012). https://doi.org/10.1007/978-1-4614-3223-4_12
17. Agarwal, N., Kumar, S., Gao, H., Zafarani, R., Liu, H.: Analyzing behavior of the influentials across social media. In: Cao, L., Yu, P. (eds.) Behavior Computing, pp. 3–19. Springer, London (2012). https://doi.org/10.1007/978-1-4471-2969-1_1
18. Cervone, G., Sava, E., Huang, Q., Schnebele, E., Harrison, J., Waters, N.: Using Twitter for tasking remote-sensing data collection and damage assessment: 2013 Boulder flood case study. Int. J. Remote Sens. **37**(1), 100–124 (2016)
19. Wei, W., Joseph, K., Liu, H., Carley, K.M.: Exploring characteristics of suspended users and network stability on Twitter. Soc. Netw. Anal. Min. **6**(1), 51 (2016)
20. Carley, K.M., Momin, M., Landwehr, P.M., Pfeffer, J., Kowalchuck, M.: Crowd sourcing disaster management: the complex nature of Twitter usage in Padang Indonesia. Saf. Sci. **90**, 48–61 (2016)
21. Morstatter, F., Pfeffer, J., Liu, H. Carley, K.M.: Is the sample good enough? Comparing data (2013)

A Computational Model of Cyber Situational Awareness

Geoffrey B. Dobson and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA 15213, USA
{gdobson,kathleen.carley}@cs.cmu.edu

Abstract. A computational model of cyber situational awareness is built using the Cyber-FIT agent-based modeling and simulation framework. This work expands the framework by adding a computational cognitive model of the agents' perception of cyber situational awareness. Virtual experiments are conducted to test the model, and determine how long it may take for a military cyber team to gain cyber situational awareness.

Keywords: Cyber situational awareness · Agent-based modeling
Cyber behavior · Military

1 Introduction

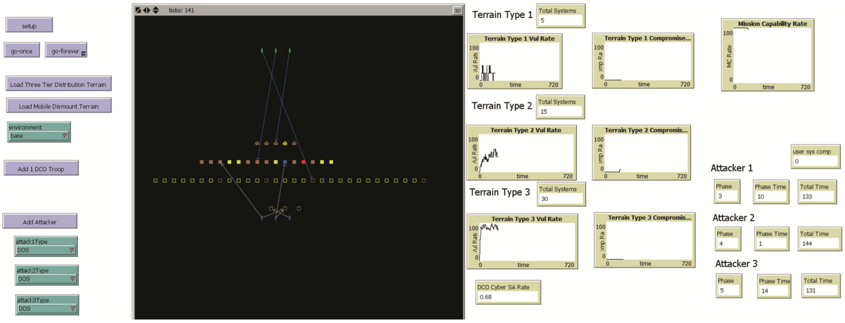
Military leaders are keenly aware of the pressing need for improved cyber situational awareness capabilities. In recent testimony to the U.S. Senate Armed Services Committee [3], Navy Vice Admiral Michael Gildaysaid: "We've extended our defensive posture to include deploying defensive cyber teams with our carrier strike groups and our amphibious readiness groups". This means that an ever growing number of military operations will have a defensive cyber force attached. When defensive cyber forces fall into an area of responsibility, they must first conduct a survey of the cyber terrain, like an infantry unit would survey the land terrain, or an air controller would examine the air space. Militaries have been conducting land terrain surveys for thousands of years, but cyber terrain surveys for less than a decade. In this paper, we simulate a cyber terrain survey using the Cyber-FIT agent-based modeling and simulation framework [2]. We show that given several realistic behaviors and constraints, the defensive cyber force can conduct a full survey in approximately two hours, but full cyber situational awareness is impossible.

2 Background

The purpose of surveying cyber terrain, whether on a corporate network, or military mission, is to gain understanding of the states of the various systems under the team's purview. Put another way, it is to gain "cyber situational awareness". There are many definitions of cyber situational awareness. Onwubiko [4] defines cyber situational awareness as "processes and technology required to gain awareness of historic, current,

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 395–400, 2018.
https://doi.org/10.1007/978-3-319-93372-6_43

and impending (future) situations in cyber”. In this paper we are modelling the knowledge of the current situation that the defender has realized. Similarly, Barford et al. [1] describe seven aspects of cyber situational awareness. The first is “Be aware of the situation. This aspect can also be called situation perception”. Our simulation software defines the perception of the agents’ knowledge of the terrain as a table of system states. This gives a computational model of the cognitive representation of cyber situational awareness for each agent, and cumulatively, for the team. By defining cyber situational awareness in this manner, we can observe the changes over time, determine what factors most affect it, and more clearly understand what the appropriate definition of cyber situational awareness is, in a given scenario (Fig. 1).



ticks: 141

environment: base

attack1Type: DOS

attack2Type: DOS

attack3Type: DOS

Terrain Type 1 — Total Systems: 5

Terrain Type 2 — Total Systems: 15

Terrain Type 3 — Total Systems: 30

DCO Cyber SA Rate: 0.68

user sys comp: 0

	Phase	Phase Time	Total Time
Attacker 1	3	10	133
Attacker 2	4	1	144
Attacker 3	5	14	131

Fig. 1. Screenshot of the Cyber-FIT dashboard

3 The Model

This work expands on previous work by Dobson and Carley [2] proposing the Cyber-FIT simulation framework. The framework defines two agent types: forces and terrain. The force agents interact with terrain agents by defending (defensive agents) or attacking (offensive agents). Terrain is one of three types: networking, servers, and client systems. The terrain becomes vulnerable over time, if not defended. Offensive agents can attack terrain in one of three ways: routing protocol attack, denial of service attack, or phishing attack. Offensive agents move through the cyber kill chain in order to conduct attacks, with behavior based on empirical observations by Rege et al. [5]. In this updated version of the framework, we added a cognitive model of cyber situational awareness. That is, at every time tick (one simulated minute), the agents store the value of the state of the system they are interacting with. This models their cognitive understanding of the cyber terrain. The team’s cyber situational awareness is the sum of their cognitive models. At time 0, they have no cyber situational awareness. As time goes on, they build and update a table of terrain states. The states are one of three: not vulnerable, vulnerable, and compromised. This is compared against the true state of the systems at every minute to give the team Cyber Situational Awareness (CSA). This is computationally defined as:

$$\text{CSA} = \frac{\sum \text{Defender1Correct States} + \sum \text{Defender2Correct States} + \sum \text{Defender3Correct States}}{\text{Number of Cyber Terrain}}$$

4 Virtual Experiments

We conducted two virtual experiments using this model. In each experiment, we hold the attack type, number of agents, vulnerability growth rate, exploit success rate, and defensive action success rate all constant. Those variables can be altered to explore the response of the model, but is not necessary for these two experiments. In these two experiments we are examining how successful the terrain survey is, as defined by team level cyber situational awareness and how quickly the agents can survey (Table 1).

Table 1. Independent and dependent variables in the virtual experiments

Independent variables		
IV	Variants	Specific values
Attack type	1	DOS
DCO forces	1	3
Agent success rate	2	10, 50
Dependent variables		
DV	Variable type	
Cyber SA rate	Continuous	
Time to survey	Integer	

4.1 What Is the Maximum Cyber Situational Awareness During a Cyber Terrain Survey?

This experiment simulates a defensive cyber force falling into contested cyber terrain under active attack. Three defensive agents survey and defend the terrain, while three offensive agents attack the terrain. The goal of this experiment is to determine how successful the survey is, and how much time should surpass until the performance levels off.

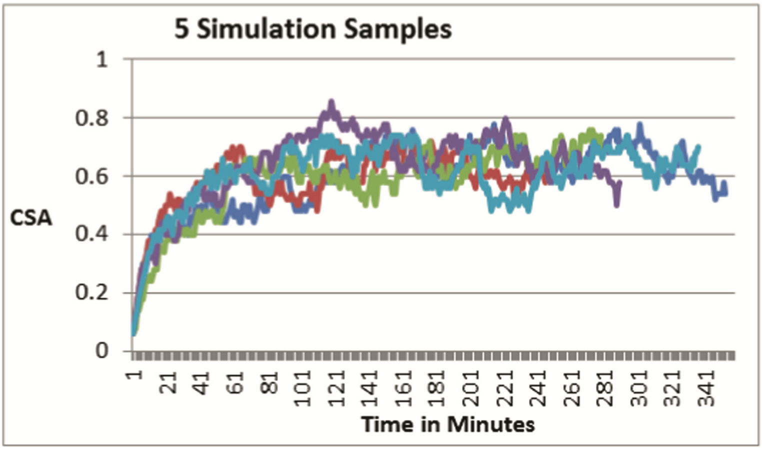


Fig. 2. A sample of five simulations to show the variance in cyber situational awareness

As shown in both Figs. 2 and 3, the performance levels off after 100 min, but there is still a fair amount of variance between runs of the experiment. After 100 min, the minimum CSA observed was 0.40, the maximum was 0.86, and the average was 0.64.

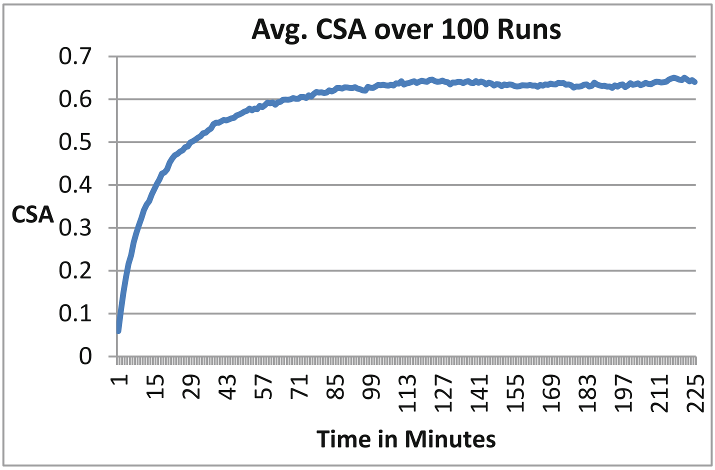


Fig. 3. The average cyber situational awareness across all 100 runs of the experiment

4.2 How Long Does It Take to Conduct a Full Survey and What Is the CSA at that Time?

In this experiment, we are running the simulation until the agents' cognitive model of cyber situational awareness covers all 50 cyber terrain points (Fig. 4).

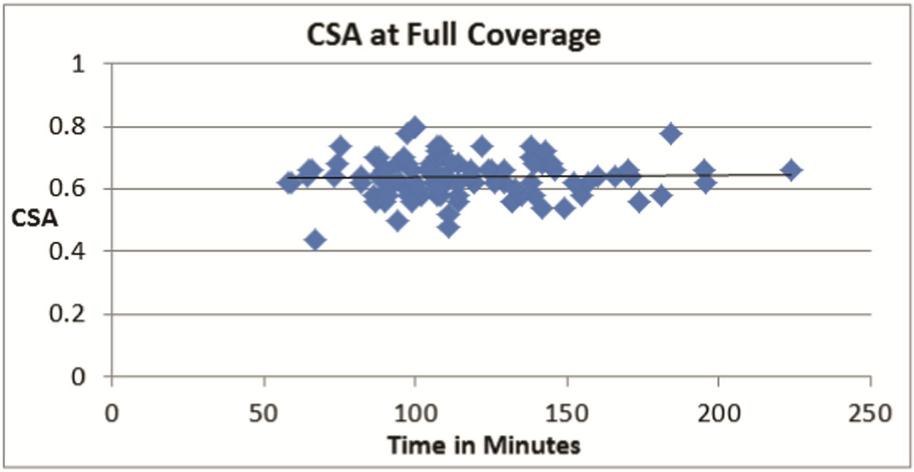


Fig. 4. Scatter plot shows no improvement in CSA when agents take longer to conduct a full terrain survey

In this experiment, we observed that the average time to complete the full survey (all 50 cyber terrain endpoints inspected) was 115.81 min and average CSA at that point was 0.64.

4.3 Virtual Experiment Discussion

The key finding of experiment 1 was that over 100 runs, the maximum CSA observed was 0.86. This is expected because agents can only inspect one piece of terrain per minute. Like in real life, as one system is being inspected, other systems may become vulnerable or compromised. Military leaders must decide how many resources to apply to a cyber terrain survey, and how much risk they will accept, given the fact that 100% cyber situational awareness is impossible. Also, when vulnerabilities are found, which should be immediately elevated, which should be immediately fixed, and which can be left to fix later? Also in experiment 1, after minute mark 100, at any given time, the CSA ranged from 0.40 to 0.86. This is a fairly large performance gap. Cyber forces should consider defining what routines and processes increase the likelihood of higher cyber situational awareness.

In experiment 2, we found that the average time to conduct a full survey is 115 min. This is based on the agents randomly selecting terrain, and switching every minute. In an operational mission, military leaders should develop detailed cyber terrain survey plans, with clear reporting instructions. This will ensure that survey missions are repeatable and measurable. Also, careful attention should be paid to the order in which terrain is surveyed. In this simulation, order of operations does not matter, which would not be the case in an operational environment.

5 Conclusion

In this paper we proposed an approach to computationally defining team cyber situational awareness as an accumulation of the agents' cognitive model of the state of cyber terrain, compared with the true state of the cyber terrain. We conducted two virtual experiments to assess the assumptions of the model and reason about the applicability of the findings. This work is part an ongoing effort to improve the state of the art of military based cyber force package modelling and simulation by Carnegie Mellon University's Computational Analysis of Societal and Organizational Systems (CASOS) Center. In future work we plan to simulate the passing of messages between agents, in order to share cyber situation awareness, and collectively act upon cyber terrain vulnerabilities and threats. Also, we'll create simulations where more realistic constraints are applied, which will force the simulated commander to make resource trade off decisions.

References

1. Barford, P., et al.: Cyber SA: situational awareness for cyber defense. In: Jajodia, S., Liu, P., Swarup, V., Wang, C. (eds.) Cyber Situational Awareness. ADIS, vol. 46, pp. 3–13. Springer, Boston (2010). https://doi.org/10.1007/978-1-4419-0140-8_1
2. Dobson, G.B., Carley, K.M.: Cyber-FIT: an agent-based modelling approach to simulating cyber warfare. In: Lee, D., Lin, Y.-R., Osgood, N., Thomson, R. (eds.) SBP-BRiMS 2017. LNCS, vol. 10354, pp. 139–148. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60240-0_18
3. Martin, A.: Military Leaders Highlight Progress in Cyber Domain during U.S. Senate Hearing. March 16, 2018. https://homelandprepnews.com/stories/27332-military-leaders-highlight-progress-cyber-domain-u-s-senate-hearing/. Accessed 19 March 2018
4. Onwubiko, C.: Understanding cyber situation awareness. Int. J. Cyber (2016)
5. Rege, A., Parker, E., Singer, B., Masceri, N.: A qualitative exploration of adversarial adaptability, group dynamics, and cyber-intrusion chains. J. Inf. Warfare **16**(3), 1–16 (2017)

Assessment of Group Dynamics During Cyber Crime Through Temporal Network Topology

Nima Asadi[1(✉)], Aunshul Rege[2], and Zoran Obradovic[1]

[1] Computer and Information Sciences Department, Temple University, Philadelphia, USA
nima.asadi@temple.edu

[2] Department of Criminal Justice, Temple University, Philadelphia, USA

Abstract. Understanding group dynamics can provide valuable insight into how the adversaries progress through cyberattacks and adapt to any disruptions they encounter. However, capturing the characteristics of such dynamics is a difficult task due to complexities in the formation and focus of the adversarial team throughout the attack. In this study, we propose an approach based on concepts and measures of social network theory. The results of experiments performed on observations at the US Industrial Control Systems Computer Emergency Response Team's (ICS-CERT) Red Team-Blue Team cybersecurity training exercise held at Idaho National Laboratory (INL) show that the team dynamics can be captured and characterized using the proposed approach. Moreover, we provide an analysis of the shifts in such dynamics due to the adversarial team's adaptation to disruptions caused by the defenders.

Keywords: Network theory · Group dynamics · Machine learning

1 Introduction

Governments and organizations worldwide are experiencing a continuously evolving threat landscape, where cyberadversaries are highly organized, sophisticated, and persistent. Defenders can only be effective if they understand how adversaries organize, make decisions, carry out attacks, and adapt to disruptions. Earlier research has examined adversarial attack paths also known as intrusion chains, time spent on the various stages of cyberattacks, and which stages adversaries focus on more when they are disrupted by defenders span [1–4].

However, little is known in the open literature about adversarial group dynamics. It is imperative to study how adversaries interact, structure themselves, change over the duration of the attack, manage disruptions by defenders, recover from their mistakes, and make decisions as they progress through cyberattacks in real-time.

R. Thomson et al. (Eds.): SBP-BRiMS 2018, LNCS 10899, pp. 401–407, 2018.
https://doi.org/10.1007/978-3-319-93372-6_44

2 Methodology

2.1 Case Studies

The dataset for our first case study was collected at a five day cybersecurity training organized by the United States Industrial Control Systems Computer Emergency Response Team (ICS-CERT) and hosted by Idaho National Laboratory (INL) in September/October 2014. The training included a Red Team/Blue Team exercise (RTBTE), where the Red team operated as the adversarial team. The Red Team consisted of ten members who had a mixed set of skills. The data for the second case study was collected at a one-day student cybersecurity competition where a team including 7 members was observed and interviewed. The data used for the case studies in this paper included time stamped observations of the Red Teams in both of the mentioned exercises.

2.2 Construction of the Temporal Network

Capturing the characteristics and patterns in the adversarial team's formation during the cyber intrusion can helps us gain important knowledge about the decision making, task scheduling and planning its process. Here we propose a methodology for capturing and analyzing such information by using concepts and measurements of network science. In order to perform such analysis, we first create the temporal network of the adversarial team based on the commonalities in activities of the team members during each time point. In other words, if, team members A and B perform the same intrusion chain stage during the time point t, a link is drawn between them in the network. For this purpose, we use the intrusion chain model proposed by [3]. Therefore, at each time point, the team members are the nodes of the network, and the links (edges) between them indicates that the nodes have been performing similar intrusion stage during that specific time point. Each time point for our case study spans for 15 min. Therefore, this criteria generates T different networks where T is the number of time points.

After creating the team network for each time point, we are able to take advantage of several informative measures for capturing and analysis of team dynamics. In the next section, we discuss our proposed measures for the analysis of team dynamics using the constructed temporal networks.

2.3 Analytical Measures

Number of Connected Components. Number of connected components is an important topological invariant of a graph [6]. In this study, a high number of connected components in a graph shows that a majority of team members work individually on non-similar intrusion stages, while a lower number of connected components shows that more members work together on similar intrusion stages. In other words, the number of connected components is an indicator of the level of connectivity and cooperation.

Edge Density. Density of the edges in the network shows the level of overall connection in the network. This measure is defined as the number of connections a node has, divided by the total possible connections a node can have [6].

Transitivity. Transitivity is the overall probability of the existence of tightly connected communities or cliques. This measure is calculated as the transitivity is the ratio of triangles to triplets in the network.

Average Shortest Path Length. Average shortest path length in a graph is calculated as the average number of stops needed to reach two distant nodes in the graph. The smaller the result, the more efficient the network in information circulation.

Average Node Degree. Average node degree is simply calculated by averaging the degrees of all of the nodes in the graph.

Modularity. Modularity quantifies the degree to which the network may be subdivided into clearly delineated groups.

After deriving the listed network characteristics, they are used to form the feature vector for detection of possible anomalies in the adversarial movement. In order to make such prediction, we train an algorithm for binary classification where the labels indicate if the condition is normal, or an anomaly is taking place, i.e. a disruption is happening. Sources of disruptions can be the Blue team or the Red team's own failures. We used support vector machine (SVM) and logistic regression as the classifiers for this study.

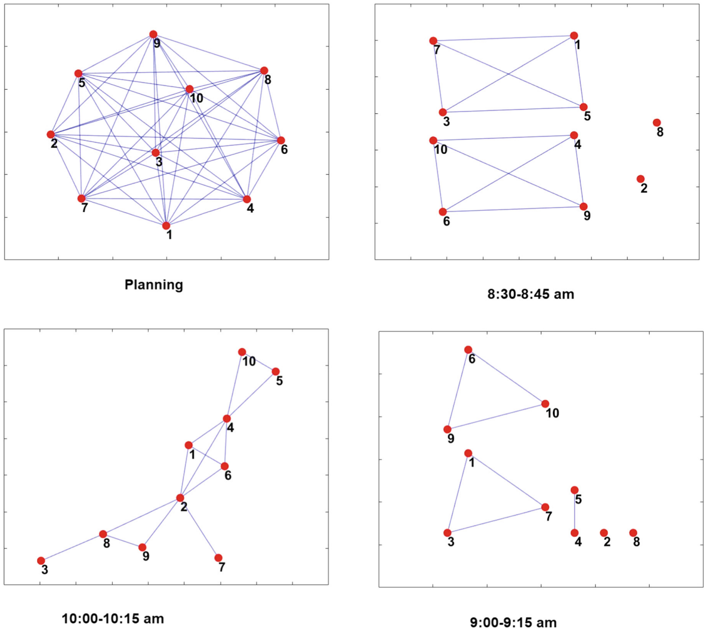
3 Results

3.1 Team Dynamics Characteristics

The team dynamics networks were created for each time point according to the descriptions provided in the previous section. The duration of the first and second RTBTE sessions were 9 and 6 hours, respectively. An example of the network created at four different time points for the first case study is provided in Fig. 1. In that figure, the top left figure indicates the graph at the very beginning of the exercise where we can observe a complete graph (each node is connected to every other node). This is due to the fact that in the team spent the very early phase of the exercise discussing the plans, meaning that the entire team was involved in one task. The three other plots display the Red Team's formation during three different periods of the exercise. For instance, in the top right figure, we can observe that team members 2 and 8 were working individually on separate intrusion chain stages while two other groups, each including four members, were involved in different intrusion chain stages.

3.2 Network Analysis Results

A plot of number of connected components and edge density for the first case study are provided in Fig. 2. An observation one can make based on that figure is



Planning

8:30-8:45 am

10:00-10:15 am

9:00-9:15 am

Fig. 1. Example graphs constructed from the case study data at four different points of the exercise.

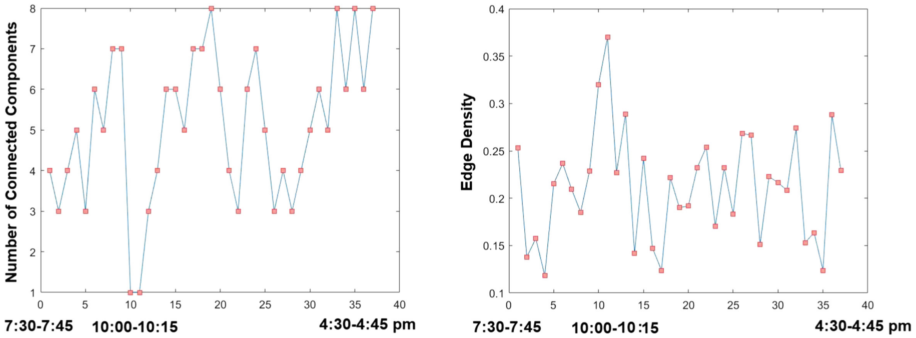


Fig. 2. Analytical results of team dynamics based on constructed temporal networks. Left: the number of connected components at each time point. Right: the edge density of the networks at each time point.

the existing anomaly in both number of connected components and edge density plots during the time point 10:00 am to 10:15 am. This can be associated to the fact that two disruptions were observed at the case study one at that time. As we can observe in Fig. 2, during and after occurrence of a disruption, the number of connected components decreased to one while the edge density was increased to above 0.3. One can interpret this decrease of the number of connected components and increase in edge density as the immediate increase in the entire Red team's focus on a few certain intrusion stages. This observation can be expanded to other network network measures as well. For anomaly detection We used the data from case study one as the training sample, and case study two as the test sample. The reason for using different case studies as the train and test datasets is to ensure the generalizability of the model. Note that each data point in our prediction is a time point at the cyber security training. The prediction results are provided as the area under the curve (AUC) in Fig. 3. We can observe that AUC of 0.782 and 0.735 were achieved using logistic regression and SVM, respectively.

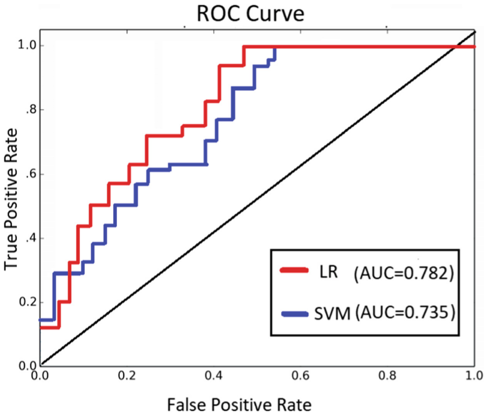


Fig. 3. Area under the curve (AUC) for anomaly prediction through characteristics of team dynamic network. LR stands for logistic regression, and SVM denotes support vector machine.

4 Conclusion

Certain limitations with this study, such as lack of generalizability are inevitable. However, the authors argue that this work intends to lay the framework for further research in the area. Moreover, the case study in this paper is based on two case studies including one of the most reputable force on force ("paintball") exercises in the United States.

The proposed network analysis offers some interesting findings about the adversarial team dynamics:

The Team Dynamics Networks Usually Contains More than One Connected Component

Except from the two time periods after the disruptions occurred, the number of connected components remained above two. This indicates that usually the adversarial subgroups perform multiple intrusion stages in parallel.

The Edge Density is Usually Low Throughout the Exercise

Except the time span when the disruptions took place, the edge density of the constructed networks was below 0.3. This further indicates the sparse and parallel performance of the subgroups of the Red Teams rather than being highly connected and focused on few intrusion stages together.

Disruptions Can Affect Team Dynamics

Topological characteristics of the team dynamic networks show deviation during disruptions. For instance, the decrease in connected components and the increase in edge density can be interpreted as a change in team dynamics towards more focus on certain intrusion stages with higher connection among team members. The results of anomaly detection using the machine learning algorithms further prove the effect of disruptions on team dynamics.

This paper offered a preliminary analysis of adversarial group dynamics during a real-time cybersecurity exercise. Future research, however, should delve deeper into other aspects of groups, such as the influence and interaction in groups, performance and functionality, divisions of labor, and subgroup decision-making and autonomy.

Acknowledgements. This material is based upon work supported by the National Science Foundation CAREER Award, Grant No. CNS1453040 and partially by National Science Foundation CPS Award, Grant No. 1446574. The authors thank the Industrial Control Systems Computer Emergency Response Team (ICSCERT) and Idaho National Laboratory (INL) for allowing data collection at their September/October 2014 Red Team/Blue Team Cybersecurity Training Exercise.

References

1. Rege, A., Obradovic, Z., Asadi, N., Singer, B., Masceri, N.: A temporal assessment of cyber intrusion chains using multidisciplinary frameworks and methodologies. In: 2017 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), pp. 1–7. IEEE, June 2017
2. Rege, A., Obradovic, Z., Asadi, N., Parker, E., Masceri, N., Singer, B., Pandit, R.: Using a real-time cybersecurity exercise case study to understand temporal characteristics of cyberattacks. In: Lee, D., Lin, Y.-R., Osgood, N., Thomson, R. (eds.) SBP-BRiMS 2017. LNCS, vol. 10354, pp. 127–132. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60240-0_16
3. Cloppert, M.: Security Intelligence: Attacking the Cyber Kill Chain (2009). http://digital-forensics.sans.org/blog/2009/10/14/security-intelligence-attacking-the-kill-chain. Accessed 2 Feb 2014
4. Colbaugh, R., Glass, K.: Proactive Defense for Evolving Cyber Threats. Sandia National Laboratories [SAND2012-10177] (2012). https://fas.org/irp/eprint/proactive.pdf. Accessed 15 Feb 2017

5. Leclerc, B.: Crime scripts. In: Wortley, R., Townsley, M. (eds.) Environmental Criminology and Crime Analysis. Routledge, Abingdon (2016)
6. Krause, J., Croft, D.P., James, R.: Social network theory in the behavioural sciences: potential applications. Behav. Ecol. Sociobiol. **62**(1), 15–27 (2007)
7. Rokach, L., Maimon, O.: Clustering methods. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, Boston (2005). https://doi.org/10.1007/0-387-25465-X_15
8. Schneider, R.: Survey of peaks/valleys identification in time series. Department of Informatics, University of Zurich, Switzerland (2011)
9. Ellens, W., Kooij, R.E.: Graph measures and network robustness. arXiv preprint arXiv:1311.5064 (2013)

Author Index