Dongwon Lee · Yu-Ru Lin
Nathaniel Osgood · Robert Thomson (Eds.)

LNCS 10354

Social, Cultural, and Behavioral Modeling

10th International Conference, SBP-BRiMS 2017
Washington, DC, USA, July 5–8, 2017
Proceedings

Springer

Lecture Notes in Computer Science 10354

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

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

Dongwon Lee · Yu-Ru Lin
Nathaniel Osgood · Robert Thomson (Eds.)

Social, Cultural, and Behavioral Modeling

10th International Conference, SBP-BRiMS 2017
Washington, DC, USA, July 5–8, 2017
Proceedings

Springer

Editors
Dongwon Lee
Penn State University
State College, PA
USA

Yu-Ru Lin
University of Pittsburgh
Pittsburgh, PA
USA

Nathaniel Osgood
University of Saskatchewan
Saskatoon, SK
Canada

Robert Thomson
United States Military Academy
West Point, NY
USA

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-60239-4 ISBN 978-3-319-60240-0 (eBook)
DOI 10.1007/978-3-319-60240-0

Library of Congress Control Number: 2017943044

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

Printed on acid-free paper

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

Preface

Improving the human condition requires understanding, forecasting, and impacting socio-cultural behavior both in the digital and non-digital 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 developing countries etc. 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 social science or only on computer science are doomed to failure. For example, Anonymous developed an approach for identifying members of terror groups such as ISIS on the social media platform Twitter using state-of-the-art computational techniques. These accounts were then suspended. This was a purely technical solution. The response was 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 goal and 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 empirically 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 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 both 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 are available for 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 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 just 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 2017 conference, you will see suggestions of how to address the challenges just described. SBP-BRiMS 2017 continued in 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 79 papers were submitted as regular track submissions. Of these, 16 were accepted as full papers for an acceptance rate of 20.2% and 27 were accepted as short papers for an acceptance rate of 42.8%.

Additionally, there were a large number of papers describing emergent ideas, late-breaking results, or responses to the challenge problem were submitted and accepted. Finally, there were nine tutorials covering a diversity of topics. This is an international group with papers submitted by authors from 15 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 including, but not limited to, machine learning, language technology, social network analysis and visualization, agent-based simulation, and statistics. Based on the author's self-selected area for each paper, the breakdown is as follows:

- Behavior and social sciences: 20 submissions, 11 accepted
- Health sciences: 5 submissions, 3 accepted
- Information, systems, and network sciences: 44 submissions, 22 accepted
- Methodology: 3 submissions, 3 accepted
- Military and intelligence applications: 7 submissions, 4 accepted

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 Committe, and the Scholarship Committee, as well as 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 Office of Naval Research – N00014-15-1-2463, N00014-16-1-2274 and N00014-17-1-2461, the National Science Foundation – IIS-1523458, and the Army Research Office – W911NF-17-1-0138. Enjoy the conference proceedings.

April 2017

Kathleen M. Carley
Nitin Agarwal

Organization

Conference Co-chairs

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

Program Co-chairs

Dongwon Lee	Penn State University, USA
Nathaniel Osgood	University of Saskatchewan, Canada
Robert Thomson	United States Military Academy
Kevin S. Xu	University of Toledo, USA
Yu-Ru Lin	University of Pittsburgh, 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 at 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

Therese L. Williams University of Central Oklahoma, USA

Local Area Coordination

David Broniatowski The George Washington University, USA

Proceedings 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 at Little Rock, USA
Sumeet Kumar Massachusetts Institute of Technology, USA
Brandon Oselio University of Michigan, USA
Justin Sampson Arizona State University, USA

Topic Area Chairs

Halil Bisgin University of Michigan-Flint, USA
Shahryar Minhas Duke University, USA
Daniel Cassenti US Army Research Laboratory, USA
Konstantinos Pelechrinis University of Pittsburgh, USA
Ayaz Hyder The Ohio State University, USA
Elizabeth Mezzacappa US Army Target Behavioral Response Laboratory, 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 at 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, USA
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, USA
Coty Gonzalez	Carnegie Mellon University, USA
Brian Gore	NASA, USA
Kristen Greene	National Institute of Standards and Technology, USA
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 Bharat Patel	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 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
Halil Bisgin
Lashon Booker
David Bracewell
David Broniatowski
Magdalena Bugajska
Jose Cadena
Brad Cain
Ernesto Carrella
Subhadeep Chakraborty
Fahmida Chowdhury
David Clark
Rachel Cummings
Peng Dai
Hasan Davulcu
Yves-Alexandre de Montjoye
Robert Demarco
Jana Diesner
Koji Eguchi
Bruno Emond
William Ferng
Michael Fire
Elizabeth Ginexi
Ariel Greenberg
Kristen Greene
Kyungsik Han
Walter Hill
Shen-Shyang Ho
Shuyuan Mary Ho
Tuan-Anh Hoang
Yuheng Hu
Robert Hubal
Terresa Jackson
Nasir Jaffery
John Johnson
Kenneth Joseph
Byeong-Ho Kang
Bill Kennedy
Masahiro Kimura
Shamanth Kumar
Kiran Lakkaraju
Othalia Larue
Trisha Lawrence
Christian Lebiere
Jongwuk Lee
Dongwon Lee
Jiexun Li
Yu-Ru Lin
Huan Liu
Lyle Long
Deryle W. Lonsdale
Andreas Luedtke
Jiebo Luo
Jonathas Magalhães
Juan F. Mancilla-Caceres
Stephen Marcus
Venkata Swamy Martha
Allen Mclean
Shahryar Minhas
Sai Moturu
Keisuke Nakao

Radoslaw Nielek
Kouzou Ohara
Byung Won On
Brandon Oselio
Nathaniel Osgood
Alexander Outkin
Konstantinos Pelechrinis
Hemant Purohit
Weicheng Qian
S.S. Ravi
Gladstone Reid
Amit Saha
Philip Schrodt
Samira Shaikh
Narjes Shojaati
Shade Shutters
Barry Silverma
Amy Sliva
David Stracuzzi
Yizhou Sun
Samarth Swarup
George Tadda
Robert Thomson
Robert P. Trevino
Melissa Walwanis
Xiaofeng Wang
Zhijian Wang
Changzhou Wang
Rik Warren
Wei Wei
Elizabeth Whitaker
Paul Whitney
Kevin S. Xu
Xiaoran Yan
Laurence T. Yang
S. Jay Yang
Yong Yang
Serpil Yuce
Reza Zafarani
Rifat Zahan
Qingpeng Zhang
Kang Zhao

Contents

Behavioral and Social Sciences

Cyber and Intelligence Applications

Information, Systems, and Network Sciences

Methodology

Behavioral and Social Sciences

Inferring Follower Preferences in the 2016 U.S. Presidential Primaries with Sparse Learning

Yu Wang[1(✉)], Yang Feng[1], Xiyang Zhang[2], and Jiebo Luo[1]

[1] Department of Computer Science, University of Rochester, Rochester, USA
{ywang176,yfeng23,jluo}@cs.rochester.edu
[2] School of Psychology, Beijing Normal University, Beijing, China
zxy2013@mail.bnu.edu.cn

Abstract. In this paper, we propose a framework to infer Twitter follower preferences for the 2016 U.S. presidential primaries. Using Twitter data collected from Sept. 2015 to Mar. 2016, we first uncover the tweeting tactics of the candidates and then exploit the variations in the number of 'likes' to infer followers' preference. With sparse learning, we are able to reveal neutral topics as well as positive and negative ones.

Methodologically, we are able to achieve a higher predictive power with sparse learning. Substantively, we show that for Hillary Clinton the (only) positive issue area is women's rights. We demonstrate that Hillary Clinton's tactic of linking herself to President Obama resonates well with her supporters but the same is not true for Bernie Sanders. In addition, we show that Donald Trump is a major topic for all the other candidates, and that the women's rights issue is equally emphasized in Sanders' campaign as in Clinton's.

Lessons from the primaries can help inform the general election and beyond. We suggest two ways that politicians can use the feedback mechanism in social media to improve their campaign: (1) use feedback from social media to improve campaign tactics within social media; (2) formulate policies and test the public response from the social media.

Keywords: Presidential primaries · Republicans · Democrats · Preference · Twitter · Sparse learning

1 Introduction

Twitter is playing a significant role in connecting the presidential candidates with voters [13,17]. Candidates increasingly formulate issue policies and attack rival candidates over Twitter. Some of the candidates' tweets have even entered into the Democratic and the Republican debates.[1] Between September 18, 2015

[1] For example, during the Democratic debate in Flint, Michigan, a tweet by Bernie Sanders targeting Hillary Clinton became the focal point. During the eleventh Republican debate, Donald Trump explicitly invited the audience to check his Twitter account.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 3–13, 2017.
DOI: 10.1007/978-3-319-60240-0_1

and March 1st, 2016, Hillary Clinton posted 1973 tweets, Bernie Sanders 2375 tweets, Donald Trump 3175 tweets, Ted Cruz 1876 tweets, and Marco Rubio 1333 tweets.[2] These tweets constitute a valuable data source because they are explicitly political in nature, they are many, and, importantly, they carry feedback information from the followers in the form of 'likes.'

In this work, we first extract the tweeting tactics of the candidates: we analyze which political figures are mentioned in these tweets and what issues are raised. Then we use L1-regularized negative binomial regression to infer followers' preference over these politicians and issues. Our study focuses on the five major candidates during the primaries: Hillary Clinton (D), Bernie Sanders (D), Donald Trump (R), Ted Cruz (R), Marco Rubio (R).[3]

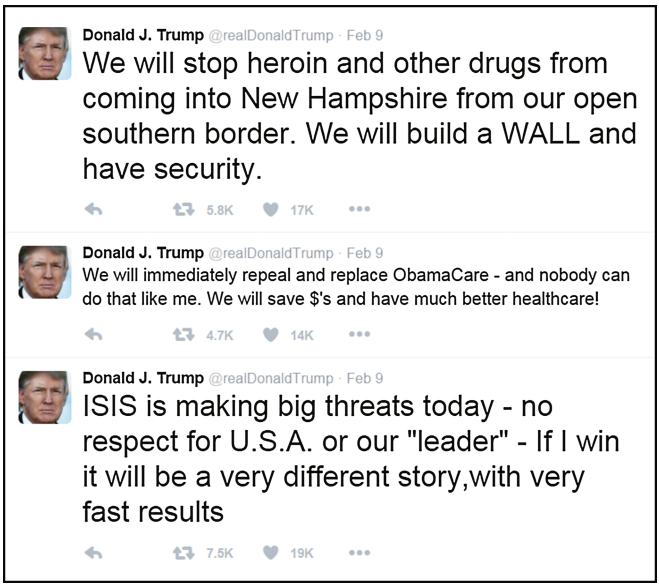


Fig. 1. Selected tweets Donald Trump (R) posted on February 9th.

Figure 1 shall illustrate our points well. It presents three tweets that Donald Trump (R) posted on February 9th, 2016, all of which are political in nature. The first tweet talks about drugs. The second tweet raises the issue of ObamaCare and points towards President Obama (D). The third tweet is about the ISIS. Trump supporters responded to these three tweets differently, assigning

[2] We do not count retweets, because retweets do not have as a feature the number of 'likes' and our focus is on the number of 'likes' in this work.

[3] The selection is based on both polling results and on the number of delegates that each candidate has. Marco Rubio (R) dropped out of the race on March 15th, 2016. Throughout, we follow the convention that Republican candidates are marked with (R) and Democratic candidates are marked with (D).

to the third tweet the most 'likes' and to the second tweet the fewest 'likes.' By connecting topics with responses, we are therefore able to infer the preferences of the followers.

2 Related Work

Our work builds upon previous research in electoral studies, behavioral studies, and sparse learning.

A number of studies have found that campaign and news media messages can alter voters' behavior [7,12]. According to Gabriel S. Lenz, public debates help inform some of the voters about the parties' or candidates' positions on the important issues [9]. In our work, we assume that tweets posted by the presidential candidates reveal their policy positions in various dimensions and that supporters reveal their policy preference by deciding whether or not to 'like' the tweets.

There is a large body of studies on using social media data to analyze and forecast election results. DiGrazia et al. [2] find a statistically significant relationship between tweets and electoral outcomes. MacWilliams [10] suggests that a candidate's number of 'likes' in Facebook can be used for measuring a campaign's success in engaging the public. According to Williams and Gulati [21], the number of Facebook fans constitutes an indicator of candidate viability. Wang, Li and Luo [19] study the growth pattern of Donald Trump's followers. Gayo-Avello, Metaxas and Mustafaraj [4] advocate that research should be accompanied with a model explaining the predicative power, which we heed to.

Substantively, our paper is closely related to two existing studies of the 2016 U.S. presidential election. Wang, Li, and Luo [18] use Twitter profile images to study and compare the demographics of the followers of Donald Trump and Hillary Clinton. Wang, Li, and Luo [20] model the number of 'likes' that each Trump tweet receives. Our work uses both the number of candidate followers (as a control variable) and the number of 'likes', and we study all the five major candidates rather than focus exclusively on Donald Trump.

There are quite a few studies modeling individual behaviors on social media. Lee et al. [8] model the decision to retweet, using Twitter user features such as agreeableness, number of tweets posted, and daily tweeting patterns. Mahmud, Chen, and Nichols [11] model individuals' waiting time before replying to a tweet based on their previous replying patterns. Wang, Li, and Luo [20] use Latent Dirichlet Allocation (LDA) [1] to extract tweet topics and model follower preferences. In this paper, we search for specific topics using predefined keywords. Compared with LDA, the limitation of our approach is that topics present in the document all share the same weight. The advantage of our approach is that our labels are more definitive and objective.

One of our innovations is to apply L1-regularization to uncover neutral topics. In implementing the L1-regularized negative binomial regression, our work

borrows heavily from [14], which introduces stochastic coordinate descent methods for LASSO [16] and for logistic regression. William Greene [5] discusses the likelihood formulation of negative binomial regression. Hastie et al. [6] give a very good discussion on using cross validation to select the penalty term and on using bootstrapping to draw inferences. To the best of our knowledge, our study is the first to implement and apply the regularized negative binomial regression to a real world problem.

3 Data

We use the dataset *US2016*, constructed by us with Twitter data.[4] The dataset contains a tracking record of the number of followers for all the major candidates in the 2016 presidential race, including Hillary Clinton (D), Bernie Sanders (D), Donald Trump (R), Ted Cruz (R), and Marco Rubio (R).[5] The dataset spans the entire period between September 18th, 2015 and March 1st, 2016 and is updated every 10 min.

Our dataset *US2016* contains all the tweets that the five candidates posted during the same period and the number of 'likes' that each tweet has received. In Table 1, we report the summary statistics of the dependent variable: 'likes.'

Table 1. Summary statistics

Variable	Mean	Std. Dev.	Min.	Max.	Observations
Donald Trump	4677.911	3742.242	722	32636	3175
Ted Cruz	548.552	608.731	8	8209	1876
Marco Rubio	590.479	879.356	3	11731	1333
Hillary Clinton	1800.052	1636.849	119	19923	1973
Bernie Sanders	2979.268	2966.524	204	57635	2375

We refer to these 'likes' as tallies and, in line with [10], we assume that the more likes the better. To visualize these tallies, we plot the density distribution for each candidate, grouped by party affiliation. We align the x axis so that it is easy to compare the distribution both across candidates and across parties. We observe that in the Democratic party, Sanders' tweets tend to receive more 'likes' than Clinton's tweets. Among Republican candidates, Trump's tweets receive more 'likes' than Cruz and Rubio. Equally important, we observe large variations in the distribution for all candidates.

[4] Some of the studies based on this dataset include [18–20].

[5] Eleven other candidates, including John Kasich (R) and Martin O'Malley (D), are also included in the dataset.

We believe part of the variations can be attributed to the topics embedded in the tweets: a more preferred topic generates more 'likes.' To operationalize this idea, we first multi-label each tweet for the following individual-based topics: President Obama (D), Hillary Clinton (D), Bernie Sanders (D), Martin O'Malley (D), Donald Trump (R), Ted Cruz (R), Marco Rubio (R), Jeb Bush (R), Ben Carson (R), Rand Paul (R), John Kasich (R), and Chris Christie (R).[6] We then multi-label each tweet for issue-based topics: ISIS, immigration, Iran, women's rights, education, drugs, gun control, abortion, economy and the Wall Street.[7]

Topic features are binary. We derive these features using keyword matching. For example, we assign to the Obama topic 1 for tweets that contain "Obama" and assign 1 to the Rubio topic for tweets containing "marcorubio" (case-sensitive) or "Rubio." For issue topics, we first transform the tweets to lowercase before the matching procedure. For example, we assign 1 to the abortion topic for tweets that either contain "abortion" or "planned parenthood."

In addition to topic feature variables, we control for the number of followers, the length of the tweet (after removing stop words), and whether or not the tweet contains an http link. In the online appendix, we report time-series follower data for all the five candidates during the primaries. One immediate observation is that Hillary Clinton (D) and Donald Trump (R) dominate their respective party in terms of Twitter followers.

4 Methodology

In this section, we first report on how we formulate our problem as an L1 regularized loss minimization problem and then we detail the coordinate descent algorithm for solving the problem, and the parameter selection procedure.

4.1 Model Formulation

Our problem starts as a standard negative binomial regression problem, linking the number of 'likes', which is count data, to the explanatory topics. In this regression, the conditional likelihood of the number of 'likes', y_j, is formulated as

$$f(y_j|v_j) = \frac{(v_j\mu_j)^{y_j}e^{-v_j\mu_j}}{\Gamma(y_j+1)}$$

[6] By the end of the New Hampshire primary, Martin O'Malley, Rand Paul, and Chris Christie have quit the race.

[7] The selection of political figures is based on the poll performance. For poll data, please refer to http://elections.huffingtonpost.com/pollster#2016-primaries. The selection of political issues follows the Bing Political Index. Available at https://blogs.bing.com/search/2015/12/08/the-bing-2016-election-experience-how-do-the-candidates-measure-up.

where $\mu_j = exp(\mathbf{x_j}\boldsymbol{\beta})$ is the link function that connects our topics to the number of 'likes' in the tweets and v_i is a hidden variable with a Gamma($\frac{1}{\alpha}$, α) distribution.[8] After plugging in the topic variables, the loss function, which is the negative unconditional log-likelihood of the 'likes' takes the form:

$$\begin{aligned} L = & -\sum_{j=1}^{N}[ln(\Gamma(m+y_j) - ln(\Gamma(y_j+1)) - ln(\Gamma(m)) \\ & + \mathrm{m}ln(p_j) + y_j ln(1-p_j)] \\ p = & \, 1/(1+\alpha\mu) \\ m = & \, 1/\alpha \\ \mu = & \, exp(\beta_0 + \beta_1 \text{Follower Count} + \beta_2 \text{Tweet Length} \\ & + \beta_3 \text{Hyperlink} + \beta_4 \text{Self Referencing} \\ & + \boldsymbol{\beta_5} \cdot \textbf{Political Figures} + \boldsymbol{\beta_6} \cdot \textbf{Political Issues}) \end{aligned}$$

where α is the over-dispersion parameter and will be estimated as well. Now we combine loss L with a penalty for the L1 norm of the coefficients $\boldsymbol{\beta}$ and arrive at the final formulation of our optimization problem.

$$\min_{\boldsymbol{\beta},\alpha} L(\alpha, \langle \boldsymbol{\beta}, \mathbf{X} \rangle, \mathbf{Y}) + \lambda ||\boldsymbol{\beta}||_1$$

4.2 Coordinate Descent Algorithm

Solving this minimization problem using coordinate descent, we first calculate the derivative of L with respective to $\boldsymbol{\beta}$ and $\log(\alpha)$ as follows:[9]

$$\begin{aligned} \frac{\partial L}{\partial \beta_i} = & -\sum_{j=1}^{N}[(y_i - m)\frac{1}{1+\alpha\mu_j}\mu_j x_i + \frac{y_j}{\mu_j}\mu_j x_i] \\ \frac{\partial L}{\partial \ln(\alpha)} = & -\sum_{j=1}^{N}[\frac{1}{\alpha^2}(\psi(m) - \psi(m+y_j) - \ln(p_j)) \\ & + [-(m+y_j)\frac{1}{1+\alpha\mu_j}\mu_j + \frac{y_j}{\alpha}]]\alpha \end{aligned}$$

where $\psi(x)$ is the digamma function, $\psi(x) := \frac{\partial \ln(\Gamma(x))}{\partial x}$.

[8] For a detailed introduction to the formulation of the negative binomial likelihood, please see [5,15].

[9] We calculate the derivative of L with respect to $\log(\alpha)$ instead of α to ensure that α stays positive throughout the optimization procedure.

We apply regularization to all the independent variables but not to the over-dispersion parameter α. Our coordinate descent algorithm is detailed as follows:[10]

Algorithm. Coordinate Descent for L1 Regularized Negative Binomial.

p: number of features, η: learning rate;
let β=**1** $\in R^p$, $\alpha = 1 \in R$, $\eta = 0.002$
for iter=1,2,...N do
  for i=1,2, ..,p do
    $L'_i = \frac{\partial L}{\partial \beta_i}$
    if $\beta_i - \eta L' > \eta\lambda$:
      $\beta_i \leftarrow \beta_i - \eta L' - \eta\lambda$
    else if $\beta_i - \eta L' < -\eta\lambda$:
      $\beta_i \leftarrow \beta_i - \eta L' + \eta\lambda$
    else:
      $\beta_i \leftarrow 0$
  end for
  $\mathrm{L}'_\alpha = \frac{\partial L}{\partial ln(\alpha)}$
  $\ln(\alpha) \leftarrow \ln(\alpha) - \eta \mathrm{L}'_\alpha$
  $\alpha \leftarrow e^{ln(\alpha)}$
end for

4.3 Model Selection

Following suggestions in [6], we use cross validation to select the penalty term and the model. Specifically, we use 5-fold cross validation, i.e. iteratively 20% of the dataset is held out for validation, to select λ_{CV} that minimizes the averaged mean squared prediction error (MSE):

$$MSE = \frac{1}{N}\sum_{j=1}^{j=N}(y_j - e^{\mathbf{x}_j\hat{\mathbf{b}}})^2$$

where y_j is the true number of 'likes' for the jth tweet and $e^{\mathbf{x}_j\hat{\mathbf{b}}}$ is the model's prediction.

When λ_{CV} equals zero, i.e. no penalty, our model is equivalent to the standard negative binomial regression, so we treat the $\lambda_{CV} = 0$ as the baseline model.[11] We report the cross validation error curve in the appendix of our online version. Based on cross validation and the MSE metric, the model we use for later point estimation and inference yields the highest predictive power.

In drawing inferences on β's distribution, we use 1,000 bootstrap realizations of $\hat{\beta}_{\lambda_{cv}}$ [3]. The detailed results are reported in Sect. 5.

[10] We have posted our codes at http://sites.google.com/site/wangyurochester.

[11] When the penalty term is zero, our results are identical to the standard outputs from Stata (http://www.stata.com) and R.

5 Empirical Results

In this section, we present our estimation of followers' preferences. We obtain the penalty terms for each candidate using 5-fold cross validation. We use bootstrapping to calculate the standard errors for each topic coefficient. The distributions of the coefficients for each candidate can be found in the online version of the paper.

We report the estimated coefficients for Donald Trump in the first column of Table 2. We find that Trump receives more 'likes' when he attacks Democrats and fewer 'likes' if he attacks fellow Republicans. This is consistent with the findings reported in [20], which uses unsupervised LDA. Trump is also particularly strong on ISIS. Due to regularization, Rand Paul (R) and abortion prove to be neutral topics.

By contrast, Hillary Clinton, reported in the last column of Table 2, clearly benefits from her association with President Obama. Clinton receives more 'likes'

Table 2. L1 Regularized negative binomial regression

	Trump	Cruz	Rubio	Sanders	Clinton
Likes					
Constant	−2.068** (0.105)	−1.791** (0.088)	−2.111** (0.13)	0.810** (0.106)	**0.000** (0.167)
Followers	6.532** (0.176)	6.653** (0.315)	5.793** (0.598)	3.828** (0.731)	1.787** (0.333)
Length	0.131** (0.022)	0.454** (0.045)	0.514** (0.086)	**0.000** (0.039)	−0.052 (0.049)
Http	0.097** (0.024)	0.143 (0.079)	0.134 (0.118)	−0.417** (0.045)	−0.171** (0.045)
Obama	0.386** (0.09)	0.073 (0.086)	−0.127 (0.078)	**0.000** (0.045)	0.445** (0.186)
Clinton	0.236** (0.046)	0.698** (0.332)	0.057 (0.079)	0.225 (0.126)	−0.378** (0.037)
Sanders	0.122 (0.126)	**0.000** (0.095)		0.319 (0.215)	0.128 (0.14)
Omalley	0.253 (0.164)			**0.000** (0.0)	
Trump	−0.311** (0.022)	0.789** (0.14)	1.704** (0.19)	0.961** (0.145)	0.336 (0.196)
Cruz	−0.098** (0.037)	−0.307** (0.047)	0.210 (0.413)		0.000 (0.009)
Jeb	−0.085** (0.036)	**0.000** (0.212)	**0.000** (0.142)	**0.000** (0.0)	**0.000** (0.009)
Carson	−0.073 (0.058)	0.169 (0.248)	**0.000** (0.241)	0.185 (0.326)	**0.000** (0.007)
Rand	**0.000** (0.087)				
Rubio	−0.072 (0.042)	0.345 (0.277)	−0.560** (0.221)	**0.000** (0.218)	**0.000** (0.055)
Kasich	−0.048 (0.068)	**0.000** (0.141)			**0.000** (0.123)
Christie	−0.007 (0.184)				
ISIS	0.414** (0.098)	0.306** (0.095)	0.725** (0.231)	**0.000** (0.025)	**0.000** (0.051)
Immigration	0.111 (0.092)	−0.053 (0.123)	**0.000** (0.108)	−0.257** (0.071)	−0.045 (0.086)
Iran	0.089 (0.106)	−0.105 (0.123)	**0.000** (0.04)	(—)	**0.000** (0.118)
Women's rights	−0.008 (0.094)	0.060 (0.172)	**0.000** (0.051)	0.389** (0.083)	0.176** (0.058)
Education	0.382 (0.33)	−0.090 (0.157)	**0.000** (0.036)	0.221** (0.062)	**0.000** (0.083)
Drugs	0.012 (0.261)			**0.000** (0.066)	**0.000** (0.081)
Gun Control	0.571 (0.292)	**0.000** (0.075)	**0.000** (0.044)	0.468 (0.297)	0.018 (0.071)
Abortion		**0.000** (0.103)	0.276 (0.224)	0.235 (0.149)	0.047 (0.07)
Economy	**0.000** (0.056)	**0.000** (0.097)	**0.000** (0.043)	−0.253** (0.048)	−0.229** (0.071)
Wall Street	−0.348 (0.185)	**0.000** (0.065)		−0.080 (0.058)	−0.330** (0.074)
α	0.0700	0.0016	0.0274	0.0708	0.1930
λ	0.00055	0.00065	0.00155	0.0018	0.0017

Standard errors in parentheses
** $p < 0.05$
Zero coefficients in bold.

on women's rights and fewer 'likes' on the economy and on Wall Street. The penalty term for Clinton (0.0017) is larger than that for Trump (0.00055). As a result, more coefficients for Hillary Clinton turn out to be zeros.

For Ted Cruz (R), he receives more 'likes' when attacking Hillary Clinton (D) and Donald Trump (R). The coefficients for other political figures are either zero or not significant. In issue areas, Cruz is strong on ISIS but not others. For Marco Rubio (R), the only two topics that earn him more 'likes' are Donald Trump and ISIS. His attacks on President Obama and Hillary Clinton do not resonate with his supporters very well. Bernie Sanders (D) is known as the candidate calling for a political revolution, and the preferences of his voters do show peculiarity.

We observe that Sanders is the only candidate who receives more 'likes' in the area of education and fewer 'likes' on immigration.[12]

Focusing on the parties, we observe that the Republicans benefit from their position on ISIS, while ISIS is a neutral topic for the Democrats. The women's rights issue wins Democratic candidates more 'likes,' but it is neutral for the Republicans. The economy is apparently hurting the Democratic candidates but not the three Republican candidates. The results also show that voters do not like to see candidates attack fellow party members.

For the control variables, we observe that the more followers one has the more 'likes' one receives, which is intuitive. In terms of the length of the tweets, we find that longer tweets tend to help Republicans win more 'likes,' but not so for the Democrats. Lastly, tweets with a hyperlink tend to receive more 'likes' for Donald Trump. The opposite is true for the Democrats.

Lastly, we emphasize that our regularized negative binomial regression outperforms the standard negative binomial, i.e. with no penalty terms, in two dimensions. First, we are able to achieve a smaller prediction error. Second, interpretation of the results is made substantially easier as many (33, to be specific) of the coefficients have shrunk to zero.

6 Conclusions

Twitter is playing an important role in connecting presidential candidates and potential voters through the tweeting and 'likes' channels. In this paper, we have proposed a framework to infer follower preferences via this feedback mechanism. Using Twitter data collected from September 2015 to March 2016, we first uncovered the tweeting tactics of the candidates and then we exploited the variations in the number of 'likes' to infer voters' preference. Besides positive and negative topics, we were also able to reveal neutral topics with sparse learning.

Methodologically, we were able to achieve a higher predictive power with sparse learning. Substantively, we showed that for Hillary Clinton the (only) positive issue area is women's rights. We demonstrated that Hillary Clinton's tactic of linking herself to President Obama resonates well with her supporters but the same is not true for Bernie Sanders. In addition, we showed that Donald

[12] For a review of Sanders' rise and his proposed policies, please see http://www.nytimes.com/2016/03/13/opinion/sunday/the-bernie-sanders-revolution.html.

Trump is a major topic for all the other candidates and that the women's rights issue is equally emphasized in Sanders' campaign as in Clinton's.

Lessons from the primaries can help inform the general election and beyond. We suggest two ways that politicians could use the feedback mechanism in social media to improve their campaign: (1) use feedback from social media to improve campaign tactics within social media; (2) formulate policies and test the public response from the social media.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. DiGrazia, J., McKelvey, K., Bollen, J., Rojas, F.: More tweets, more votes: social media as a quantitative indicator of political behavior. PLoS ONE **8**(11), e79449 (2013)
3. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, New York (1993)
4. Gayo-Avello, D., Metaxas, P.T., Mustafaraj, E.: Limits of electoral predictions using twitter. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011)
5. Greene, W.: Functional forms for the negative binomial model for count data. Econ. Lett. **99**, 585–590 (2008)
6. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton (2015)
7. Iyengar, S., Kinder, D., Matter, N.T.: Television and American Opinion. University of Chicago Press, Chicago (1987)
8. Lee, K., Mahmud, J., Chen, J., Zhou, M., Nichols, J.: Who will retweet this? detecting strangers from twitter to retweet information. ACM Trans. Intell. Syst. Technol. **6**(3), 1–25 (2015)
9. Lenz, G.S., Learning, O.C., Priming, N.: Reconsidering the Priming Hypothesis. Am. J. Polit. Sci. **53**, 821–837 (2009)
10. MacWilliams, M.C.: Forecasting congressional elections using facebook data. PS: Polit. Sci. Politics **48**(04), October 2015
11. Mahmud, J., Chen, J., Nichols, J.: When will you answer this? estimating response time in twitter. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (2013)
12. Riker, W.H.: The Art of Political Manipulation. Yale University Press, USA (1986)
13. Sanders, B., Revolution, O.: A Future to Believe In. Thomas Dunne Books (2016)
14. Shalev-Shwartz, S., Tewari, A.: Stochastic methods for L1-regularized loss minimization. J. Mach. Learn. Res. **12**, 1865–1892 (2011)
15. Stata.com. Negative binomial regression
16. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. J. Roy. Stat. Soc. Ser. B (Methodol.) **58**(1), 267–288 (1996)
17. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
18. Wang, Y., Li, Y., Luo, J.: Deciphering the 2016 U.S. presidential campaign in the twitter sphere: a comparison of the trumpists and clintonists. In: The 10th International AAAI Conference on Web and Social Media (ICWSM-16), Cologne, Germany, May 2016

19. Wang, Y., Li, Y., Luo, J.: To follow or not to follow: analyzing the growth patterns of the Trumpists on Twitter. In: News and Pulic Opinion Workshop at the 10th International AAAI Conference on Web and Social Media (ICWSM-16), Cologne, Germany, May 2016
20. Wang, Y., Luo, J., Li, Y., Hu, T.: Catching fire via 'Likes': inferring topic preferences of trump followers on twitter. In: The 10th International AAAI Conference on Web and Social Media (ICWSM 2016), Cologne, Germany, May 2016
21. Williams, C.B., Gulati, G.J.: The political impact of Facebook: evidence from the 2006 midterm elections and 2008 nomination contest. Politics Technol. Rev. **1**, 11–21 (2008)

Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter

Zhiwei Jin[1,2(✉)], Juan Cao[1,2], Han Guo[1,2], Yongdong Zhang[1,2], Yu Wang[3], and Jiebo Luo[3]

[1] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing 100190, China
{jinzhiwei,caojuan,guohan,zhyd}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] University of Rochester, Rochester, NY 14627, USA
ywang.tsinghua@gmail.com, jluo@cs.rochester.edu

Abstract. The 2016 U.S. presidential election has witnessed the major role of Twitter in the year's most important political event. Candidates used this social media platform extensively for online campaigns. Meanwhile, social media has been filled with rumors, which might have had huge impacts on voters' decisions. In this paper, we present a thorough analysis of rumor tweets from the followers of two presidential candidates: Hillary Clinton and Donald Trump. To overcome the difficulty of labeling a large amount of tweets as training data, we detect rumor tweets by matching them with verified rumor articles. We analyze over 8 million tweets collected from the followers of the two candidates. Our results provide answers to several primary concerns about rumors in this election, including: which side of the followers posted the most rumors, who posted these rumors, what rumors they posted, and when they posted these rumors. The insights of this paper can help us understand the online rumor behaviors in American politics.

1 Introduction

In the 2016 U.S. presidential election, Twitter became a primary battle ground: candidates and their supporters were actively involved to do campaigns and express their opinions by tweeting [13]. Meanwhile, the fact that various rumors were spreading on social media during the election became a serious concern. Among all the 1,723 checked rumors from the popular rumor debunking website Snopes.com, 303 rumors are about Donald Trump and 226 rumors are about Hillary Clinton. These rumors could potentially have negative impacts on their campaigns.

In this paper, we aim to understand the rumor spreading behaviors of candidates' followers. A rumor is defined as a controversial and fact-checkable statement [2]. Existing machine learning methods for rumor detection [1,7,14] commonly require extensive labeled training data, which is expensive to label for the rumor detection problem. Besides, it is difficult to tell what rumors are posted

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 14–24, 2017.
DOI: 10.1007/978-3-319-60240-0_2

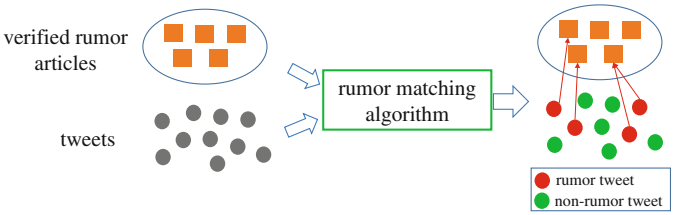


Fig. 1. Rumor detection as a text matching task.

as their binary results are not easily interpretable. Considering these limitations, we use the checked rumors from Snope.com as the objective golden samples and propose to detect rumors as a text matching task (Fig. 1). In this scheme, a set of verified rumor articles are collected as standard samples for reference. Each tweet is compared with these verified rumors to see if they match closely. Compared with existing approaches, our approach requires minimal human labeling and the matching results can be easily interpreted.

In order to find the best matching algorithm, we conduct a comparative study of several competing algorithms. These algorithms are executed on a reasonably sized set of 5,000 manually labeled tweets to provide a fair performance comparison. We then detect rumors with the selected most effective matching algorithm on over 8 million of tweets from 14,000 followers of the two leading presidential candidates: Hillary Clinton and Donald Trump. We inspect the rumor detection results from different aspects to answer following questions: which side posted the most rumors? who posted these rumors? what rumors did they post? when did they post rumors? These insights help us understand the rumor tweeting behaviors of different groups of followers and can be helpful for mining voters' real intentions and accurately detecting rumors during political events in the future.

2 Related Work

Online social media have gained huge popularity around the world and become a vital platform for politics. However, the openness and convenience of social media also fosters a large amount of fake news and rumors which can spread wildly [3]. Compared with existing rumor detection works that are focused on general social events or emergency events [5], this paper presents a first analysis of rumors in a political election.

Most existing rumor detection algorithms follow the traditional supervised machine learning scheme. Features from text content [1], users, propagation patterns [14] and multimedia content [6,8] are extracted to train a classifier on labeled training data. Some recent works further improve the classification result with graph-based optimization methods [4,5,7].

Although machine learning approaches are very effective under some circumstances, they also have drawbacks. The supervised learning process requires a large amount of labeled training data which are expensive to obtain for the rumor detection problem. They derive features in a "black box" and the classification results are difficult to interpret. In [15], a lexicon-based method is proposed for detecting rumors in a huge tweet stream. They extracted some words and phrases, like "rumor", "is it true", "unconfirmed", for matching rumor tweets. Their lexicon is relatively small, thus the detection results tend to have high precision but low recall of rumors.

In this paper, we formulate the rumor detection as a text matching task. Several state-of-the-art matching algorithms are utilized for rumor detection. TF-IDF [12] is the most commonly used method for computing documents similarity. BM25 algorithm [11] is also a term-based matching method. Recent research in deep learning for text representation embeds words or documents into a common vector space. Word2Vec [10] and Doc2Vec [9] are two widely used embedding models at the word and paragraph levels, respectively.

3 Dataset

We collect a large-scale dataset for analyzing rumors during the 2016 U.S. presidential election from Twitter. For reliable rumor detection, we obtain a set of verified rumor articles from Snopes.com. We also manually construct a testing set to fairly evaluate the rumor detection methods.

Using the Twitter API, we collect all the users who are following the Democratic presidential candidate Hillary Clinton and the Republic presidential candidate Donald Trump. We randomly select about 10,000 followers from each candidate's follower list, which contains millions of followers. We then collect up to 3,000 most recent tweets for each user using the Twitter API. Altogether, we get 4,452,087 tweets from 7,283 followers of Clinton and 4,279,050 tweets from 7,339 followers of Trump in our dataset.

We collect a set of verified rumor articles from Snopes.com as gold standard samples for rumor matching. Snopes.com is a very popular rumor debunking website. Social media users can nominate any potential rumor to this site. The employed analysts then select some of these controversial statements to fact-check them as rumors or truth. An article is presented for each checked rumor by these professional analysts, which gives conclusion of the rumor followed by full description, source, origin, suporting/opposing evidences of the rumor story. We collect the articles of all the 1,723 checked rumors on this website to form the verified rumor article set.

To quantitatively evaluate the performance of rumor detection methods, we build a manually labeled tweet set. We randomly select 100 rumors from the verified rumor set. For each verified rumor article, we search the large tweet set with keywords extracted from the article. Each tweet in the search result is manually examined to check if it matches the rumor article. After these procedures, we obtain a set of 2,500 rumor tweets from 86 rumor articles. We then

randomly sample the same number of unrelated tweets as negative samples. In this set, not only is each tweet labeled as rumor or not, but the rumor tweets are also labeled with their corresponding verified rumor articles. Therefore, we can perform both general rumor classification and fine-grained rumor identification with this dataset. The following is an example of a verified rumor article and three associated tweets.

Verified rumor article[1]:

Shaky Diagnosis. A montage of photos and video clips of Democratic presidential candidate Hillary Clinton purportedly demonstrates she has symptoms of Parkinson's disease. Photos and video clips narrated by a medical doctor demonstrate that Democratic presidential candidate Hillary Clinton likely has Parkinson's disease.....

Associated rumor tweets:

1. *Hillary collapse at ground zero! game over, Clinton! Parkinson's blackout!*
2. *Wikileaks E-mails: Hillary looked into Parkinson's drug after suffering from "decision fatigue".*
3. *Exclusive Report: How true is this?? Hillary Clinton has Parkinson's disease, doctor confirms.*

4 Rumor Detection

We formulate rumor detection on Twitter as a matching task in this paper (Fig. 1). With reliable rumor articles collected from Snopes.com, the key part of this scheme is the matching algorithm. Compared with the traditional rumor classification algorithms, our rumor matching scheme not only outputs a tweet as rumor or not but also identifies which rumor article it refers to if it is a rumor tweet. We perform comparative studies of different matching algorithms on both the classification and the identification task of rumor detection.

4.1 Rumor Detection Algorithms

We compare the performance of five matching algorithms with respect to the rumor detection task. The first set of methods includes two widely used term-based matching methods: TF-IDF and BM25. The second set includes two recent semantic embedding algorithms: Word2Vec and Doc2Vec. The third set is a lexicon-based algorithm for rumor detection on Twitter stream.

TF-IDF [12] is a widely used model in text matching. In this model, both the tweets and the verified rumor articles are represented as a v-dimensional vector, where v is the size of the dictionary of the corpus. Each element in the vector stands for the TF-IDF score of the corresponding word in the text.

[1] The full article is available at: http://www.snopes.com/hillary-clinton-has-parkinsons-disease/.

TF is the term frequency. IDF score is the inverse document frequency, which is calculated on the whole corpus.

BM25 [11] is also a text similarity computing algorithm based on the bag-of-words language model. It is an improvement of the basic TF-IDF model by normalizing on term frequency and document length. Both TF-IDF and BM25 have been widely used in many related studies.

Word2Vec [10] represents each word in a corpus with a real-valued vector in a common semantic vector space. Compared with traditional lexical-based matching models, this algorithm evaluates the quality of word representations based on their semantic analogies. We use the pre-trained Word2Vec model on a corpus of 27 billion tweets. The word dimension is 200. To aggregate a presentation for a whole text, we take the average of word vectors in the text.

Doc2Vec [9] is also an embedding algorithm on the semantic space, which can directly learn the distributed representations of documents. We use all the tweets and rumor articles for the unsupervised training of the model after standard pre-processing. We use the default parameter settings as in [9]. After training, tweets and verified rumors are represented as 400-dimensional vectors.

For Word2Vec and Doc2Vec, the matching score between a tweet and a rumor article is computed based on the cosine distance of their vector represention.

Lexicon matching [15] is a lexicon-based rumor detection algorithms for efficiently detecting in huge tweet streams. It mines a couple of signal words or phrases for recognizing prominent rumor tweets. We use the same set of regular expression patterns as in [15] to match rumor tweets.

4.2 Evaluation on Rumor Classification Task

TF-IDF, BM25, Word2Vec and Doc2Vec represent texts as numeric vectors. The similarity between a tweet and a verified rumor is computed as their matching score. By setting a threshold h for each method, we classify tweets with matching scores larger than h as rumor tweets. We can achieve different precision and recall of rumor classification by varying the threshold. We test all the four methods on the 5,000 labeled tweet set. Figure 2 illustrates the precision-recall curves

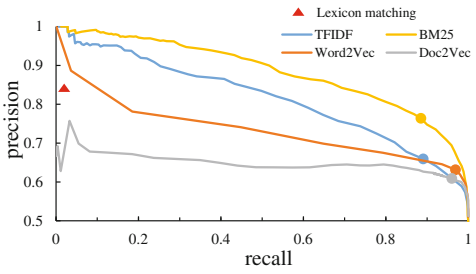


Fig. 2. The comparative performance of four matching algorithms.

of these four algorithms. The lexicon matching algorithm detects rumors by keywords matching, thus its result is actually fixed (as a single point in Fig. 2).

The highlighted round points on each curve in Fig. 2 are points where the F1-measures are maximized, at 0.758, 0.82, 0.764 and 0.745 for TF-IDF, BM25, Word2Vec and Doc2Vec, respectively. The red triangle is the fixed result of lexicon matching. These results show that BM25 reaches the best performance among all the five rumor classification methods under different metrics. The two term-based methods (TF-IDF and BM25) outperform the semantic-embedding and lexicon-based methods. For semantic-embedding, Word2Vec is slightly better than Doc2Vec. Lexicon matching can reach a rumor classification precision of 0.862, but its recall (0.008) is too low.

4.3 Evaluation on Rumor Identification Task

One extra advantage of our proposed rumor matching scheme is its ability to identify what rumor article a rumor tweet refers to, apart from classifying it as a rumor tweet. To compare the rumor identification performance of the four algorithms, we compute the similarity score between each pair of tweet and verified rumor article for the 2,500 labeled rumor tweets and 1,723 verified rumor articles. If the most similar rumor article of a tweet is exactly the same labeled rumor article for it, then this is an accurate rumor identification (Table 1).

Table 1. The accuracy of rumor identification task.

	TF-IDF	BM25	Word2Vec	Doc2Vec
Accuracy	0.795	**0.799**	0.557	0.658

From the overall rumor identification accuracy of each rumor matching methods, the BM25 algorithm achieves the best accuracy of 0.799. The accuracy of BM25 is only slightly better than that of TF-IDF, although it has major advantage in the rumor classification task. This is probably because BM25 can distinguish non-rumor tweets much better than TF-IDF. Another interesting finding is that Doc2Vec actually performs better on the rumor identification task than Word2Vec, although the latter has slightly better performance on the rumor classification task.

5 Analyzing Rumor Tweets Pertaining to the Election

This paper analyzes rumor tweets related to the 2016 U.S. presidential election. For rumor analysis at a large scale, in this section, we use the proposed rumor detection algorithm to detect rumor tweets from over 8 million tweets collected from the followers of Hillary Clinton and Donald Trump. Specifically, we match each rumor tweet with corresponding rumor articles in the verified set with

BM25 algorithm. To conduct a reliable and accurate analysis, we prefer a high precision for our rumor detection result. We set the similarity threshold $h = 30.5$ so that we can achieve a very high rumor classification precision of 94.7% and the recall of 31.5% on the test set. Based on the results, we obtain insights into the rumor tweeting behaviors from various aspects.

5.1 Which Side Posted the Most Rumors?

Twitter became an online battle field during the election. The number of rumor tweets reflects the involvement of candidates' followers in the election campaign. Which side of followers were involved most in spreading rumor tweets? To answer this question, we use rumor classification method to detect rumors in the subset of tweets of the two candidates, respectively. Given our focus on rumors during the election period, we also analyze rumor tweets posted from April up to the present.

Table 2. Rumor tweet ratio of two candidate's follower groups.

	Clinton's followers	Trump's followers
Entire time	1.20%	1.16%
Election period	1.26%	1.35%

From the results in Table 2, we find that:

- For entire time, Clinton's followers are slightly more active in posting rumor tweets than Trump's followers. 1.2% tweets are rumor tweets from Clinton's followers, which is about 4% more than that of Trump's followers.
- People tend to post more rumor tweets in the election time than in the whole time, especially for Trump's followers. Comparing their election period and all time rumor tweeting, Trump's followers have a rumor tweet ratio of 1.35% during the election, which is 18% higher than that in all time.
- During the election time, Trump's followers are more active in rumor tweeting than Hillary's followers. As the figure suggests, Trump's followers become much more involved in posting rumors at the election time, compared with Clinton's followers.

5.2 Who Posted These Rumors?

Who are behind the rumors spreading on Twitter? We investigate this issue by analyzing rumor tweets posted by individual followers of the two candidates.

We rank users by the total number of rumor tweets they posted. We find that the majority of rumors are posted by only a few users: for both Trump's and Clinton's followers, the top 10% users posted about 50% rumor tweets and the top 20% users posted about 70% of all rumor tweets.

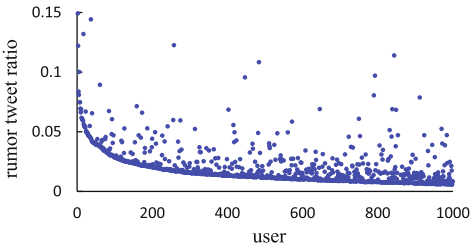


Fig. 3. Rumor tweet ratio of Clinton's followers.

Are these rumor-prolific followers just active in tweeting rumors or active in general tweeting as well? To understand this, we calculate the ratio of rumor tweets in all tweets posted by a user. We rank users based on the rumor ratio in their tweets. In Fig. 3, we show the top 1000 users from Clinton's followers. We observe that followers who post more rumor tweets also tend to have a larger rumor tweet ratio. This means the rumor-prolific users did not randomly post any tweets; they were actually more concentrated on posting rumor tweets than the users who occasionally post a few rumor tweets.

Case Study. After analyzing rumor spreaders at a large scale, we can also conduct a detailed analysis for a specific user.

Take one of Trump's followers, for example. This user posted 3,211 tweets in our dataset, 307 of which are detected as rumors. The rumor tweet ratio is as high as 9.6%, which means this user is very active in rumor tweeting. By examining the top keywords in all tweets posted by the user (Table 3), we find this person is very focused on posting tweets about the 2016 presidential election: "Clinton", "Sanders", "Trump" and "election" are the most mentioned words in the tweets. After rumor detection, we find that the rumor tweets of this user are mainly about Clinton and Sanders rather than Trump: 15% tweets about Clinton and 28% tweets about Sanders are rumor tweets, while only 10% tweets about Trump are rumors.

Table 3. The number of keywords in the tweets posted by one follower of Trump.

	Clinton	Sanders	Trump	election	Democratic	FBI
Rumor	1,106	620	275	271	97	88
Nonrumor	197	247	34	30	54	15

5.3 What Rumors Did They Post?

During the election, most rumors are focused on the candidates. By analyzing what people from different groups tweeted about in rumors, we can understand their intentions in this election. We use BM25 to identify the content of each

rumor tweet by matching it with the verified rumor articles from Snopes.com. Given our focus on the two primary presidential candidates, Hillary Clinton and Donald Trump, we only analyze rumor tweets related to them. After normalizing the number of candidate-related rumor tweets with the total number of rumor articles for this candidate in our dataset, we plot the rumor content spread by Trump's and Clinton's followers in Table 4. We offer some analysis of this figure based on the normalized rumor tweet number.

Table 4. Normalized number of rumors posted by followers of Trump and Clinton.

	Clinton's followers	Trump's followers
Rumors about Clinton	50.31	54.50
Rumors about Trump	53.95	51.94

First, both follower groups post rumors about their favored candidate as well as the opponent candidate. Supporters of one candidate would spread rumors about the opponent as a negative campaign tactic and debunk rumors about their favored candidate. For example, we show two tweets about the rumor "*Hillary Clinton has Parkinson's disease*" from our dataset:

Tweet 1: *Medical experts watching debate said Hillary showed "Telltale Signs" of Parkinson's Disease.*

Tweet 2: *"I know her physician; I know some of her health history which is really not so good" Trump's MD on Hillary—her MD shared her info with him?*

The first tweet comes from a follower of Trump. It is spreading the rumor by quoting medical experts. The second tweet comes from a follower of Clinton. It is questioning the truthfulness of the rumor.

Second, users would post more rumor tweets about the opponent candidate than their favored candidate. Clinton's followers post 8% more rumor tweets about Trump than rumors about Clinton. Trump's followers post 5% more rumor tweets about Clinton than rumors about Trump. Moreover, Trump's followers are more active in this rumor tweeting behavior towards both Clinton and Trump. The numbers of rumor tweets about the two candidates posted by Trump's followers are both larger than those of Clinton's followers.

5.4 When Did They Post These Rumors?

Analyzing the time patterns of rumor tweeting can reveal insights of online campaign. We plot the rumor tweeting of Clinton's followers over six months (April 2016 to September 2016) in Fig. 4. We annotate the key events for some rumor peaks in the figure to understand the inherent reason behind them. We find that rumors are peaked in three types of occasions: (1) key point in the presidential campaign, such as "the presidential debate" and "official nominee"; (2) controversial emergency events, including "the Orlando shooting"; (3) events

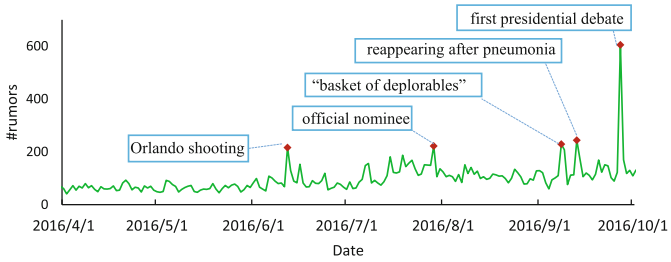


Fig. 4. Rumor tweet timeline of Clinton's followers.

triggering rumors, such as "reappearing after pneumonia". This insight reminds us to pay more attention to rumors during these types of events in future political campaigns.

6 Conclusions

This paper studies the rumors spreading phenomenon on Twitter during the 2016 U.S presidential election. We propose a reliable and interpretable approach to detecting rumor tweets by matching them with verified rumor articles. We conduct a comparative study of five algorithms for this rumor matching approach. With a rumor detection precision of 94.7%, we use this method to detect rumors in over eight million tweets collected from the followers of the two primary presidential candidates. We provide a thorough analysis on the detected rumor tweets from the aspects of people, content and time. We would benefit from the discovery in the paper to understand rumors during political events and build more effective rumor detection algorithms in the future.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800403 and the National Nature Science Foundation of China (61571424, 61525206). Jiebo Luo and Yu Wang would like to thank the support from the New York State through the Goergen Institute for Data Science. Zhiwei Jin gratefully thanks the sponsorship from the China Scholarship Council.

References

1. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW), pp. 675–684. ACM (2011)
2. DiFonzo, N., Bordia, P.: Rumor psychology: Social and organizational approaches. American Psychological Association (2007)
3. Friggeri, A., Adamic, L.A., Eckles, D., Cheng, J.: Rumor cascades. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (2014)

4. Gupta, M., Zhao, P., Han, J.: Evaluating event credibility on twitter. In: Proceedings of the SIAM International Conference on Data Mining, p. 153. Society for Industrial and Applied Mathematics (2012)
5. Jin, Z., Cao, J., Jiang, Y.G., Zhang, Y.: News credibility evaluation on microblog with a hierarchical propagation model. In: 2014 IEEE International Conference on Data Mining (ICDM), pp. 230–239. IEEE (2014)
6. Jin, Z., Cao, J., Zhang, Y., Yongdong, Z.: Mcg-ict at mediaeval 2015: Verifying multimedia use with a two-level classification model. In: Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop (2015)
7. Jin, Z., Cao, J., Zhang, Y., Luo, J.: News verification by exploiting conflicting social viewpoints in microblogs. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, USA, 12–17 February 2016
8. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and statistical image features for microblogs news verification. IEEE Trans. Multimedia **19**(3), 598–608 (2017)
9. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. In: ICML, vol. 14, pp. 1188–1196 (2014)
10. Mikolov, T., 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, pp. 3111–3119 (2013)
11. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Now Publishers Inc., Hanover (2009)
12. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval. J. Documentation **28**(1), 11–21 (1972)
13. Wang, Y., Luo, J., Niemi, R., Li, Y., Hu, T.: Catching fire via "likes": Inferring topic preferences of trump followers on twitter. In: ICWSM (2016)
14. Wu, K., Yang, S., Zhu, K.Q.: False rumors detection on sina weibo by propagation structures. In: IEEE International Conference on Data Engineering, ICDE (2015)
15. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of the 24th International Conference on World Wide Web, pp. 1395–1405 (2015)

How Emotional Support and Informational Support Relate to Linguistic Alignment

Yafei Wang, David Reitter, and John Yen(✉)

College of Information Sciences and Technology,
The Pennsylvania State University,
University Park, PA 16802, USA
{yxw184,reitter}@psu.edu, jyen@ist.psu.edu

Abstract. Linguistic alignment in text-based communication means that people tend to adjust their language use to one another both in terms of word choice and sentence structure. Previous studies about linguistic alignment suggested that these two forms of adaptation are correlated with each other, and that they build up to alignment at a higher representational level, such as pragmatic alignment for support functions. Two types of social support have been identified as important for online health communities (OHCs): emotional and informational support between support seekers and support providers. Do the two lower-level alignment measures (lexical and syntactic) relate to these two types of social support in the same way or, are they different? Our hypothesis was that they are similar, due to their correlation relationship. However, we found that, based on an analysis of a 10-year online forum for cancer survivors, the lower-level alignment measures have distinct relationships to the two higher-level support functions. In this paper, we describe this finding and its implications regarding potential refinement of the Interactive Alignment Model.

1 Introduction

For people living with a chronically illness like cancer, online "peer-to-peer" healthcare, such as exchanging information and social support, is an essential pathway for patients getting social support, helping each other, and increasing the quality of life [12]. A study from Pew Research Center [13] shows that 34% internet users have read others' health-related experience online, and 18% internet users have tried to find other patients with similar experiences online. A primary way of "peer-to-peer" healthcare interactions is using social network websites, such as Facebook groups, online health communities, etc. Online health communities (OHCs) serve prominent function of exchanging social support.

Generally, there are four types of social support: emotional support, informational support, tangible support and appraisal support [21]. However, the social support people seeking and providing in web-based conversation is either informational or emotional support [31]. This is because patients with chronic disease can benefit from not only information provided by their peers, such as

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 25–34, 2017.
DOI: 10.1007/978-3-319-60240-0_3

side effects of chemo therapy, but also emotional support, like understanding and caring. Moreover, anecdotal information and first-hand accounts from peers can be useful to patients in terms of reassurance, when patients do not weigh it against the evidence-based, scientific medical information obtained from professionals. Previous studies showed that receiving appropriate social support in thread-based conversations is correlated with members' commitment of online health support groups [31]. Similarly, [28] reported that members have higher satisfaction of online dialogues when they got the social support they need.

Another line of work in this paper is communication accommodation. Communication accommodation [15] is a communication phenomenon showing that people tend to adapt their gestures [4], speech [6], and language use [16], to accommodate other participants while interaction. Among these accommodation phenomena, linguistic alignment [24] is one of the useful models of accommodation, analyzing accommodation on language usage, with respect to words [16], syntax [6], and more, although general alignment may extend beyond language use. In *web-based communication*, linguistic alignment is the primary form of accommodation which is found under various settings [9,22] as well as in OHCs [29].

Although linguistic alignment is well-studied, the reason of linguistic alignment is still arguable. This phenomenon may be subject to more or less explicit, conscious control, while many studies suggest an automated, implicit process [20,26]. Lexical alignment, i.e., adaptation to a conversation partner's choice of words, is arguably more under a speaker's conscious control. [17] suggested that syntactic alignment is due to lexical repetition. Thus, under that finding, we expect consistency, and very little difference in correlation of syntactic and lexical priming with secondary variables.

We choose the higher-level representation of a conversation, such as the outcome of a given conversation as a secondary variable in this paper. Theoretically, Interactive Alignment Model [24] suggests that linguistic alignment at lexical and syntactic levels build up to alignment at a higher representation level, such as pragmatic alignment [30] and mutual understanding, at the conversational level. Therefore, understanding how linguistic alignment at different levels interacts with social support is important. Also, since emotional support and informational support are two important types of support in dialogues from OHCs, it is essential to understand and predict linguistic phenomena such as alignment.

Hence, the research question of this paper is whether different types of discussions influence the linguistic alignment behavior of the support provider? To be specific, are there differences regarding the relationship between lexical alignment and syntactic alignment with regard to emotional support and informational support threaded discussions? Our goal of research is to uncover the relationship between lexical alignment and syntactic alignment in two types of posts in OHCs: emotional support post and informational support post.

This paper is organized as follows. We firstly review previous studies on linguistic alignment and its mechanism. Then, we introduce social support measurements and linguistic measurements including lexical alignment, syntactic alignment, and the corpus we used in this paper. Further, we conduct an

experiment analyzing how linguistic alignment correlates to social support types at post level. We end by comparing the experiment results to analyze different correlations between social support and linguistic alignment measurements.

The contributions of this paper include: (1) providing a deeper understanding about the different relationship between lexical and syntactic alignment and emotional and informational support based dialogues in online health communities, (2) motivating further studies to investigate the explanations or casual factors of these different relationships, and (3) making theoretical progress on psycholinguistic models regarding linguistic alignment.

2 Related Work

While linguistic alignment is a well-studied linguistic phenomenon, its mechanism is still debatable among researchers. Previous studies [6,27] considered that linguistic alignment happened at syntactic level is due to the internal mechanism of conversation, which happens unconsciously. In other words, people do not consciously adapt the sentence structure they used during conversation. While, many studies argued that alignment is due to the purpose of the dialogues, such as building common ground [8], a higher level strategy [10], or an artificial task that requires such mutual understanding [27].

Also, linguistic alignment is a useful predictor driving the higher level representation, such as the outcome of a given conversation. For example, linguistic convergence helps conversation participants win decision-making games [14], and even the outcome of speed dating [18].

Furthermore, linguistic alignment is a meaningful signal of revealing sociological relationship among conversation participants. [9], and [11] used linguistic alignment measures to identify power and power dynamics on a conversational dataset from Twitter. Generally, conversational participants with a lower power tend to align with participants with higher power. This phenomenon is also identified in Wikipedia editors' discussion groups [23] and general online communities [19].

3 Automated Classification of Social Support Types

We discuss two main types of social support in OHCs, emotional support and informational support [31]. Specifically, *Emotional support* is a type of social support that an individual provides a support seeker the provision of "understanding, empathy, encouragement and concern" [2]. In contrast, *informational support* provides support seeker "facts, advice, referral, teaching, personal experiences, and information" [2].

Table 1 shows an anonymized threaded interaction between support seeker and support provider. The message from support provider, i.e. the message in the second row, includes both emotional support, such as encouragement, and informational support, like personal experience. In OHCs, patients often responds to support requests in the forum by disclosing his/her personal experiences together

Table 1. An anonymized threaded interaction between support seeker and support provider. The sentences in gray and black in the second row are providing emotional and informational support, respectively. The support provider tends to use similar words and syntax as the support seeker.

A:	"Newly diagnosed at age XX, went to BS and they want to remove the cells., ... Everything I've read says it is very slow growing and surgery is not always necessary. She said I'm stage 1., I'm petrified of any surgery. ... My mind is overwhelmed and just won't stop!, I know I'm lucky it was caught so early and I'm lucky it's only stage 1 but ....., Thank you in advance."
B:	"You are lucky to have caught this early. You are probably in shock right now so take a few days to think. I'm with your family. Better to get a few cells removed that wait until it becomes a tumor! Breast surgery is so routine these days. My sister got similar surgery and felt fine the same day. Good luck!"

with showing understanding and giving encouragement. These two types of social support are often intertwined.

Therefore, we distinguish these two types of social support by leveraging an automated classifier at the sentence level developed by [5]. As reported in [5], the sentence classifier was built on 1,066 hand-tagged sentences selected from Cancer Survivor's Network, which is the same corpus in this paper; and the initial agreement between two taggers was 89%. The features of building the sentence classifier include combined words, part-of-speech, subjective words, cancer-related words, linguistic patterns about emotional support and informational support and etc. [5]. The macro-averaged precision, recall and F-1 score of that model reached 0.841, 0.842, and 0.840, respectively. Based on the result of this sentence-level classification, we further determine the support type of a post – if the number of emotional support sentences is larger than that of informational support sentences, the post is an *emotional support post*; otherwise, it is considered an *informational support post*. Following [5], the amount of one type of support, emotional support and informational support, in a reply post is quantified as: $Index_{Type} = SentNum_{Type}/SentNum_{Classified}$.

4 Linguistic Alignment Measures

There are multiple metrics, such as indiscriminate local linguistic alignment (LLA) [14], subtractive conditional probability (SCP) [9], linguistic style matching (LSM) [22], word-based hierarchical alignment model [10], and so on, quantifying linguistic alignment phenomenon. We choose LLA, which was used to measure linguistic alignment in online communities [29,32], measures linguistic alignment among messages which computes lexical and syntactic alignment in a similar way.

Indiscriminate Local Linguistic Alignment, implemented in [29], measures the linguistic alignment at lexical and syntactic levels. Generally, LLA at the lexical level measures the normalized word repetition in both prime and target posts

in the same conversation (in the same thread). For example, it is computed as the number of words which occur in both the target post (in the second row of Table 1) and the prime post (in the first row of Table 1) normalized by the numbers of words in prime and target posts. Formally, *Lexical Indiscriminate Local Linguistic Alignment (LILLA)* is calculated as:

$$LILLA(\text{target}, \text{prime}) = \frac{\sum_{word_i \in \text{target}} 1_{\text{prime}}(word_i)}{length(\text{prime}) \times length(\text{target})} \quad (1)$$

where length(*post*) is the number of words in the post *post*, $1_{prime}(word_i)$ is an indicator function that the outcome is 1 if $word_i$ is in the prime message.

Similarly, *Syntactic Indiscriminate Local Linguistic Alignment (SILLA)* measures the percentage of syntactic rule repetition which appears in both prime and target posts in the same conversation. For each sentence in a post, we annotated it as a collection of syntactic rules using phrase structure trees generated by Standford CoreNLP Parser[1]. Then, we compute normalized syntactic rule repetition for SILLA.

5 Corpus

We use a collection of online threads from Cancer Survivor's Network (CSN) (csn.cancer.org), with more than 166,000 registered users and 41 sub-communities [25]. In CSN, cancer patients and cancer survivors with the same disease and under similar situations are in the same sub-communities. Members in the same sub-community often exchange their personal feelings and experiences of being under similar difficult personal circumstances and the associated emotional burden. Thus, most conversations happened in CSN are support-oriented conversations, which are either seeking and offering emotional support, informational support, or a mixture of social support. We use all the threads taken place from the two largest sub-forums in CSN, Breast cancer and Colorectal cancer, between June 2000 to October 2010, as our two corpora.

Table 2. The number of Emotional and Informational Support Posts in two sub-communities

	Breast	Colorectal
Emotional Support Posts	111,495	93,355
Informational Support Posts	20,610	16,099

A conversational *thread* is an initial post (normally seeking social support) followed by a sequence of replies (normally offering emotional or informational

[1] http://stanfordnlp.github.io/CoreNLP/.

support) in temporal order. Formally, a conversational thread is shown as $< P_0, P_1, \cdots, P_i, \cdots, P_n >$, where P_0 denotes *initial post*, and the author of initial post is called *thread initiator*. The post *distance* indicates how much information has been discussed between two posts. Thus, given a post pair, P_i and P_j, the *distance* is calculated as $j - i$. Following [5], each sentence in a reply post is classified as providing either emotional or informational support. In this paper, we only consider either emotional support post or informational support post. The distribution of these two types of posts is shown in Table 2.

6 Alignment and Support Type at Post Level

Armed with linguistic alignment measures and social support classifier, we then evaluates the correlation between linguistic alignment and social support types at the post level. We look for repetition between posts from the initial author and later posts, which are classified as either emotional or informational support. We examine whether lexical and syntactic alignment between support seekers and support providers in the post pairs can be used to predict support type. Do lexical and syntactic alignment act similarly or differently?

6.1 Methods

To examine this question, we fit a generalized mixed effects linear regression model with a binomial kernel to predict the emotional support index of a target post. The covariates of the predicting model include lexical alignment (i.e., LILLA), syntactic alignment (i.e., SILLA) in *logit* space, post distance, and interaction terms. Because different types of social support could be influenced by various topics and authors, we also include these variables grouped by *ThreadID* as random effects.

The generalized mixed effects linear regression model is estimated with the lme4 R package [3]. We then use step-wise Akaike Information Criterion (AIC) [7], a measure of the quality of the current logistic regression model, to select the best model without overfitting. Overall, the full model with all the features has the best performance. Table 3 presents the main effects of covariates, and Table 4 presents the main effects and interaction terms.

6.2 Results

Initially, we focus on the effect of lexical alignment. The regression model (Table 3) shows that the lexical adaptation (LILLA) between support providers and support seekers is a reliable indicator of emotional support in both forums. In other words, messages with more emotional support tend to repeat more words in support seekers' posts. We note that this result may also be interpreted by properties of emotional support in both sub-forums. Emotional support presents understanding and empathy, including similar words from support seekers' posts. Also, the model (Table 3) shows that emotional support index generally increases

Table 3. Predicting Emotional Support Index in posts using linguistic alignment between posts from initial author and other replies. This model only includes main effects.

Predictor	Breast cancer sub-forum			Colorectal cancer sub-forum		
	β	SE	p	β	SE	p
Intercept	1.593	0.065	0.000	1.971	0.072	0.000
Lexical alignment	0.127	0.010	0.000	0.188	0.011	0.000
Syntactic alignment	−0.168	0.013	0.000	−0.180	0.014	0.000
Distance	0.002	0.000	0.000	0.002	0.002	0.001

Table 4. Predicting Emotional Support Index in posts using linguistic alignment between posts from initial author and other replies. The model includes main effects and interactions.

Predictor	Breast cancer sub-forum			Colorectal cancer sub-forum		
	β	SE	p	β	SE	p
Intercept	1.357	0.076	0.000	1.849	0.085	0.000
Lexical alignment	0.080	0.011	0.000	0.171	0.013	0.000
Syntactic alignment	−0.160	0.016	0.000	−0.185	0.016	0.000
Distance	0.021	0.003	0.000	0.015	0.004	0.145
Lexical alignment × distance	0.003	0.000	0.000	0.001	0.001	0.012
Syntactic alignment × distance	−0.0003	0.000	0.676	−0.0004	0.001	0.598

with post distance. I.e., conversations tend to shift to emotional support. Another noteworthy result is the effect of syntactic alignment between messages from support seekers and providers. Less syntactic rule repetition occurs when emotional support is given.

Table 4 adds the interaction terms between the post distances and the measure of lexical and syntactic alignment between posts (Lexical Alignment × Distance and Syntactic Alignment × Distance). Compared to the models in Table 3, the effects and directions of predictors are similar. The effects of lexical alignment and post distance in both datasets show that lexical alignment is also predictive for emotional support as distance increases. However, the effect size is small. Furthermore, the effects of syntactic alignment and post distance is not reliable in either dataset. According to the models, syntactic alignment differs from lexical alignment. We do not see a positive correlation between lexical and syntactic alignment in either dataset.

7 Discussion

In this paper, we showed that two types of support relate to the two lower-level linguistic alignment measurements in different ways in two forums. When peers support one another at the emotional level, they tend to align more with

the support seeker at the lexical level than those dialogues that are primarily providing information support. This may be due to active adaptation, or it may be a consequence of the available lexicon associated with the language register (emotional vs. informational): a smaller set of possible words might imply more overlap. This does not necessarily mean that people tailor their message to an audience (in fact, they show very little adaptation in that respect: [29]). However, it is possible that lexical choice is influenced by a desire to demonstrate empathy when the intent is to provide emotional support.

Importantly, however, this principle does not extend to adaptation in syntactic structure. Support providers align with support seekers at the syntactic level when providing informational as opposed to emotional support. This pattern of effects can potentially be explained by a model of communication that suggests different control over the social-level message at the lexical level versus than at the level of sentence structure. Syntactic adaptation is increased in the informational-task situation, which complements previous corpus-based comparisons that showed increased syntactic priming in task-oriented dialogue compared to non-consequential chat [27]. It also complements recent findings of reduced (perceived) phonetic alignment in the speech of conversation partners that are subjected to increased cognitive load [1].

We believe the findings reported by this paper is important because no previous study has suggested that different types of social support dialogue may be associated with different relationship to the lower-level linguistic alignment phenomena. It not only opens the door of studying potential relationships between the higher level functions of dialogue and lower-level alignments, but also suggests potential directions in which existing alignment theories can be enriched through further studies. For example, can we separate the effect of lower-level linguistic alignments to previous replies on a threaded discussion from the effect of linguistic alignment to the thread initiator (i.e., the support seeker)? If so, is the relationship between support types and lower-level alignment measures still hold?

8 Conclusion and Future Work

Social support in online health communities, especially in the forms of emotional support and informational support, benefits chronically ill patients. This support can come from peers, as our dataset demonstrates, and probably benefits support seekers as well as support providers. That is why we think this study is important in improving health care and well-being of patients.

Our analysis models how users, especially support provider, provide social support by adapting to each other. The pattern of adaptation is interesting, even from a theoretical perspective. We observed reliable lexical adaptation, and relatively smaller syntactic adaptation. Adaptation differs with type of support, with emotional support attracting greater lexical and less syntactic adaptation, and informational support being correlated with the opposite.

This study makes potential theoretical progress revealing the relationship between different levels of linguistic adaptation. The result implies that lexical

and syntactic adaptation have different adaptation levels and are influenced in different ways by the higher-level support function of the dialogue. This difference may be due to different cognitive bases between these two types of adaptation. These results motivate further research and experimentation regarding the mechanism of linguistic adaptation in web-based conversations.

Acknowledgements. This work was supported by a collaborative agreement with American Cancer Society, which made the data of CSN available for this Research. The authors would like to thank K. Portier, G. Greer of the American Cancer Society, current and former members of the Cancer Informatics Initiative at the Pennsylvania State University for useful discussions and comments.

References

1. Abel, J., Babel, M.: Cognitive load reduces perceived linguistic convergence between dyads. Lang. Speech (2016). doi:10.1177/0023830916665652
2. Bambina, A.: Online Social Support: The Interplay of Social Networks and Computer-mediated Communication. Cambria Press, Youngstown (2007)
3. Bates, D., Maechler, M., Bolker, B., Walker, S.: Lme4: linear mixed-effects models using Eigen and S4, R package version 1.1-7 (2014)
4. Bergmann, K., Kopp, S.: Gestural alignment in natural dialogue. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society, pp. 1326–1331 (2012)
5. Biyani, P., Caragea, C., Mitra, P., Yen, J.: Identifying emotional and informational support in online health communities. In: Proceedings of COLING, Dublin, Ireland, pp. 827–836 (2014)
6. Bock, J.K.: Syntactic persistence in language production. Cogn. Psychol. **18**(3), 355–387 (1986)
7. Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, New York (2002)
8. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L.B., Levine, J.M. (eds.) Perspectives on Socially Shared Cognition, pp. 127–149 (1991)
9. Danescu-Niculescu-Mizil, C., Gamon, M., Dumais, S.: Mark my words!: linguistic style accommodation in social media. In: Proceedings of the 20th International Conference on the World Wide Web, pp. 745–754. ACM (2011)
10. Doyle, G., Frank, M.C.: Investigating the sources of linguistic alignment in conversation. In: Proceedings of ACL, pp. 526–536 (2016)
11. Doyle, G., Yurovsky, D., Frank, M.C.: A robust framework for estimating linguistic alignment in twitter conversations. In: Proceedings of the 25th International Conference on World Wide Web, pp. 637–648. ACM (2016)
12. Fox, S.: Peer-to-peer healthcare (2011). http://www.pewinternet.org/2011/02/28/peer-to-peer-health-care-2/
13. Fox, S.: The social life of health information (2011). http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/
14. Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., Tylén, K.: Coming to terms quantifying the benefits of linguistic coordination. Psychol. Sci. **23**(8), 931–939 (2012)
15. Giles, H., Coupland, N., Coupland, J.: Accommodation theory: communication, context, and consequences. In: Contexts of Accommodation: Developments in Applied Sociolinguistics, pp. 1–68 (1991)

16. Gries, S.T.: Syntactic priming: a corpus-based approach. J. Psycholinguist. Res. **34**(4), 365–399 (2005)
17. Healey, P.G., Purver, M., Howes, C.: Divergence in dialogue. PloS One **9**(6), e98598 (2014)
18. Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., Pennebaker, J.W.: Language style matching predicts relationship initiation and stability. Psychol. Sci. **22**(1), 39–44 (2011)
19. Jones, S., Cotterill, R., Dewdney, N., Muir, K., Joinson, A.: Finding Zelig in text: a measure for normalising linguistic accommodation. In: Proceedings of COLING, pp. 455–465. University of Bath (2014)
20. Kaschak, M.P., Kutta, T.J., Jones, J.L.: Structural priming as implicit learning: cumulative priming effects and individual differences. Psychon. Bull. Rev. **18**(6), 1133–1139 (2011)
21. Malecki, C.K., Demaray, M.K.: What type of support do they need? Investigating student adjustment as related to emotional, informational, appraisal, and instrumental support. Sch. Psychol. Q. **18**(3), 231–252 (2003)
22. Niederhoffer, K.G., Pennebaker, J.W.: Linguistic style matching in social interaction. J. Lang. Soc. Psychol. **21**(4), 337–360 (2002)
23. Noble, B., Fernández, R.: Centre stage: how social network position shapes linguistic coordination. In: Proceedings of CMCL, pp. 29–38 (2015)
24. Pickering, M.J., Garrod, S.: The interactive-alignment model: developments and refinements. Behav. Brain Sci. **27**(02), 212–225 (2004)
25. Portier, K., Greer, G.E., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P., Yen, J.: Understanding topics and sentiment in an online cancer survivor community. JNCI Monogr. **2013**(47), 195–198 (2013)
26. Reitter, D., Keller, F., Moore, J.D.: A computational cognitive model of syntactic priming. Cogn. Sci. **35**(4), 587–637 (2011)
27. Reitter, D., Moore, J.D.: Alignment and task success in spoken dialogue. J. Mem. Lang. **76**, 29–46 (2014)
28. Vlahovic, T.A., Wang, Y.C., Kraut, R.E., Levine, J.M.: Support matching and satisfaction in an online breast cancer support community. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1625–1634. ACM (2014)
29. Wang, Y., Reitter, D., Yen, J.: Linguistic adaptation in conversation threads: analyzing alignment in online health communities. In: Proceedings of CMCL, pp. 55–62 (2014)
30. Wang, Y., Yen, J., Reitter, D.: Pragmatic alignment on social support type in health forum conversations. In: Proceedings of CMCL, pp. 9–18 (2015)
31. Wang, Y.C., Kraut, R., Levine, J.M.: To stay or leave?: The relationship of emotional and informational support to commitment in online health support groups. In: Proceedings of CSCW, pp. 833–842. ACM (2012)
32. Xu, Y., Reitter, D.: An evaluation and comparison of linguistic alignment measures. In: Proceedings of CMCL, pp. 58–67 (2015)

Gender Politics in the 2016 U.S. Presidential Election: A Computer Vision Approach

Yu Wang(✉), Yang Feng, and Jiebo Luo

Department of Computer Science, University of Rochester,
Rochester, NY 14627, USA
ywang176@ur.rochester.edu

Abstract. Gender plays an important role in the 2016 U.S. presidential election, especially with Hillary Clinton becoming the first female presidential nominee and Donald Trump being frequently accused of sexism. In this paper, we introduce computer vision to the study of gender politics and present an image-driven method that can measure the effects of gender in an accurate and timely manner. We first collect all the profile images of the candidates' Twitter followers. Then we train a convolutional neural network using images that contain gender labels. Lastly, we classify all the follower and unfollower images. Through a case study of the 'woman card' controversy, we demonstrate how gender is informing the 2016 presidential election. Our framework of analysis can be readily generalized to other case studies and elections.

Keywords: Gender politics · Computer vision · Hillary Clinton · Donald Trump

1 Introduction

Gender has always played an important role in American elections (e.g. Ronald Reagan's re-election in 1984, George W. Bush's election in 2000 [18], Barack Obama's election in 2008 and re-election in 2012 [9]). It is set to play an important role again in the 2016 election cycle. On the Democrats' side, Hillary Clinton has become the first female presidential nominee for a major political party in the U.S. history. On the Republican side, Donald Trump is frequently accused of sexism, with his controversies against Carly Fiorina, Megyn Kelly, Heidi Cruz, and Hillary Clinton. Naturally, being able to measure the effects of gender in an accurate and timely manner becomes crucial.

Recent advances in computer vision [5,12] have made object detection and classification increasingly accurate. In particular, face detection and gender classification [4,8] have both achieved very high accuracy, largely thanks to the adoption of deep learning [13] and the availability of large datasets [7,10,17].

In this paper, we introduce computer vision to the study of gender politics and present an image-driven method to analyze how gender is shaping the 2016

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 35–45, 2017.
DOI: 10.1007/978-3-319-60240-0_4

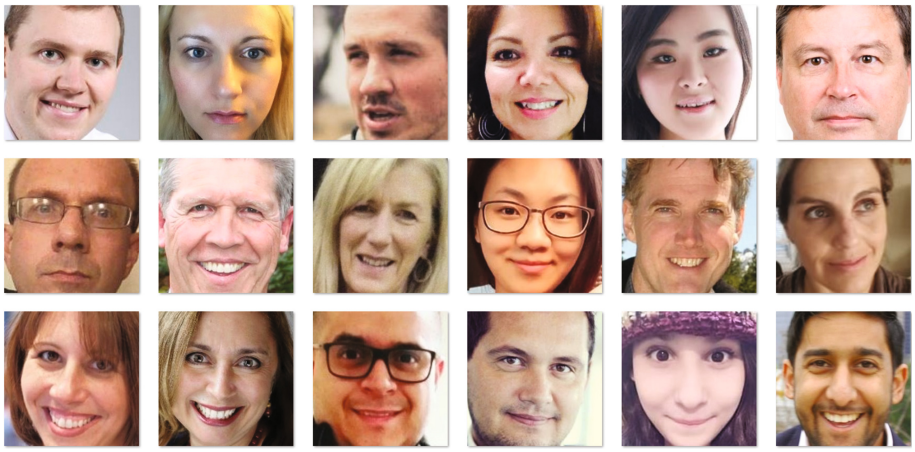


Fig. 1. Top row: Hillary Clinton unfollowers; Middle: Donald Trump followers; Bottom: Bernie Sanders followers.

presidential election. We first collect the profile images of the candidates' followers and unfollowers (Fig. 1).[1] Then we select the images with gender labels to train a convolutional neural network and we use the trained network to classify all the images for Hillary Clinton, Bernie Sanders, and Donald Trump. Lastly, we construct a gender affinity model and test statistically whether or not an event of interest has disturbed the prior gender balance.

To illustrate the effectiveness of our method, we select a case study that carries great importance: the effects of the 'woman card' controversy. The 'woman card' controversy refers to the incident where Trump accused Hillary Clinton of playing the 'woman card' against him. We provide more background information in Sect. 4.

We show that the 'woman card' controversy has made women more likely to follow Hillary Clinton and less likely to unfollow her. Our framework of analysis, which marries gender politics with computer vision, can be readily generalized to study other cases such as Sanders followers jumping ship for Trump and even other elections, such as the French presidential election in 2017.

2 Related Literature

Our work builds on previous literature in electoral studies, data mining, and computer vision.

In eletoral studies, researchers have argued that gender constitutes an important factor in voting behavior. One common observation is that women tend to vote for women, which is usually referred to as gender affinity effect [1,3,11].

[1] By 'unfollower', we mean people who previously followed a candidate on Twitter and later unfollowed him/her.

Another observation is that pre-election polls tend to underestimate support for female candidates [19]. In the 2016 presidential election, Hillary Clinton explicitly portrays herself as a champion "fighting for women's healthcare and paid family leave and equal pay." It is also widely reported that male Sanders supporters are more likely to vote for Trump than female Sanders supporters. Our work will test the strength of this gender affinity effect by constructing a random utility model.

In data mining, there is a burgeoning literature on using social media data to analyze and predict elections. In particular, several studies have explored ways to infer users' preferences. According to [16], tweets with sentiment can potentially serve as votes and substitute traditional polling. [22] exploits the variations in the number of 'likes' of the tweets to infer Trump followers' topic preferences. [14] uses candidates' 'likes' in Facebook to quantify a campaign's success in engaging the public. [21] uses follower growth on the dates of public debate to measure candidates' debate performance. Our work also pays close attention to the number of followers, but we go further by investigating the gender composition of these followers.

Our work ties in with current computer vision research. In this dimension, our work is related to gender classification using facial features. [8] collects 4 million weakly labeled images to train an SVM classifier and has achieved an accuracy of 96.98%. [20] uses user profile images to study and compare the social demographics of Trump followers and Clinton followers.

3 Data and Methodology

3.1 Data

In this section, we describe our dataset *US2016*, the pre-processing procedures and our CNN model. One key variable is *number of followers.* This variable is available for all three candidates and covers the entire period from September 18, 2015 to Oct 19, 2016. Compared with other candidates who have dropped out of the race, the two presidential nominees (and Bernie Sanders) also have the most Twitter followers. This variable is updated every 10 min in our program. In Figs. 2 and 3, we present the cumulative number of Trump followers and Clinton followers and the net follower gain respectively.[2] By recording the follower IDs in a timely manner, we are able to identify the new followers and the unfollowers:

$$\#(\text{net follower gain}) = \#(\text{new followers}) - \#(\text{unfollowers})$$

Besides the number of followers, our dataset *US2016* also contains the detailed follower IDs for Trump, Clinton and Sanders on specific dates, including March 24th, April 17th and May 10th. This information enables us to track the evolution of the election dynamics.

[2] For a detailed analysis of follower growth patterns, see [21].

3.2 Modeling Gender Affinity

We hypothesize that gender-related events have asymmetrical effects on men and women. When such an event occurs, it will disturb the gender balance previously observed. In the context of Twitter following, this can be modeled formally as:

$$\begin{aligned} U_m &= \beta X_m + \lambda_m E + \epsilon \\ U_w &= \beta X_w + \lambda_w E + \epsilon \end{aligned}$$

where X is a vector of static variables related to one's following inclination, such as education, age, and income, λ_m represents the impact of the event E on a man, E denotes the occurrence of an event and is binary, λ_w is the utility impact on a woman, $\epsilon \sim Normal(0, 1)$, and U denotes the utility of following. Individuals will follow a candidate if and only if their utility of following is positive.

This translates into a probability of following for men and women respectively as follows:

$$\begin{aligned} Pr(Y_m = 1) &= \Phi(\beta X_m + \lambda_m E) \\ Pr(Y_w = 1) &= \Phi(\beta X_w + \lambda_w E) \end{aligned}$$

Therefore, the gender distribution of new followers prior to an event is calculated as:

$$\frac{N'_m \Phi(\beta X_m)}{N'_w \Phi(\beta X_w)}$$

where N'_m is the number of prospective male followers and N'_w is the number of prospective female followers for the period before the event. After the event has occurred, the gender distribution of new followers becomes:

$$\frac{N''_m \Phi(\beta X_m + \lambda_m)}{N''_w \Phi(\beta X_w + \lambda_w)}$$

where N''_m is the number of prospective male followers and N''_w is the number of prospective female followers in the period immediately after the event.

Finally, the disturbance in the gender distribution that is attributed to the event is calculated as:

$$D(Event) = \frac{N''_m}{N''_w} \times \frac{\Phi(\beta X_m + \lambda_m)}{\Phi(\beta X_w + \lambda_w)} - \frac{N'_m}{N'_w} \times \frac{\Phi(\beta X_m)}{\Phi(\beta X_w)}$$

3.3 Statistical Testing

For an event that could disproportionately affect women, such as the 'woman card' controversy, we expect that there will be a positive disturbance towards men in the gender distribution among Trump followers and that the disturbance will tilt towards women for Hillary Clinton. The statistical significance of disturbance $D(Event)$ can then be calculated using the two-sample z-test:

$$z = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\hat{p}(1-\hat{p})(1/n_2 + 1/n_1)}}$$

where

$$\hat{p}_2 = \frac{N''_m \times \Phi(\beta X_m + \lambda_m)}{N''_m \times \Phi(\beta X_m + \lambda_m) + N''_w \times \Phi(\beta X_w + \lambda_w)}$$

$$\hat{p}_1 = \frac{N'_m \times \Phi(\beta X_m + \lambda_m)}{N'_m \times \Phi(\beta X_m) + N'_w \times \Phi(\beta X_w)}$$

$$n_2 = N''_m \times \Phi(\beta X_m + \lambda_m) + N''_w \times \Phi(\beta X_w + \lambda_w)$$

$$n_1 = N'_m \times \Phi(\beta X_m) + N'_w \times \Phi(\beta X_w)$$

$$\hat{p} = \frac{n_1 * \hat{p}_1 + n_2 * \hat{p}_2}{n_1 + n_2}$$

With large n_1 and n_2, by the central limit theorem z is approximately standard normal. The null hypothesis is that the event of interest has not disturbed the gender balance among Twitter followers and unfollowers. A large z in absolute terms (2.1 for example) will be strong evidence that there exists a disturbance and that the null hypothesis should be rejected.

3.4 Gender Inference by Computer Vision

We collect the profile images based on follower IDs. Our goal is to infer an individual's gender based on the profile image and to test the hypothesis that gender is affecting the following and unfollowing behavior of the presidential candidates' Twitter followers.

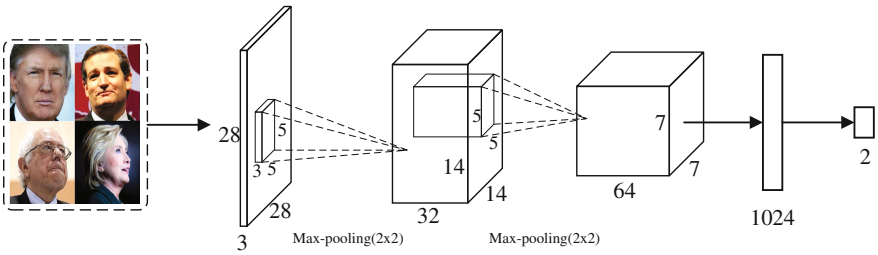


Fig. 2. The CNN model consists of 2 convolution layers, 2 max-pool layers and a fully connected layer.

To process the profile images, we first use OpenCV to identify faces, as the majority of profile images only contain a face.[3] We discard images that do not contain a face and the ones in which OpenCV is not able to detect a face. When multiple faces are available, we choose the largest one. Out of all facial images thus obtained, we select only the large ones. Here we set the threshold to 18kb. This ensures high image quality and also helps remove empty faces. Lastly we resize those images to (28, 28, 3).

To classify profile images, we train a convolutional neural network using 42,554 weakly labeled images, with a gender ratio of 1:1. These images come from Trump's and Clinton's current followers. We infer their labels using the followers' names. For example, David, John, Luke and Michael are male names, and Caroline, Elizabeth, Emily, Isabella and Maria are female names.[4] For validation, we use a manually labeled data set of 1,965 profile images for gender classification. The validation images come from Twitter as well so that we can avoid the cross-domain problem. Moreover, they do not intersect with the training samples as they come exclusively from individuals who unfollowed Hillary Clinton before March 2016.

Table 1. Summary statistics of CNN performance

Architecture	Precision	Recall	F1	Accuracy
2CONV-1FC	91.36	90.05	90.70	90.18

The architecture of our convolutional neural network is illustrated in Fig. 2, and we are able to achieve an accuracy of 90.2%, which is adequate for our task (Table 1).

4 Case Study: The Woman Card

During his victory speech on April 26, 2016, Donald Trump accused Hillary Clinton of playing the 'woman card,' and said that she would be a failed candidate if she were a man. Clinton fired back during her victory speech in Philadelphia and said that "If fighting for women's health care and paid family leave and equal pay is playing the 'woman card,' then deal me in." The 'woman card' subsequently became the meme of the week and its effects are much debated. According to *CNN, New York Times, Washington Post* and *The Financial Times*, this exchange between the two presidential nominees signaled a heated general election clash over gender.[5]

By leveraging the gender classifier and the detailed information on followers, we can easily measure the effects of the 'woman card' exchange on the gender

[3] http://opencv.org.

[4] The full list of label names together with the validation data set and the trained model, is available at the first author's website.

[5] See, for example, http://www.nytimes.com/2016/04/29/us/politics/hillary-clinton-donald-trump-women.html.

composition of new followers and unfollowers for both Hillary Clinton and Donald Trump. Specifically, here we examine whether this exchange has made women more likely to follow Hillay Clinton and more likely to leave Trump.

Our dataset *US2016* contains the detailed IDs of Trump's and Clinton's followers. Specifically for this case study, we are able to use these IDs to identify all the new followers and the unfollowers of Donald Trump first between April 19 and April 26 and then between April 26 and May 1 (Fig. 5). Similarly, we have information on Hillary Clinton's new followers and unfollowers first between April 20 and April 27 and then between April 27 and May 2. This enables us to examine in a definitive manner the gender composition of new followers and unfollowers one week before the 'woman card' exchange (April 26) and one week after. We report the summary statistics in Table 2.

Table 2. Mobility in the Candidates' followers

	Hillary Clinton		Donald Trump	
'Woman Card'	Before	After	Before	After
New followers	72,266	54,820	116,456	115,246
Unfollowers	9,572	8,393	18,376	18,292

New Followers. In Fig. 3, we report on the gender composition of Clinton's new followers one week before the 'woman card' exchange and one week after. We observe a 1.6% increase in percentage of women followers. Our sample size is 14,504 in the first week and in the second 11,147.

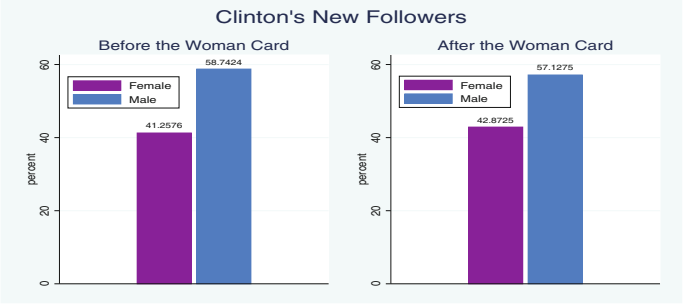


Fig. 3. Gender composition of Hillary Clinton's new followers.

In Fig. 4, we report on the gender composition of Trump's new followers one week before the 'woman card' exchange and one week after. We observe a 0.6717% increase in percentage of women followers. Our sample size is 20,204 in the first week and in the second 21,187. While our main focus is the time-series

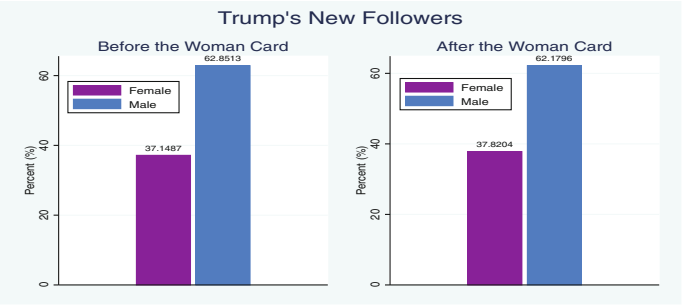


Fig. 4. Gender composition of Donald Trump's new followers.

variations for the candidates, it is interesting to note that across candidates, Clinton attracts more new female followers proportionally than Trump.

Using score test (Table 3), we are able to show that for Clinton the surge of female presence among her new followers is statistically significant. The same does not hold for Donald Trump.

Table 3. New followers' gender composition

Null hypothesis	Clinton		Trump	
	z statistic	p value	z statistic	p value
p_{before}=p_{after}	2.597	0.0093	1.411	0.1582

Unfollowers. In Fig. 5, we report on the gender composition of Clinton's unfollowers one week before the 'woman card' exchange and one week after. We observe a 3.7728% decrease in the percentage of women unfollowers. Our sample size is 2039 in the first week and 1587 in the second.

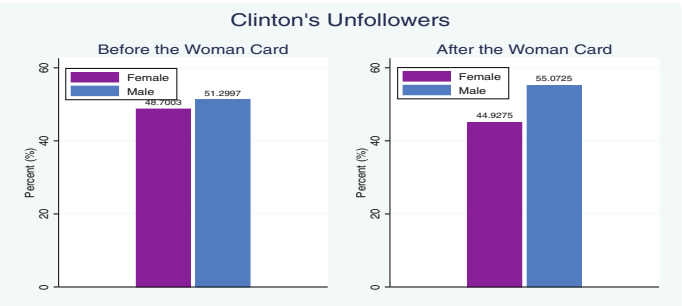


Fig. 5. Gender composition of Clinton's unfollowers.

In Fig. 6, we report on the gender composition of Trump's unfollowers one week before the 'woman card' exchange and one week after. We observe a 0.2786% decrease in the percentage of women unfollowers. Our sample size is 3682 in the first week and 3036 in the second.

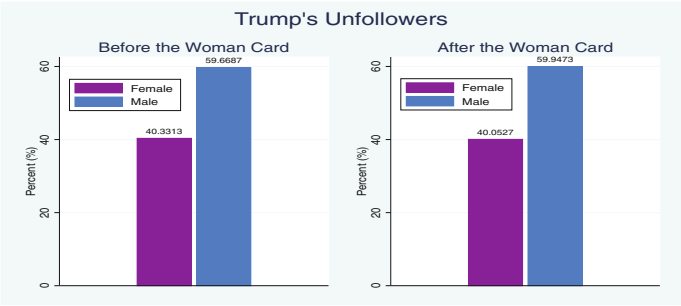


Fig. 6. Gender composition of Trump's unfollowers.

Using score test, we are able to show that for Clinton the decrease of female presence among her unfollowers is statistically significant at 95% confidence interval. While Donald Trump also observes a decrease in the percentage of female unfollowers, the decrease is not statistically significant.

5 Limitations and Future Research

First, our work is built on the assumption that Twitter users, campaign followers in particular, are representative of the demographics of the U.S. population. This assumption may not exactly hold as various demographic dimensions such as gender, race and geography are skewed in Twitter [15]. Second, in our analysis we have deliberately removed empty profile images, as they are not informative with regards to gender. Posting an empty profile image might be correlated to one's following behavior. So both cases could potentially produce selection bias and affect our estimation [6]. Nonetheless, we believe the direction of our estimates will remain consistent, especially if calibrated by reliable polls.

6 Conclusions

Gender has been playing an important role in the U.S. presidential elections. Recent advances in computer vision, on the other hand, have made gender classification increasingly accurate. In this paper we introduced computer vision to the study of gender politics.

We first collected all the profile images of the candidates' Twitter followers. Then we trained a highly accurate convolutional neural network using images

that contain gender labels. Lastly, we classified all the follower and unfollower images. Through the woman card case, we demonstrate how gender has informed the 2016 U.S. presidential election.

Our framework of analysis, which marries gender politics with computer vision, can be readily generalized to study other cases and other elections, such as the upcoming French presidential election. Our study has focused exclusively on images and we have demonstrated its effectiveness. We suggest, however, incorporating text-based analysis (e.g. of user names and tweets) could be beneficial as well [2,23].

References

1. Brians, C.L.: Women for Women? Gender and Party Bias in Voting for Female Candidates, American Politics Research (2005)
2. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (2011)
3. Dolan, K.: Is There a "Gender Affinity Effect" in American Politics? Affect, and Candidate Sex in U.S. House Elections. Political Research Quarterly, Information (2008)
4. Farfade, S.S., Saberian, M., Li, L.-J.: Multi-view face detection using deep convolutional neural networks. In: ICMR (2015)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
6. Heckman, J.J.: Sample selection bias as a specification error. Econometrica **47**(1), 153–161 (1979)
7. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts (2007)
8. Jia, S., Cristianini, N.: Learning to classify gender from four million images. Pattern Recogn. Lett. **58**, 35–41 (2015)
9. Jones, J.M.: Gender gap in 2012 vote is largest in gallup's history (2012). http://www.gallup.com/poll/158588/gender-gap-2012-vote-largest-gallup-history.aspx
10. Ricanek, Jr., K., Tesafaye, T.: Morph: a longitudinal image database of normal adult age-progression. In: 7th International Conference on Automatic Face and Gesture Recognition (FGR 2006) (2006)
11. King, D.C., Matland, R.E.: Sex and the grand old party: an experimental investigation of the effect of candidate sex on support for a republican candidate. Am. Politics Res. **74**(3), 633–640 (2003)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
13. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
14. MacWilliams, M.C.: Forecasting congressional elections using facebook data. PS: Polit. Sci. Politics **48**(04), 579–583 (2015)
15. Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., Rosenquist, J.N.: Understanding the demographics of twitter users. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (2011)

16. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (2010)
17. Phillips, P.J., Wechslerb, H., Huangb, J., Raussa, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image Vis. Comput. **16**(5), 295–306 (1998)
18. Saad, L.: Big gender gap distinguishes election 2000 (2000). http://www.gallup.com/poll/2884/big-gender-gap-distinguishes-election-2000.aspx
19. Stout, C.T., Kline, R.: Political Behavior (2010)
20. Wang, Y., Li, Y., Luo, J.: Deciphering the 2016 U.S. Presidential campaign in the twitter sphere: a comparison of the Trumpists and Clintonists. In: Tenth International AAAI Conference on Web and Social Media (2016)
21. Wang, Y., Luo, J., Niemi, R., Li, Y.: To follow or not to follow: analyzing the growth patterns of the Trumpists on Twitter. In: Workshop Proceedings of the 10th International AAAI Conference on Web and Social Media (2016)
22. Wang, Y., Luo, J., Niemi, R., Li, Y., Hu, T.: Catching fire via 'Likes': inferring topic preferences of trump followers on Twitter. In: Tenth International AAAI Conference on Web and Social Media (2016)
23. Wang, Y., Zhang, X., Luo, J.: When follow is just one click away: Understanding twitter follow behavior in the 2016 U.S. Presidential Election. arXiv:1702.00048 (2017)

Agent-Based Modeling Approach in Understanding Behavior During Disasters: Measuring Response and Rescue in *eBayanihan* Disaster Management Platform

Maria Regina Justina E. Estuar(✉), Rey C. Rodrigueza, John Noel C. Victorino, Marcella Claudette V. Sevilla, Marlene M. De Leon, and John Clifford S. Rosales

Ateneo Social Computing Science Laboratory, Department of Information Systems and Computer Science, Ateneo de Manila University, 1108 Quezon City, Philippines
restuar@ateneo.edu

Abstract. Development of a disaster management system is as complex as the environment it mimics. In 2015, the *eBayanihan* disaster management platform was launched in Metro Manila, Philippines. It is designed to be an integrated multidimensional and multi-platform system that can be used in managing the flow of information during disaster events. Since its development, usage of the system varies depending on the agent who uses the system and which area is affected by what type of disaster. As a complex problem, behavior of disaster agents, such as official responders, volunteers, regular citizens, is best understood if the system can capture, model, and visualize behavior over time. This study presents the development and implementation of an agent-based approach in understanding disaster response and rescue by automatically capturing agent behavior in the *eBayanihan* Disaster Management Platform. All user activities are logged and converted into behavior matrices that can be saved and imported into the Organizational Risk Analyzer (ORA) tool. ORA is used to generate the agent-based model which can be viewed in the *eBayanihan* platform. Actual behavior (ABehM) is compared against perceived (PBM) and expected behavior (EBM) during rescue and response. Results show that EBM networks are fully connected while PBM during rescue and response are granular and vast. Both however show centrality at the provincial and municipal level. ABehM on the other hand shows concentration only at the municipal level with more interactions with ordinary volunteers and citizens.

Keywords: Agent-based models · Behavioral analysis · Disaster informatics

1 Introduction

In 2015, the *eBayanihan* disaster management platform, a government funded project developed at the Ateneo Social Computing Science Laboratory, was

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 46–52, 2017.
DOI: 10.1007/978-3-319-60240-0_5

launched in the Philippines as a free web and mobile tool that will capture and manage the flow of information between and among rescue and response clusters and the public. The development of the platform was highly influenced by the noticeable increase of Twitter posts from citizens during disaster events. As a new source of valuable information, there was a need to provide emergency response clusters from the National level down to the Local Government level with an organized view of verified social media reports. Aside from social media data, there was also a need to create a public platform that can be used the general public in reporting real time disaster related incidents.

The end goal of *eBayanihan* is to provide a top down and bottom up coordinated communication platform for information sharing during emergency events between rescue and response agencies as well as with the public. Registered users from response clusters as well as the public provide basic demographic information including skills and resources that can be volunteered or performed during emergency events. Each user is also required to select the most appropriate role based on the community organizational chart. The idea behind the role assignment is to be able to keep track of actual behavior during simulated and actual disaster events.

However, understanding the behavior of disaster agents, namely: response cluster agencies, volunteers as well as the public, require capturing movements and interactions between and among users of the system. (*eBayanihan*) captures user activities and interactions which can automatically be converted to behavior matrices. The system makes use of Organizational Risk Analyzer (ORA), a meta-network organizational analysis tool developed at the Center for Computational Analysis of Social and Organizational Systems (CASOS), as an extension to the *eBayanihan* platform. System logs can be automatically translated to behavior matrices that are uploaded in ORA for social network analysis. Results generated in ORA can be viewed in the (*eBayanihan*) platform. Having this feature allows for the understanding of the complexity of behavior of agents during disaster events. The system allows for the visualization of social network behavior: who interacts with whom, who is knowledgeable on specific tasks, who performs what task, in relation to disaster management and mitigation. Actual behavior is compared to expected and perceived behavior to gain a better understanding of which coordination activities are relevant or become deterrent to saving lives.

2 Literature Review

2.1 Agent-Based Modeling in Understanding Behaviors During Disaster and Emergency Events

Agent-based modeling (ABM) is an approach to evaluate complex systems where independent and interacting agents make up its domain [6]. As examples, ABM was used to simulate crowd evacuation [4,5,7], to aid flood management incident [2], for evaluating positioning during outbreaks [3], earthquake and tsunami

evacuation simulation [4,5], preventing fire and flood [7] and for disaster management [1]. ABM is also used in studies that focused on human behavior during earthquake [4] and tsunami [5].

2.2 Organizational Chart of the Philippine Rescue and Response Cluster

The Philippines makes use of a decentralized approach in the management of disaster and emergency events. At the regional and national level, the Office of Civil Defense (OCD) has formed the National Disaster Risk Reduction and Management Council (NDRRMC). At the provincial level, the elected Governor provides a team under the Provincial Disaster Risk Reduction and Management Council (PDRRMC). At the local level, Local Government Units (LGUs) led by the elected Mayor forms its own Municipal Disaster Risk Reduction and Management Council (MDRRMC). Following the 2005 Humanitarian Reform Agenda of the United Nations (UN), the formation of the council members at each level all follow the cluster approach allowing for coordination and management of humanitarian needs including: food security, health, shelter, telecommunications, rescue and response, to name a few.

2.3 *eBayanihan* as a Social Networking Platform for Disaster Management

eBayanihan stems from the Filipino root word *bayan*, which means country. The word *bayanihan* is a social representation of the collective helping behavior of Filipinos where an individual or a family in need of assistance becomes a concern of the community. The community then crowdsources and provides possible solutions. The *eBayanihan* disaster management system was therefore designed as an online social networking platform to allow crowdsourcing of information from the public as well as provide avenue to manage response and rescue by line agencies during disaster events.

As part of the user registry, *eBayanihan* requires a user to select a role from a list that represents each cluster role based on duties and responsibilities assigned during disaster events as well as a volunteer or an ordinary citizen for users that do not have official roles. Table 1 shows an aggregated listing of roles and corresponding descriptions.

eBayanihan includes features for viewing reports by location through a timeline which can be filtered by location or viewed on a map. Community leaders are recommended to submit community profiles. Data in this form are used for risk assessment and serve as guide for understanding community-based protocols during disaster events. Official volunteers, such as shelter managers, can profile and update status of evacuation sites in real-time. Lastly, ordinary citizens can submit real-time reports on disaster incidents as they experience it.

Table 1. User roles and descriptions

User role	Description
PDRRMC and MDRRMC	Oversees rescue and response, submits reports
Barangay DRRM Council (BDRRMC)	Produces barangay profile, views reports and requests
Line Agencies (Police, Fire, Health)	Safety, security, response and rescue
Official Volunteers	Shelter, response and rescue
Ordinary Volunteers	Report incidents, verify and validate incidents

3 Methodology

An *eBayanihan* training session was conducted to capture actual behavior of disaster agents during a simulated disaster event. Participants were grouped according to designated work titles. The study compares interactions of individual agents in three ways: (1) as it is perceived by the users, (2) as expected based on mandatory disaster protocols and (3) actual behavior captured and measured by the system. The Perceived Behavior Model (PBM) is generated by creating a behavior matrix from results of a survey asking two questions: *Who do you work with before/during a disaster?* which captures Agent × Agent Network and *What tasks are assigned to you?* which captures Agent × Task. The Expected Model (EBM) is constructed by generating a behavior matrix from the roles and responsibilities stated in the standard local disaster management contingency protocols. The Actual Behavior Model (ABehM) is obtained from the behavior matrix that is generated by *eBayanihan*. Matrix comparison and standard network analysis were performed on the datasets.

4 Results

The Expected Behavioral Model (EBM) during a disaster show that most of the agents interact directly with one another showing evidence of a strongly connected network as seen in Fig. 1.

As seen in Fig. 2, PBM shows an expanded view detailing coordination between and among LGUs and Non-Government Organizations (NGOs), official and ordinary volunteers. Interactions between and among agents significantly increase. Government agencies receive more requests for services.

Figure 3 shows the ABehM interaction between and among agents as captured by the *eBayanihan* system during a simulated earthquake event. It shows that there are lesser agents than EBM which indicates lesser use of the system than expected.

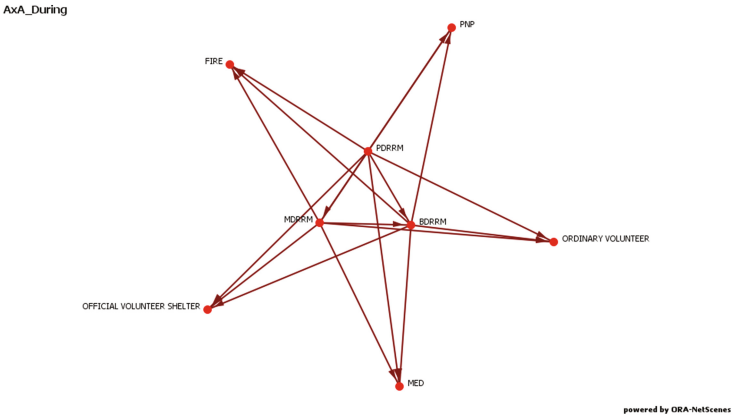


Fig. 1. Expected A × A Behavior Model (EBM) during disaster

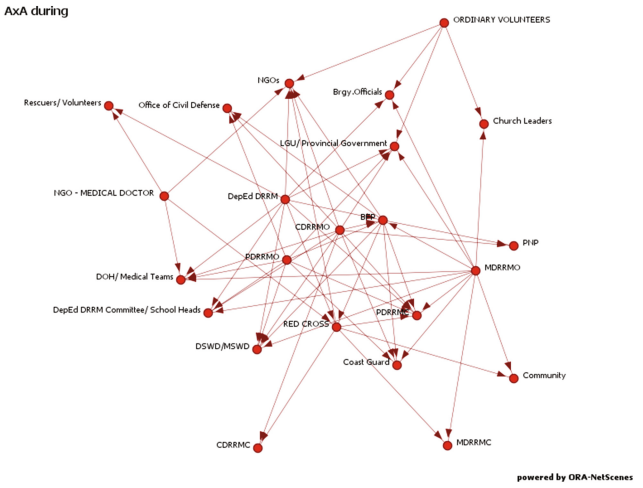


Fig. 2. Perceived A × A Behavior Model (PBM) during disaster

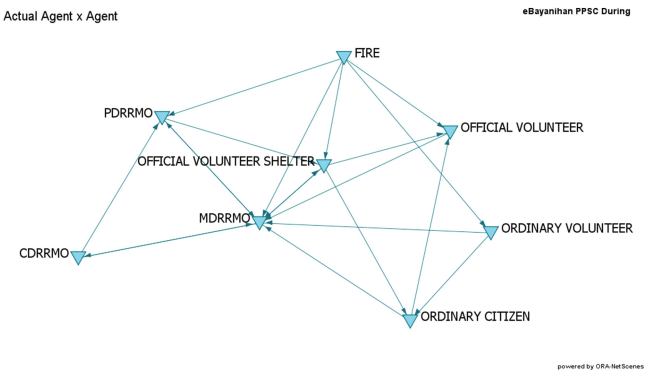


Fig. 3. Actual A × A Behavior Model (ABehM) during disaster

5 Conclusion

This study contributes findings in understanding behavior during rescue and response by capturing perceived behavior (PBM), expected behavior (EBM) and actual behavior (ABehM) in a disaster information management platform. Expected rescue and response behavior of agents shows a fully connected graph with LGUs at the heart of the operations, consistent to the devolved and cluster approach. However, the EBM does not depict the granularity of coordination that is perceived by disaster agents. Actual behavior shows that not all perceived activities are actually implemented during simulated disaster events. There is a need therefore to revisit disaster coordination protocols to ensure that expected behavior, perceive behavior and actual behavior remain consistent during the different phases of disaster, most especially during critical operations of rescue and response.

Acknowledgments. Acknowledgment is given to Philippine Council for Industry, Energy and Emerging Technologies Research and Development (PCIEERD), Department of Science and Technology (DOST), for funding this research.

References

1. Altay, N., Pal, R.: Information diffusion among agents: implications for humanitarian operations. Prod. Oper. Manage. **23**(6), 1015–1027 (2014)
2. Dawson, R.J., Peppe, R., Wang, M.: An agent-based model for risk-based flood incident management. Nat. Hazards **59**(1), 167–189 (2011)
3. De Leon, M.M., Estuar, M.R.E.: Rapid application development of ebayanihan patroller: a crowdsourcing sms service and web visualization disaster reporting system. In: International Conference on IT Convergence and Security (ICITCS), pp. 1–4. IEEE (2014)
4. D'Orazio, M., Spalazzi, L., Quagliarini, E., Bernardini, G.: Agent-based model for earthquake pedestrians' evacuation in urban outdoor scenarios: behavioural patterns definition and evacuation paths choice. Saf. Sci. **62**, 450–465 (2014)

5. Erick, M., Suppasri, A., Imamura, F., Koshimura, S.: Agent-based simulation of the 2011 great East Japan Earthquake/Tsunami evacuation: an integrated model of tsunami inundation and evacuation. J. Nat. Disaster Sc. **34**(1), 41–57 (2012)
6. Macal, C.M., North, M.J.: Tutorial on agent-based modelling and simulation. J. Simul. **4**(3), 151–162 (2010)
7. Wagner, N., Agrawal, V.: An agent-based simulation system for concert venue crowd evacuation modeling in the presence of a fire disaster. Expert Syst. Appl. **41**(6), 2807–2815 (2014)

An Agent-Based Model of Posting Behavior During Times of Societal Unrest

Krishna C. Bathina[1], Aruna Jammalamadaka[2(✉)], Jiejun Xu[2], and Tsai-Ching Lu[2]

[1] Indiana University, Bloomington, IN 47408, USA
bathina@indiana.edu
[2] HRL Laboratories, LLC, Malibu, CA 90265, USA
{ajammalamadaka,jxu,tlu}@hrl.com

Abstract. Social media is increasingly monitored during periods of societal unrest to gauge public response and estimate the duration and severity of related protest events. To this end, we build an agent-based simulation model that accurately describes the shift in posting behavior of users as related to a real historical event. First we define an appropriate indication that an agent has become an "activist", or someone who disseminates protest-related posts during times of unrest. We then build an agent-based model based on parameters estimated from before and during the protest. We validate our model using a complete collection of Tumblr data from six months prior to the Ferguson protest of 2014, until the state of emergency was lifted. Validation is performed by visual inspection of the similarity of simulated distributions of established emergent metrics to the empirically observed data. Our results show that our model has potential for predicting posting behavior during future protests.

Keywords: Agent based model · Information diffusion · Social media · Political unrest · Tumblr

1 Introduction

Agent-based models (ABMs) are computational models with autonomous agents, an environment, and mechanistic behaviors that can be used to represent and simulate emergent behavior from complex, non-linear mathematical systems [1]. In this paper, we focus on building an ABM that accurately models social media behavior during an actual protest by switching rules for agents during the protest depending on whether or not they become an "activist". A cartoon representation of our model is shown in Fig. 1. In the upper left corner, an agent (i.e., an online social media user) chooses to adapt a meme #4 from its neighbor. Subsequently, in the upper right corner, the same agent updates its memory so that

Krishna C. Bathina: This work was conducted while the first author was doing an internship at HRL Laboratories, LLC.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 53–59, 2017.
DOI: 10.1007/978-3-319-60240-0_6

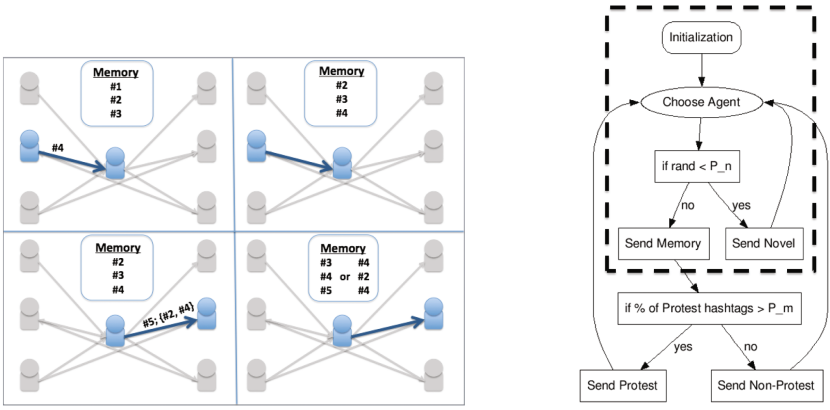


Fig. 1. The cartoon, on the left, is a visualization of the agent-based model for online social media users before a protest event. The flowchart, on the right, represents the *During* model. The *Before* model is shown in the dashed box. Both algorithms are described in the text.

the oldest meme is removed and #4 is added. In the lower left corner, the agent can choose to post either a novel meme (e.g., #5) or chose existing memes from memory (e.g., #2, #4). Finally, in the bottom right quadrant, similar to above, the memory of the agent is updated, either with #5 or with #2 and #4. Such a diffusion style is inspired by the work in [5]. We choose the Ferguson protest of 2014 as the topic of our case study due to its national popularity.

2 Data Collection

We used data from Tumblr, a popular microblogging social network, for our analysis. Users can upload blog posts, unilaterally follow other users, and re-blog content. Typical social media APIs only allow for a partial data collection, or implement waiting times that make the collection of a complete dataset very difficult. Our case study differs in that we have the full dataset from 2012 to 2014, and thus our model can be accurately validated.

For our experiment, we first obtained a list of memes such as "justiceformike-brown", "handsupdontshoot", and "fergusonshooting" that were in support of the Ferguson protest (for a full list of the memes used please refer to [4]). We then defined two time periods in our dataset and for our model: May to August 8th as *Before* the protest and August 9th to September 3rd as *During* the protest. *Before* corresponds to three months before the protest (used for modeling non-protest behavior) and *During* corresponds to the period from the killing of Michael Brown to the day the national emergency was lifted (for modeling protest behavior). We then found all users that used any of the above protests memes at least once in *During*. Once the user population was collected, we extracted all of their posts and re-blogs from both *Before* and *During*. We label

all memes that are not in the list above as *non-protest memes* in the context of this study. We also extracted all posts and re-blogs for the same time periods from 10,000 random users that never used one of the above protest memes as a control group, in order to first test if there is a statistical difference in the behavior of this group and the protest-meme-using group. In total, we extracted 220 million posts and 764 million memes. During the protest, about 1.7% of the posts and 2.1% of the tags were about the Ferguson protest. From this dataset, we are able to extract the full re-blog network and analyze every blog and re-blog. This network consists of 413,867 nodes and around 23 million total edges.

2.1 Preliminary Analysis

To describe our data, we use the four emergent metrics for quantifying social media behavior shown in Table 1. The Meme Time, Meme Popularity, User Entropy, and User Attention are averaged over only days that had posts; days without any posts were ignored.

Table 1. Emergent metrics for validation, from [5].

Metric	Definition
Meme Time	Longest consecutive number of days a meme was posted
Meme Popularity	Average number of posts of a meme per day
User Entropy	Average entropy of the memes posted by a given user per day
User Attention	The average number of re-blogs per user per day

In order to determine whether a change in ABM rules during the protest was needed, we performed four different preliminary statistical analyses, results of which are shown in Table 2. Protesters are individuals who posted at least one protest meme during the time period of the study. Non-protesters, which are only used for analysis A, were chosen by finding all users that did not use any of the protest hashtags, followed by randomly sampling a set of 10,000 users to prevent any bias. $\Delta\widetilde{x}$ is the difference in median between both groups, or the effect size, and Z is the test statistic from the Kolmogorov-Smirnov test. The *Meme Time* was normalized over the total number of days, allowing the continuous assumptions of the KS test to hold. P-values are not reported because $-\log p > 30$ for each test, and thus were significant. For most comparisons, the $\Delta\widetilde{x}$ were very small. Analyses C and D, on the other hand, show that the protest memes had much more *Popularity* than all non-protest memes. Also, Analyses A and B show that the *User Attention* for both non-protesters and protesters before the protest was larger than protesters during the protest. Our results show that a difference exists between all compared groups, especially the *Popularity* of memes and the *Attention* of users.

3 Model Description

Our model, built using Python, is meant to mimic the natural posting patterns and influence of connected users during protest and non-protest periods. The

Table 2. Preliminary analyses showing statistical differences in posting behavior.

Comparison	Popularity		Time		Attention		Entropy	
	$\Delta\widetilde{x}$	Z	$\Delta\widetilde{x}$	Z	$\Delta\widetilde{x}$	Z	$\Delta\widetilde{x}$	Z
A. Non-protesters during vs protesters during	−0.19	0.02	0.07	0.10	−12.79	0.28	−2.14	0.31
B. Protesters before vs protesters during	−0.05	0.03	−0.03	0.89	−10.65	0.01	−1.63	0.02
C. Non-protest memes during vs protest memes during	2.35	0.78	0.52	0.89				
D. Non-protest memes before vs protest memes during	12.95	0.82	0.47	0.89				

model consists of Tumblr users as agents and the re-blog network as their environment, where directed edges represent the *flow* of memes. Each time step in our model represents one day. The total number of posts in the *Before* and *During* simulation periods are equal to the observed total number of posts during those periods, and an equal number of posts occur on each day of the simulation. Agents have a finite-sized *Memory* that contains a list of memes with repetitions. The memory is finite to model the limited attention that is evident among social media users [3]. If new memes are added to the memory, the oldest memes are removed from the list, representing the discovery that the number of memes to which a user can pay attention is bound, and therefore the injection and survival of new memes comes at the expense of others [5]. In line with this work, we utilize the following five model parameters:

- $\mathbf{P_n}$: probability of posting a novel meme
- $\mathbf{P_r}$: probability of posting multiple memes per post before the protest
- $\mathbf{P_{rn}}$: probability of posting multiple non-protest memes during the protest
- $\mathbf{P_{rp}}$: probability of posting multiple protest memes during the protest
- $\mathbf{P_m}$: proportion of protest memes needed in memory to post about the protest

The novel aspects of our model come from splitting the model into two time periods; *Before* and *During* the protest. At initialization, the largest connected component (containing 412,803 nodes) of the re-blog network for the data is loaded into the model. The agents' memories are then loaded with random hashtags. At each iteration, an agent is chosen to post with a probability proportional to their out-degree [2]. This agent then either posts a novel hashtag with probability P_n, or posts a set of hashtags from memory. If the agent is posting from memory, each hashtag in memory is added to the post with probability P_r. After every post, the agent's memory, along with the memories of its neighbors are updated with the posted memes as shown in Fig. 1. The *During* model is initialized with the agent attributes and network from the end of the *Before* model. An initial number of agents, equal to the number of actual protestors on the first day of the protest, are randomly chosen as protesters, and protest memes with frequencies

proportional to the observed counts on the first day, are added to their memory. The model itself is identical to the *Before* model until an agent chooses to post from memory. If the percentage of protest memes in their memory is greater than P_m, the agent has become an *activist*, and consequently, this agent posts only protest memes. Each protest meme is chosen with probability P_{rp} with each post containing at least one meme. If the percentage in memory is not greater than P_m, the agents posts only non-protest memes, each with probability P_{rn}. To clarify, all agents in our model are protesters; they become activists once more than P_m percent of their memory is filled with protest memes. Again, after every post, the agent's memory along with the memories of its neighbors are updated accordingly. All parameters are found empirically. P_n $(= 0.2657)$ is calculated by finding the average number of posts with a new meme per unit time (day). The P_r parameter family is calculated by the average number of memes per post divided by the length of the agent's *Memory*. P_r (0.2624) represents the average number of memes per post before the protest, while P_{rp} (0.3145) represents the average number of memes per post during the protest for posts that include protest memes and similarly, P_{rn} (0.2622) is calculated by the average number of memes per post during the protest for posts that do not include protest memes. P_m (0.6) and the size of the *Memory* (10) are tunable parameters.

4 Results and Discussion

Overall, our ABM metrics show that results from the *Before* and *During* model are quite similar to empirical results from observed data. In this section we choose to focus on the emergent results from the *During* model because it is the major part of our contribution. These results are shown in the normalized histograms of Fig. 2. Figure 2a shows that the *User Attention* from our model did not match that from data as well as we expected. Our model shows a linear distribution because of our proportionality assumption. However, the results indicate that the number of Tumblr users with a moderate average number of posts per day are higher than we expected. Even with this mismatch, we believe our assumption is reasonable based on previous studies, and we hesitate to overfit our model by incorporating the observed data. The *Entropy* (Fig. 2b), of the model did match the data with a slight increase to a peak around 1.0, and then a rapid decrease afterwards. This suggests that most users tended to post with very little variety per day. In the model, *Entropy* is a factor of the rate of novel memes, protest memes, and non-protest memes. Increasing the rate of novel memes, P_n, would increase the average user entropy while increasing the rate of protest and non-protest memes would decrease the entropy. Figure 2c shows that although the model and data distributions have similar shapes, the model tended to overestimate the *Meme Popularity*. This is most likely due to our posting behavior assumption since *Meme Popularity* is a function of what memes are posted, and thus, which users are posting. But, with such a low difference in probabilities and a similar distribution shape, we believe the deviances are reasonable. The *Meme Time* for the model and data in Fig. 2d show similar

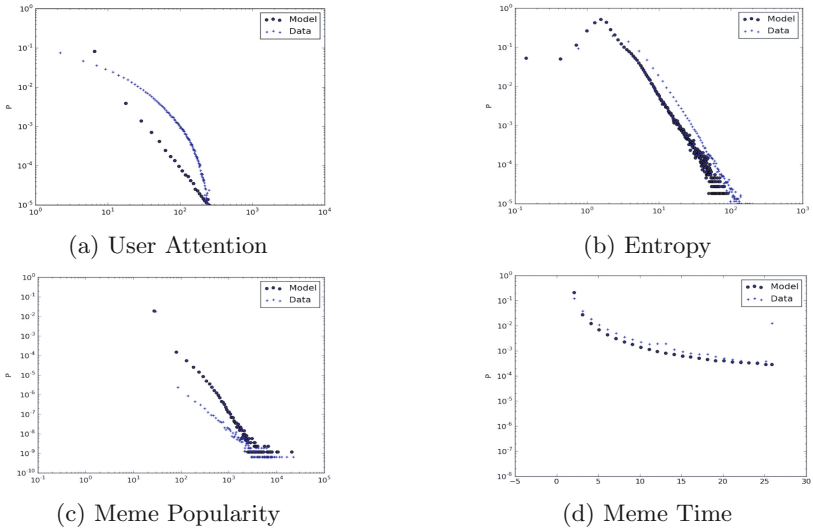


Fig. 2. The above plots show the comparison between model results and observed data via normalized histograms of the defined metrics. All plots are shown on a log-log scale except for (d) Meme Time, which is shown on a linear-log scale.

distributions, with both flattening out as the time increases, suggesting that the majority of Tumblr hashtags are not re-blogged. *Meme Time* is a function of the re-blog parameters, P_r, P_{rp}, P_{rn}; increasing their values would cause an increase in the lifetime of the meme.

The results from the *During* model in Fig. 2 may look very similar to the data simply due to a large proportion of non-protest memes. Therefore, to capture the true effect of the model, we also computed unnormalized histograms for only the protest memes. We found similar behavior between *Meme Times* in the "During" model and the data, with the model tending to slightly underestimate the times. Similarly, the *Meme Popularity* showed that the shape of the model results and data distributions match well, but the model tends to overestimate the popularity by about a factor of 10. Overall, the difference in model and empirical results are small, therefore we believe that our model successfully and accurately describes the full Tumblr dataset.

In summary, we built a new ABM which uses a real protest event to represent a radical change in behavior of a sub-population of the agents. We then validated our model empirically by analyzing Tumblr data during the Ferguson protest of 2014. We acknowledge that this is an empirical study, and validation on an entirely new protest dataset is required in order for the model to be proven usable for prediction and simulation of future events. We chose not to perform cross-validation by sub-sampling the network used in this study, opting instead for the more realistic analysis of posting interactions that ensues from analyzing the full Tumblr re-blog network. However, we believe that our model, and the

extensions of it described below, can still be useful in quantifying and simulation social media posting behavior during times of protest.

References

1. Bruch, E., Atwell, J.: Agent-based models in empirical social research. Sociol. Methods Res. **44**(2), 186–221 (2015)
2. Chung, F., Lu, L.: The average distances in random graphs with given expected degrees. Proc. Natl. Acad. Sci. **99**(25), 15879–15882 (2002)
3. Hodas, N.O.: How limited visibility and divided attention constrain social contagion. In: SocialCom, Citeseer (2012)
4. Jules, B.: Hashtags of ferguson (2014). https://medium.com/on-archivy/hashtags-of-ferguson-8f52a0aced87.vbia14pwu
5. Weng, L., et al.: Competition among memes in a world with limited attention. Sci. Rep. **2**, 1–8 (2012). doi:10.1038/srep00335

‘They All Look the Same to Me.’ An Agent Based Simulation of Out-Group Homogeneity

Ansgar E. Depping(✉), Nathaniel Osgood, and Kurt Kreuger

Department of Computer Science, University of Saskatchewan,
176 Thorvaldson Bldg, 110 Science Place, Saskatoon, Canada
ansgar.depping@usask.ca

Abstract. Group memberships can dramatically affect the way people perceive each other. One of the effects group membership elicits is called out-group homogeneity. It is the tendency to judge members of out-groups as more similar to one another than in-group members. Research on out-group homogeneity faces some challenges that are difficult to overcome in experimental or field settings. We propose that simulation models can help us further understand these principles and test hypotheses that are difficult to test in classical research settings. Our model simulates trust developments using a classic social dilemma based on the prisoner’s dilemma. The patterns that emerge in our model are coherent with what literature would suggest. We also shed some light onto so far unexplored territory such as the longitudinal analysis of trust development in the context of group perceptions. We further discuss limitations and possible future directions.

Keywords: Out-group homogeneity · Prisoner’s dilemma

1 Background

A large area of social psychology deals with the effects of group membership, with decades of research showing that dividing people into groups dramatically affects social perception and behavior. A commonly researched effect associated with group membership is out-group homogeneity (OGH) [5]. Tajfel & Turner conceptualize it as the asymmetrical accentuation of intragroup similarities in favor of OGH [7]. This means that people judge members of outgroups as more similar to one another than they do members of their in-group. To understand how animosities between groups commence, evolve and persist, out-group biases have been heavily researched [5, 7]. However, the research on out-group homogeneity, being mostly limited to experimental and field studies, is facing major difficulties.

First, research on OGH has mostly investigated the relationship between individuals and their attitudes toward out-groups [5]. It is assumed that these biases will affect group interactions on a population level. However, with current research methods the effect OGH has on a societal level remain unclear. Second, while OGH can be observed, it is difficult if not impossible to meaningfully manipulate group perception for research purposes. Presently, researchers cannot experimentally look at group dynamics with OGH ‘on’ or ‘off’. Third, lab experiments or field questionnaires can

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 60–64, 2017.
DOI: 10.1007/978-3-319-60240-0_7

only ever get a temporary snapshot of how OGH relates to other constructs. What is missing is a longitudinal view at how group perceptions shape inter-group relationships over time. In the present study we propose simulation models as a viable research tool to overcome these challenges.

2 Model Description

2.1 Model Scope and Architecture

We simulated a population of agents who repeatedly interact. We utilized a classic investment dilemma in which two agents of a population repeatedly 'trade' with each other. In these interactions, the agents can cooperate or compete [2]. We draw from previous literature on trust formation and assume that cooperation is determined by the perceived trustworthiness of the trade partner; agents will only cooperate if they anticipate cooperation from their partner [3, 6]. Inter-agent trust is tracked over time.

To introduce the idea of OGH we implemented three features. First, agents are given a group membership, and were randomly assigned to group A or B. Second, each agent has a generalized idea of how trustworthy the out-group is, implementing the belief that out-group members behave similarly. Third, this belief is used to affect future behavior with out-group agents.

Using the simulation program Anylogic, [1] we adopted a simulation model on group interactions [4]. The simulation creates an environment with a population of 100 agents. The agents engage in randomly assigned 'trades' with other agents. Each agent waits on average one unit of time before starting the next trade. The more agent A trusts agent B, the higher the chance that agent A cooperates and vice versa. The perceived trustworthiness of the partner is a value between -100 for complete distrust and 100 for complete trust. The model assumes an inherent tendency towards initial trust; when perceived trustworthiness is at its neutral position of 0, the chance of cooperation is 60%. The agent-to-agent trust matrix, S, contains all trust perceptions; agent A's trust value of agent B is S_{ab}, being updated after each trade. If agent B cooperates, A increases S_{ab} by the value of 10 (trust build). If B competes, A adjusts S_{ab} by the value of -15 (trust break). Trust break is greater than trust build to simulate the assumption that trust builds slowly and breaks easily. In the beginning of each simulation the initial trust values are drawn from a uniform distribution between -50 and 50 $(\mu = 0)$, simulating the findings that people make naïve assumptions about someone's trustworthiness even without and personal history [6].

Simulating Out-Group Homogeneity. The first step to conceptualizing OGH in trust-requiring interactions is to assign groups. The population was split into two equal groups. The second step was to make agents aware of group membership and have them consider group membership into their decision process. Specifically we want to implement the heuristic of out-group homogeneity in the decision process. Given that agents A and B are in different groups and they engage in a trade, in addition to the inter-individual trust between agent A for B (S_{ab}), agent A also includes the average experience they had with every member of agent B's group (*Group perception,* G_{ab}). Including the average trust towards a group when deciding how to interact with a

specific group member represents the assumption that agent B will behave like other agents from their group. How heavily the group perception weighs into the decision process is determined by the value *groupSalience* (ranging between 0 and 1). The overall perceived trustworthiness (T) can be represented as:

$$T = (1 - groupSalience) * S_{ab} + groupSalience * G_{ab}$$

3 Experimental Runs

The simulation ran multiple times with different values for *groupSalience*. For each parameterization, 20 replications were run. Each realization ran for 500,000 time steps.

As an indicator for intergroup trust, we measured the ratio of trusting relationships (trust > 0) between groups over all relationships between groups. The positive relationship score (*PR*) is this ratio, where a value of 1 means every between-group-relationship is positive and 0 meaning every between-group-relationship is negative.

We also present a visual representation of *S*, with the columns representing the trusters, and the rows the trustees. The possible range of values from −100 to 100 is represented by the scale form black (distrust) to white (trust).

3.1 A World Without Out-Group Homogeneity

Figure 1a shows the S matrix when groupSalience is 0. All agents make trust decisions purely on their individual trade history with their partner. The example shows that roughly half of squares are full white and the other half are full black. There are no grey cells. This is coherent with the self-reinforcing nature of trust observed in literature [6]. Positive experiences increase the chance for cooperative behavior in the future and vice versa. Over the course of these prolonged relationships, trust develops into either full trust or full distrust.

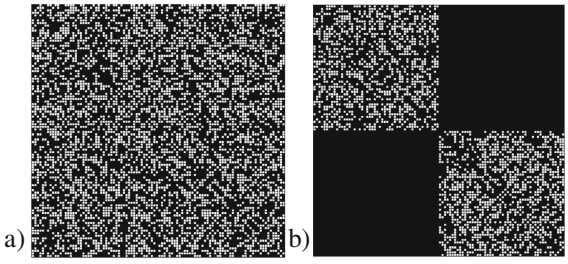


Fig. 1. (a) S-matrix with groupSalience = 0 (Group A: agents 1–50, Group B: agents 51–100). (b) S-matrix with groupSalience = 0.5. (Group A: agents 1–50, Group B: agents 51–100)

3.2 Introducing Out-Group Homogeneity

Figure 1b shows a matrix for a realization with groupSalience at 0.5 (50%). All agents perceive every outgroup-agent as completely untrustworthy. The perception of in-group members remains the same. To get a closer look at the effect of OGH on trust formation we present the PR values over time given different groupSalience values (see Fig. 2). We ran our model, varying the groupSalience in 5% iterations from 50% down to 10%. Each scenario was run 20 times. The scores were averaged. The chart shows how quickly the PR ratio drops depending on the groupSalience. The decline towards complete distrust between the groups strongly depends on group salience.

Due to the longitudinal view on intergroup relations we can also observe how the decline towards group animosity is not linear but rather exponential with a strong decline in the beginning of our simulations. The tilt towards distrust is especially strong in the beginning of the simulation in the absence of any personal history where decisions are only informed by a generalized group perception. The stronger this group perception weighs in, the stronger the initial burst of distrust. Intergroup relations appear to be especially vulnerable to OGH effects in early interactions.

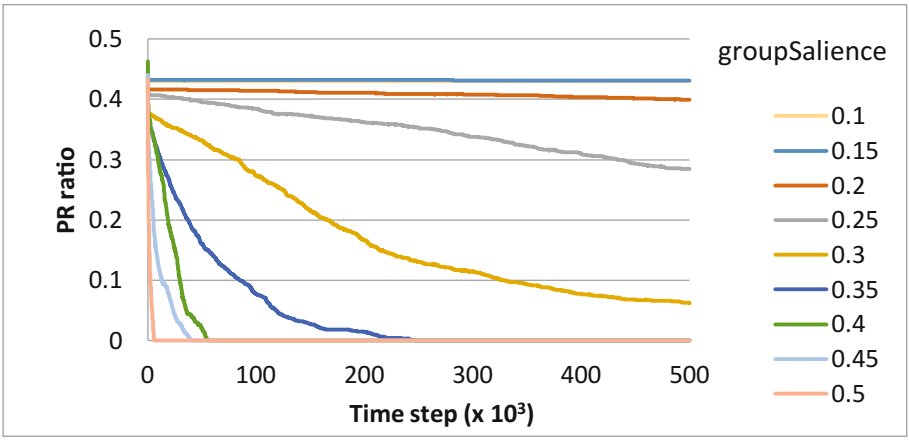


Fig. 2. PR scores over time for different groupSalience values.

4 Discussion and Conclusion

We demonstrated a simple individual behavioral feature that leads to the emergence of a population-wide phenomenon widely seen in human populations. We were able to address some of the challenges face by traditional research methods on OGH. By giving agents only the expectation that out-group members will behave similarly, we have drastically changed population-wide dynamics with mostly devastating effects on group segregation and intergroup trust. We experimentally varied the salience of OGH resulting in expected differences in outcomes. Our approach also revealed a nonlinear development of intergroup relations indicating a sensitivity towards OGH in the absence of personal history.

Our work has some obvious limitations, the most obvious of which are the simplifying assumptions made. Notably our model does not include other well documented stereotyping mechanisms such as in-group favoritism. Similarly, we assume group perception to be based solely on personal experience, avoiding descriptions of human cultural values. Both points can be further investigated in variations of the present model. Secondly, while the model behavior matches qualitatively the behavior of human populations, it is too stylized to be able to validate against real human data. Nonetheless, the insights gleaned from this work could direct future human studies. We do not argue simulation models to be a superior method, but rather a complementary, even synergistic, strategy to understand complex social phenomena. Effects predicted in simulation models need to be corroborated by data, generated perhaps from field studies.

References

1. AnyLogic Software, version 7.0.1. http://www.anylogic.com
2. Axelrod, R.: The evolution of strategies in the iterated prisoner's dilemma. In: The Dynamics of Norms, pp. 1–16 (1987)
3. Dirks, K.T., Ferrin, D.L.: The role of trust in organizational settings. Organ. Sci. **12**(4), 450–467 (2001)
4. Esfahbod, B., Kreuger, K., Osgood, N.: Gaming the social system: a game theoretic examination of social influence in risk behaviour. In: Agarwal, N., Osgood, N. (eds.) SBP 2015. LNCS, vol. 9021, pp. 296–301. Springer, Heidelberg (2015). doi:10.1007/978-3-319-16268-3_33
5. Ostrom, T.M., Sedikides, C.: Out-group homogeneity effects in natural and minimal groups. Psychol. Bull. **112**(3), 536 (1992)
6. Rusman, E., Van Bruggen, J., Sloep, P., Koper, R.: Fostering trust in virtual project teams: towards a design framework grounded in a TrustWorthiness Antecedents (TWAN) schema. Int. J. Hum. Comput. Stud. **68**(11), 834–850 (2010)
7. Tajfel, H., Turner, J.C.: The social identity theory of intergroup behavior (2004)

Cultural Dimension Theory Based Simulations for US Army Personnel

Brian An[1(✉)], Donald E. Brown[1(✉)], Riannon M. Hazell[2(✉)], and Peter Grazaitis[2(✉)]

[1] University of Virginia, Charlottesville, VA 22904, USA
[2] U.S. Army Research Laboratory, Aberdeen Proving Ground, Adelphi, MD 21005, USA
baa4cb@virginia.edu
http://www.sys.virginia.edu/

Abstract. US Military Personnel are constantly operating in culturally diverse operational theaters and among multinational coalitions around the world. Though some cultural and linguistic criteria are considered when filling deployments, cultural missteps continue to plague the success of our combat operations. In order to address the increasing need for cross-culturally competent personnel the Department of Defense (DoD) requires scalable evaluative methods that supplement current measures primarily focused on evaluating linguistic skills for cross-cultural competence. This work investigates the integration of Cultural Dimension Theory and immersive avatar-based gaming systems with the goal of measuring and predicting cross-cultural competence. The objective of this effort is to assess the applicability of Cultural Dimension Theory as a means to interpret perceived cultural differences and to introduce a novel framework by which a Cultural Dimension-based simulation can be developed.

Keywords: Cultural simulation · Avatars · Cultural dimension theory

1 Introduction

Cross-Cultural Competence (C3) has been identified by both corporate and military establishments as a necessary requirement for success in both modern business and military operations [1,2]. DoD Instruction 5160.70, Management of the Defense Language, Regional Expertise, and Culture (LREC) program, identifies regional expertise and culture as mission critical competencies for the Defense LREC program. Specifically, the Department of Defense (DoD) has recommended the creation of a developmental model for C3 expertise that prescribes the progression of competency development [3]. For this reason, we are developing an objective and quantifiable means of measuring cultural competence through the adoption of Cultural Dimension Theory and an avatar-based simulation. This paper describes exploratory survey results to identify the applicability of Cultural Dimension Theory. Additionally, this paper discusses the proposed Cultural Simulation Design Process which incorporate Cultural Dimension Theory in a simulation intended to measure C3.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 65–70, 2017.
DOI: 10.1007/978-3-319-60240-0_8

2 Literature Review

2.1 Cultural Dimensions

Over the past several decades, several research efforts have investigated theories to characterize the abstract concept of culture. This research can be broadly characterized as an effort to define cultures across a common set of parameters and values that allow for cross-cultural comparison.

Despite broad research in the area, none have paralleled the widely accepted six dimensional model of Geert Hofstede [4]. After 30 years of critique and evaluation, this model or variations of this model continue to appear as the most utilized cultural dimension framework [5]. Hofstede characterizes culture across six dimensions: Individualism vs Collectivism (IDV); Power Distance (PDI); Uncertainty Avoidance (UAI); Masculinity vs Femininity (MAS); Long Term vs Short Term Orientation (LTO); Indulgence vs Restraint (IVR) [4]. Sixty-three countries have been evaluated across these dimensions. Other notable frameworks include Trompenaar's Cultural Dimensions and Kluckhohn's Value Orientations which, despite their research validity, were not considered for this research [6].

2.2 Virtual Reality C3 Simulations

The proliferation and recent sophistication of virtual environment development tools such as Unity3D and Unreal Engine have facilitated the growth of realistic, dynamic, and immersive simulations. The research efforts described below represent the notable efforts in the cultural domain. Despite the advances made in these efforts, they all appear to lack a formal foundation in Cultural Psychology.

Sandia National Laboratory in collaboration with the U.S. Army John F. Kennedy Special Warfare Center developed a multiplayer cross-cultural game to simulate joint host-nation operations [7]. Though non-player characters were used for simple game-progression interactions, human players played the avatars in order to create the most realistic interaction possible. The evaluation of a player's performance and C3 was performed by real-time observations by Subject Matter Experts (SME).

The Cultural Awareness in Military Operations (CAMO) project at the Norwegian University of Science and Technology developed a Virtual Afghan village in Second Life (SL) with the objective of simulating a military security operation. Participants and SMEs evaluated the realism and effectiveness of the simulation though performance evaluations of the participants have yet to be published [8].

3 Cultural Dimension Theory Applicability

In an effort to determine the applicability of Cultural Dimension Theory as a means of measuring the Cross-Cultural Competence of DoD personnel, we conducted a perspective-taking survey of United States Military Academy (USMA) cadets. Seventy-four USMA cadets whom had recently returned from semester exchanges were surveyed.

3.1 Objective

The objective of this survey was to determine whether the cadets' perceived cultural differences were consistent with the published Hofstede's Cultural Dimension studies of the same countries. Similar methods have been used in other cross-cultural studies with mixed results though no studies were found to have used the Hofstede's Values Survey Module 2013 Inventory in this fashion [9].

3.2 Method

All the cadets were administered Hofstede's Values Survey Module 2013. Once complete, the cadet's were immediately administered the same survey with the instruction to answer the questions as they would expect the people in their exchange country to respond. The cadet semester exchange countries were concentrated in five countries (Taiwan, France, Germany, Mexico, Brazil). All the cadets designated American as their primary culture. They were not given any descriptive information about the specific dimensions.

3.3 Results

Figure 1 depicts a comparison between the difference of means observed in this study and the difference of means observed in the published Hofstede results of the same countries. The scales of each of the dimensions from the higher values to the lower values are as follows: High to Low Power Distance, Individualistic to Collectivistic, Masculine to Feminine, High to Low Uncertainty Avoidance, Long-Term to Short-Term Orientation, Indulgent to Restraint. The numeric scale in Fig. 1 is the absolute difference between the mean score of the USMA cadets and the mean scores as published in previous Hofstede publications. This is intended to show whether the observed difference of means in this study trend in the same direction as Hofstede's results.

3.4 Discussion

In examining Fig. 1, twenty-four of the thirty calculated cultural dimensions spanning the five countries trended in the same direction as that of Hofstede's published results. However, a couple notable dissonances were observed. Though several hypotheses are proposed for these dissonances, we are unable to definitively explain the root cause without further investigation.

In the case of Taiwan, we observed that the survey participants' perceptions of Taiwan were generally more masculine than feminine in contrast to the more neutral perspective reported in Hofstede's work. Other studies have replicated the result we observed [10,11]. Wu et al. explained the difference due to the rapid change in the U.S. workforce and gender roles since Hofstede's results were published which in turn would change the comparative result.

Another notable difference was in Power Distance for France. Through discussions with the participants, it was noted that they observed a less rigid rank

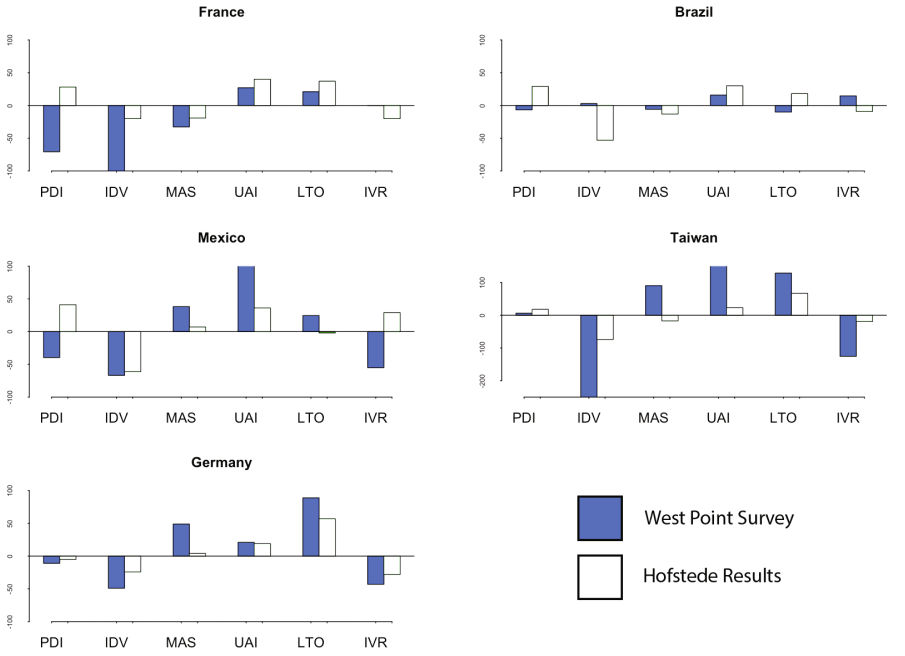


Fig. 1. Comparison of USMA results and Hofstede results for mean difference

structure among the French military than what they were accustomed to at USMA. This may be more representative of French military culture rather than French culture as a whole.

This result lends some evidence that the cultural dimensions can be correctly perceived which presumably would allow an individual to better interpret behaviors and respond accordingly. To provide further support to this conclusion, we plan to correlate these findings to established self-report measures of C3.

4 Cultural Simulation Design Process

Previous iterations of this research [12] and other related efforts have not attempted to incorporate Cultural Dimension Theory into their simulation design processes. As such, these simulations failed to present the users anything beyond rudimentary cultural interactions. As an evolution of the previously developed avatar-based cultural training simulation [12] and the results of the USMA study, this research effort systematically incorporates the previously discussed Cultural Dimension Theory into the simulation design process to address DoD requirements [3].

4.1 Overview

The Cultural Simulation Design Method extends existing Human-Computer Interface (HCI) design principles as described in other simulation design efforts [7]. This specific addition uses Cultural Dimension Theory to influence the development of the branched dialogues in order to increase the efficacy of the pedagogical value and assessment of the simulation.

Dimension Selection. In order to determine which Cultural Dimensions could be used to evaluate cross-cultural competence, we first determined the participant's most affiliated culture as well as the target assessment culture of the simulation. As an example, a participant considers himself most closely aligned with American culture and the simulation is designed to reflect Chinese culture.

We then determined which Cultural Dimensions were the most different between the participant's culture and the simulation culture. Hofstede's previous work shows that Power Distance, Individualism, Long-Term Orientation, and Indulgence have the most notable differences [13]. From these remaining dimensions, a subset was selected as the target dimensions in the simulation.

Dialogue Generation. In-game branched dialogues with avatars are the sole means of progressing through the simulation. These dialogues are contextually developed based on a predetermined storyline.

In response to an avatar, participants must select a response from a list of predetermined responses. The spectrum of these responses is crafted to capture the spectrum of the target cultural dimension. This methodology is repeated throughout the entirety of the dialogue trees in order to maximize the number of times a cultural dimension is exposed to the participant.

Cross-Cultural Competence Assessment. Given the integration of Cultural Dimensions into the dialogue structure, cross-cultural competence is assessed based on the number of culturally appropriate responses selected. We hypothesize that one's ability to recognize and respond to the appropriate cultural dimensions is reflective of a higher cross-culture competence.

4.2 Simulation Development

Using Unity3D as the simulation development engine and the previously described design process, we developed a five scenario simulation set in a Chinese university. The initial background model was purchased from the Unity3D marketplace and was subsequently tailored for our storyline. The avatars were each individually developed in Blender and Mixamo Fuse. Using the Pixel Crushers Dialogue System, we created branching text/voice dialogue trees.

5 Conclusion

This project builds upon previous work [12] to assess and improve a soldier's cross-cultural competence through the use of avatar-based simulation systems.

The Cultural Dimension survey of USMA cadets provides initial validity to the use of Cultural Dimensions to measure cross-cultural competence. Additionally, the novel Cultural Simulation Design Process introduces a systematic methodology to develop cultural simulations grounded in current theory.

References

1. Gallus, J.A., Gouge, M.C., Antolic, E., Fosher, K., Jasparro, V., Coleman, S., Selmeski, B., Klafehn, J.L.: Cross-cultural competence in the department of defense: an annotated bibliography. Technical report, DTIC Document (2014)
2. Mor, S., Morris, M.W., Joh, J.: Identifying and training adaptive cross-cultural management skills: the crucial role of cultural metacognition. Academy Manag. Learn. Educ. **12**(3), 453–475 (2013)
3. McGuire, G., McGinn, M.G., Weaver, M.N.: Developing and managing cross-cultural competence within the department of defense: Recommendations and learning and assessment (2008). https://www.deomi.org/culturalreadiness/documents/racca_wg_sg2_workshop_report.pdf
4. Hofstede, G.H., Hofstede, G.: Culture's consequences: Comparing values, behaviors, institutions and organizations across nations. Sage (2001)
5. Fang, T.: Yin yang: a new perspective on culture. Manage. Organ. Rev. **8**(1), 25–50 (2012)
6. Trompenaars, F., Hampden-Turner, C.: Riding the waves of culture: Understanding diversity in global business. Nicholas Brealey Publishing (2011)
7. Raybourn, E.M.: Applying simulation experience design methods to creating serious game-based adaptive training systems. Interact. Comput. **19**(2), 206–214 (2007)
8. Tasdemir, S.A., Prasolova-Førland, E.: Visualizing afghan culture in a virtual village for training cultural awareness in military settings. In: 2014 18th International Conference on Information Visualisation (IV), pp. 256–261. IEEE (2014)
9. Heine, S.J., Lehman, D.R., Peng, K., Greenholtz, J.: What's wrong with cross-cultural comparisons of subjective likert scales? the reference-group effect. J. Pers. Soc. Psychol. **82**(6), 903 (2002)
10. Wu, M.: Hofstede's cultural dimensions 30 years later: a study of taiwan and the united states. Intercultural Commun. Stud. **15**(1), 33 (2006)
11. Fernandez, D.R., Carlson, D.S., Stepina, L.P., Nicholson, J.D.: Hofstede's country classification 25 years later. J. Soc. Psychol. **137**(1), 43–54 (1997)
12. Moenning, A., Turnbull, B., Abel, D., Meyer, C., Hale, M., Guerlain, S., Brown, D.: Developing avatars to improve cultural competence in us soldiers. In: 2016 IEEE Systems and Information Engineering Design Symposium (SIEDS), pp. 148–152. IEEE (2016)
13. Hofstede, G.: Asian management in the 21st century. Asia Pac. J. Manag. **24**(4), 411–420 (2007)

Socio-Cultural Cognitive Mapping

Geoffrey P. Morgan[1(✉)], Joel Levine[2], and Kathleen M. Carley[1]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
gmorgan@cs.cmu.edu
[2] Dartmouth College, Hanover, NH, USA

Abstract. We introduce Socio-cultural Cognitive Mapping (SCM), a method to characterize populations based on shared attributes, placing these actors on a spatial representation. We introduce the technique, taking the reader through an overview of the algorithm. We conclude with an example use-case of the Hatfield-McCoy feud. In the Hatfield-McCoy case, the SCM process clearly delineates members of the opposing clans as well as gender.

1 Introduction

Frequently in social science research, we have multiple attributes of a sample of a population of interest, but we want to understand the implicit communities within the sample. Are there multiple group of actors or is the sample relatively homogenous? What attributes show strong delineation between communities? Answers to these questions may be fruitful in understanding different community reactions to change, as well as to develop a more nuanced understanding of a group of interest, in that there may be multiple important communities within that group-label. To explore these questions, we introduce Socio-cultural Cognitive Mapping (SCM).

In the SCM process, the user identifies information of interest, reified either as attribute data on a node-set or network data. This information is then used to inform constraints between nodes – these constraints identify optimal distances between the nodes. We then place nodes at random in a 1D, 2D, or 3D space with various geometries. Nodes move in a greedy but non-local fashion to the best available position based on their constraints. After all nodes have moved to best available positions without improvement in overall fit, a Chi-Square score for goodness of fit is calculated and reported. We do multiple iterations of each geometry of interest. The best fitting space is then returned to the user for visualization and analysis.

SCM bears features in common with several other multi-dimensional scaling techniques that embed nodes in a space, thereby visualizing clustering, separation and dimensions of differentiation. SCM, like these other techniques, is a dimension reduction procedure. At this level, SCM is part of a class of tools that define a clustering among nodes based on their similarity in some other space – e.g., similarity in attribute or their connections to other nodes.

The canonical example is MDS (Multi-Dimensional Scaling). Unlike classical MDS (Torgerson 1958), SCM does not rely on eigenvector decomposition for dimension reduction. In that sense, SCM is closest to general MDS (Borg and Groenen 2005). A key difference between SCM and MDS is that MDS takes the attributes as

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 71–76, 2017.
DOI: 10.1007/978-3-319-60240-0_9

given; whereas, in SCM these attributes are first converted to a set of binary attributes thus giving equal weight to each "category" of information. Like MDS, SCM can identify a set of dimensions that best discriminate these clusters. Another key difference is that even in general MDS the user must specify the distance metric (e.g., Euclidean or Manhattan) and the number of dimensions (e.g. 2 or 3). In SCM, the distance metrics as well as the dimensions are part of the optimization. A third difference is that in SCM the nodes can vary in how much "constraint" they have on the position of the other nodes, whereas in MDS procedures all nodes contribute equally. Further that contribution is also optimized over. Finally, in SCM similarity and dissimilarity can be simultaneously taken into account; whereas, in MDS only dissimilarity is considered.

A second example is principal components analysis (Jolliffe 2002). Principal components analysis presents variables as linear combinations of all other variables, the dimensional reduction rotates the space to visualize a small number of dimensions in which distance and variation approximates that of the original space. As ordinary two-variable regression reduces a two dimensional space to a one dimensional space on which variance approximates the whole, ordinary reduction techniques reduce a high dimensional space to a low dimensional space in which variance still approximates the whole.

By contrast, SCM reduction is built from a different base: closer to the data and shedding assumptions. Typically, a "variable" is reconceived as a collection of attributes and their joint distribution is attended to directly, rather than relying on a single number proxy for the whole distribution. Conceiving "input" as a collection of attributes, SCM is not restricted to number-valued variables.

For example, "height-weight" data are sometimes used to demonstrate standard techniques. For these data, standard procedure would have us improve prediction by adding variables. SCM procedure looks at the detail. It automatically notes that weight for a given height is not normally distributed but more like a Laplacian. And, with proper modeling SCM makes the joint frequencies of height and weight are more predictable — without the complexity of higher dimensions.

With high-variable data, assumptions about the space itself are shed. For example, in some cases substitution of a "Manhattan metric" for a Euclidean metric will enhance the prediction of joint frequencies — without additional parameters and without higher dimensions.

We continue this paper by describing the algorithm in more detail, and then following that explanation with a case-study example of the Hatfields and McCoys. We conclude with a summary of key points from the paper and next steps for SCM.

2 Algorithm

The SCM process has multiple steps. For additional technical details on the SCM algorithm, see Levine and Carley (2016). The user selects data to be used to inform similarity, identifies how similarity should be assessed, selects a set of geometries for the nodes to be placed on, and then allows the process to proceed, with the tool returning the best fitting positions across all geometries as coordinates. We go into more detail on each of these steps in the following sub-sections.

2.1 Selecting Data

The SCM supports multiple types of data, and one of the goals of the SCM process is to make it easier to consider node attributes and network matrices as more interchangeable. The three types of input (attributes, binary network data, weighted network data) it evaluates are each processed differently.

Attribute data is first pre-processed to identify whether it is binary, categorical, or quantitative. Categorical and quantitative variables are converted into a set of binary variables.

Binary network data is treated as a set of binary attributes. For the SCM process to recognize that a network is binary, the checkbox "binary values" must have been selected on the network tab in ORA. Otherwise, the network data is treated as weighted network data.

Weighted network data is treated explicitly as a set of constraints for the SCM process, while the previous two inputs are used to generate a similarity matrix (more details will be given on how the similarity matrix is calculated shortly). Link weights are assumed to be event rather than distance counts, and so high counts indicate closer distance constraints for the SCM process. If your values are instead a distance metric (e.g., number of miles to a given city), the values will need to be inverted before use in the SCM process.

2.2 Generating the Similarity and Constraint Matrices

A constraint matrix informs the ideal position of points in the various evaluated geometries. The constraint matrix is calculated as a transformation of a similarity matrix or a weighted network if that option is selected. The similarity matrix is generated based on the binarized SCM attributes rather than the network or node-attributes from which they spring. Multiple control variables can inform the generation of the similarity matrix calculation, including the removal of redundant and mutually exclusive attributes, the use of negative similarity, and whether completely dis-similar nodes should be placed far away from each other or can ignore each other.

Once the similarity matrix has been generated, we convert the similarity matrix into the constraint matrix. There are multiple ways of generating a constraint matrix, but for the examples in this work, we use a simple inverted similarity. Future iterations will include other transformations.

2.3 Moving Nodes to Satisfice Constraints

After calculating the constraint matrix, we run a number of iterations across each geometry and attenuation setting, note that the SCM supports 1D, 2D, and 3D node placements. Typical geometry and attenuation settings are 0.7, 1.0, 2.0, and 3.0. We confine ourselves to 2D Euclidean spaces for this paper.

2.4 Evaluating Fit and Returning Results

Once all nodes have been unable to find a better position, the SCM evaluates the overall fit based on the constraint set and calculates a Chi-Square. We are using the Chi-Square in this fashion as a goodness of fit, and not as a statistical evaluation. We calculate the Chi-Square, and report the best fit per geometry and attenuation setting, the standard deviation of Chi-Square per geometry and attenuation setting, and the best Chi-Square over all.

For this best fitting Chi-Square, we return (by default) the coordinate positions for all nodes, but we can also report back the Similarity Matrix, the Constraint Matrix, and the Chi-Square Error Cell Matrix. Because we return the Error Matrix, we can visualize the level of satisficing and conflict in the node's current positions, providing more confidence in the ultimate groupings. We can also return the transformed SCM binary attributes to each node. Once we have node coordinate positions, we can use visualizations to examine the resulting groupings.

3 The Hatfield-McCoy Case Study

In the Hatfield-McCoy case study, we wanted to demonstrate the technique's utility on a historical scenario. The Hatfields and McCoys were two rural families living across the Big Sandy River on the West Virginia and Kentucky sides respectively. The two families were in a bitter feud from 1863 to 1891. "Devil Anse" Hatfield led the Hatfields of West Virginia while Randolph McCoy led the McCoys of Kentucky. The Hatfields were wealthier, owning a timbering operation and serving in local government, while the McCoys were farmers. Both families made and sold moonshine. Intermarriage between the families happened before and even during the feud, which was first kicked off by Asa McCoy's death. The McCoys eventually killed a mutual relative of both Hatfields and McCoys, Bill Stanton, who sided with the Hatfields over the ownership of a hog. The feud escalated over time until eventually most of the McCoys moved to Pikesville (20 or 30 miles from the border) to escape the violence and "Devil Anse" Hatfield was arrested after an armed shootout.

We used multiple sources to generate this dataset, since the exact composition of the clans is difficult to determine, and several sources contradict. This dataset is primarily intended for illustrative purposes[1]. We identified 66 members of the two families, and for each individual, we identified the binary attributes listed in Table 1.

Given the nature of the feud, it seems clear that Hatfield and McCoy would clearly differentiate the nodes. We would also expect that gender, given the era, would be clearly differentiated. Those that are intermarried are probably not clearly differentiated. We would hope that Devil Anse and Randolph family members are widely separated.

We ran the SCM removing redundant but not mutually exclusive attributes, counted both positive and negative similarity, and used a Euclidean space. The Chi-Square was 907.2 out of 2145° of freedom. The standard deviation of the Chi-Square in this

[1] The Hatfield-McCoy data can be downloaded as DynetML from the CASOS website: http://www.casos.cs.cmu.edu/tools/datasets/internal/index.php#HatfieldMcCoy.

Table 1. Binary attributes of the Hatfield-McCoy case study

Attribute	%Sample	Attribute	%Sample
Man	75.8%	Woman	24.2%
Hatfield	45.5%	McCoy	54.5%
"Devil Anse" Family	18.2%	Randolph Family	24.2%
Harmed McCoy	6.1%	Harmed Hatfield	7.6%
Intermarried	10.6%	Killed in Feud	16.7%

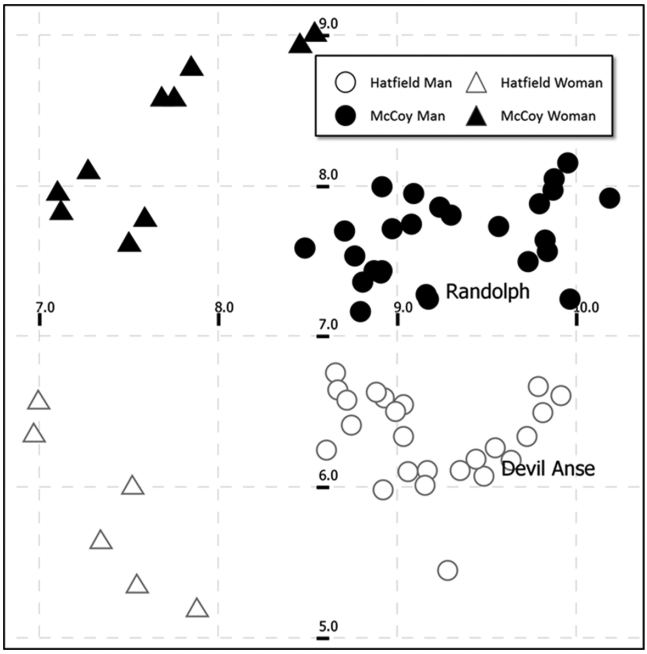


Fig. 1. Position of nodes reflects clean separation between clans (color) and gender (shape). The leaders of the two clans are labeled.

geometry was 225.3. This model has removed a substantial amount of noise – the Wilson-Hilferty Z-Score approximation (Wilson and Hilferty 1931) is −24.58 (Fig. 1).

4 Discussion and Conclusion

In this work, we have introduced Socio-cultural Cognitive Modeling (SCM), a technique we developed to characterize populations and identify implicit groups and illustrated the technique with a historical scenario.

To support use, SCM is available in ORA (Carley 2014). ORA is an analysis and visualization toolkit for high dimensional network and social network data. Network

images shown here in (e.g., Fig. 1) were done in ORA. Consequently, SCM is currently available for use by the community.

More work remains to be done. We plan to add a sophisticated optimization package, rather than relying on greedy stochasticity. This should support finding an SCM configuration for complex data with an improved overall goodness of fit. We are also interested in evaluating our best fit positions in other ways than a Chi-Square. We also plan to add the ability to support counterfactual simulation with the tool.

Nonetheless, the process already provides a novel way to take advantage of information available either as attributes or network data to produce an estimate of each node's appropriate position in relation to each other. Implicit groups of significant import may be discovered.

References

Borg, I., Groenen, P.J.: Modern Multidimensional Scaling: Theory and Applications. Springer, New York (2005)

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, Heidelberg (2014)

Jolliffe, I.: Principal Component Analysis. Wiley Online Library (2002)

Levine, J.H., Carley, K.M.: SCM System. (CMU-ISR-16–108) (2016)

Torgerson, W.S.: Theory and methods of scaling. Wiley, New York (1958)

Wilson, E.B., Hilferty, M.M.: The distribution of chi-square. Proc. Natl. Acad. Sci. **17**(12), 684–688 (1931)

Cyber and Intelligence Applications

A Cognitive Model of Feature Selection and Categorization for Autonomous Systems

Michael Martin[1(✉)], Christian Lebiere[1], Maryanne Fields[2], and Craig Lennon[2]

[1] Carnegie Mellon University, Pittsburgh, PA, USA
mkm@andrew.cmu.edu, cl@cmu.edu
[2] Army Research Laboratory, Aberdeen, MD, USA
{mary.a.fields22.civ, craig.t.lennon.civ}@mail.mil

Abstract. We describe a computational cognitive model intended to be a generalizable classifier that can provide context-based feedback to semantic perception in robotic applications. Many classifiers (including cognitive models of categorization) perform well at the task of associating features with objects. Underlying their performance is an effective selection of the features used during classification. This Feature Selection (FS) process is usually performed outside the boundaries of the model that learns and performs the classification task, often by a human expert. In contrast, the cognitive model we describe simultaneously learns which features to use, as it learns the associations between features and classes. This integration of FS and class learning in one model makes it complementary to other Machine-Learning (ML) techniques that automate the FS process (e.g., deep learning methods). But their integration in a cognitive architecture also provides a means for creating a dynamic context that includes disparate sources of information (e.g., environmental observations, task knowledge, commands from humans); this richer context, in turn, provides a means for making semantic perception goal-directed. We demonstrate automated FS, integrated with an Instance-Based Learning (IBL) approach to classification, in an ACT-R model of categorization by labeling facial expressions of emotion (e.g., happy, sad) from a set of relevant, irrelevant, distinct, and overlapping facial action features.

Keywords: Autonomous systems · Feature selection · ACT-R · Cognitive architectures · Machine-learning · Classification · Categorization · Instance-based learning

1 Introduction

Robots tend to process information with perceptual systems that feed forward to cognitive systems that do something intelligent (e.g., planning, reasoning). Thus, cognitive systems tend to provide little to no feedback to autonomous perception. We assume that autonomous perception can be improved by establishing a feedback loop between perceptual and cognitive systems. The feedback loop would make autonomous perception more of an interactive process than it is today by providing a means for

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 79–89, 2017.
DOI: 10.1007/978-3-319-60240-0_10

exploiting cognitive context (e.g., goals) to augment perceptual processes dominated by stimulus-driven information.

Semantic perception in autonomous systems is a complex process that generally involves parsing sensor data into features, grouping features into likely objects, and then assigning labels to objects. Semantic perception thus anchors recognized objects in the environment to labels that refer to those objects. The labels are placed in semantic maps for data interchange, and then passed to cognitive systems that can operate on the symbols.

More often than not, semantic perception refers to a computer vision system with independently trained object detectors. These object detectors tend to operate on local spatial features in the images they process. This means anchoring, the process of grounding semantic labels in sensor data, is frequently based on a local, stimulus-driven context that varies because of noisy sensor data, occlusion, lighting conditions and so on. To reduce classification error, some detectors employ additional context obtainable from global image features such as intensity, predominant color, etc. Others incorporate expectations about objects based on their location within a scene.[1]

Including more context in autonomous perceptual systems is one approach to improving anchoring. Another approach is to get additional context from cognitive systems and iteratively integrate it with stimuli. This is the approach we have been exploring. The basic idea is to create a general model of perception in a cognitive architecture (i.e., a computational instantiation of a unified theory of cognition) that can encode features or labels from object detectors; combine that environmental context with task-relevant information; and then provide the resulting cognitive context (e.g., expectations, points of ambiguity, points of interest, etc.) in a feedback channel that can be used by perceptual systems to augment accuracy (and efficiency).

Our efforts thus far are focused on the feedforward signal from semantic perception to cognition. As a first step, Fields et al. [1] addressed the issue of whether a cognitive architecture (ACT-R) can encode the kinds of information provided by semantic perception. In addition to object labels, semantic maps may include feature information and continuous variables (e.g., confidence, intensity). Fields et al. [1] demonstrated that Instance-Based Learning (IBL) in ACT-R performs similarly to a k-Nearest Neighbors classifier, while ACT-R's partial matching mechanism supports classification using continuous features.

Fields et al. [1] also demonstrated that adding goal-directed meaning to scene recognition processes improves accuracy. When a robot's goal is to classify public spaces (e.g., conference rooms vs dining rooms), performance improves by using features representing the general notion of social immediacy rather than counts of common objects (i.e., chairs and tables) found in the rooms. Thus the goal-directed selection of features improved recognition performance. The categorization model described below (CAT) performs goal-directed FS as our next step toward a mechanism that supports context-sensitive semantic perception.

[1] In computer vision there are both local features such as edges, Scale-Invariant Feature Transform (SIFT), or Speeded Up Robust Features (SURF); and global features such as parts, Histogram of Oriented Gradients (HOG), textures or contours. We refer to all of these as local features since they focus on the object and not the environment.

2 Approach

CAT is an ACT-R classifier that simultaneously learns: (1) associations between features and classes, and (2) which features to encode for which classes. In ML terms, it combines an IBL paradigm for the classification problem with Reinforcement Learning (RL) for the FS problem. IBL allows CAT to use memory of past examples of a class to generalize to novel members of the class (for a discussion of IBL theory in ACT-R, see Gonzalez et al. [2]). RL allows CAT to select features as a function of their effectiveness. The domain-independence of RL and IBL means CAT may generalize across classification problems; allowing interactions with perceptual classifiers regardless of the classes or features involved. Indeed, the approach generalizes to any decision process that includes decision instances and feedback.

2.1 A Cognitive Theory of Categorization

In terms of cognitive theory, CAT maps mechanisms of categorization phenomena in humans to mechanisms in ACT-R described as underlying cognitive phenomena in general (e.g., [3]). Anderson and Betz [3] mapped the Exemplar-Based Random Walk (EBRW) model of categorization [4] onto ACT-R mechanisms. Here, we describe CAT as a comparable mapping of the EBRW Response Time (EBRW-RT) model of categorization [5], even though we initially developed CAT based on the computational mechanisms of the ACT-R framework. EBRW-RT extends EBRW from a model focused only on class decision processes to one that includes feature-encoding processes.

EBRW assumes features are encoded in a single step because it is concerned with classification tasks that involve integral features, which are encoded simultaneously. EBRW-RT extended EBRW to classification tasks that involve separable features, which are encoded sequentially. Thus, while EBRW and EBRW-RT are similar information accumulation models, EBRW-RT associates the random walk process of EBRW with iterative feature-encoding processes. Accordingly, features are encoded one at a time until a category decision is made. The decision about category membership depends on similarity between an evolving stimulus representation and instances in memory.

2.2 ACT-R Implementation

Cognitive architectures provide a software implementation of cognitive theory. As software, ACT-R can be viewed as a set of asynchronous modules (e.g., memory, perceptual, motor) that store and process information, plus a production system that stores or retrieves information from buffers associated with each module. Creating a cognitive model of a task involves writing a set of productions that performs the steps required by the task. Productions encode condition-action rules that perform actions on the buffers anytime they match conditions on the contents of the buffers. Each buffer contains at most one chunk of information, where a chunk is a data structure with a set of slot-value pairs.

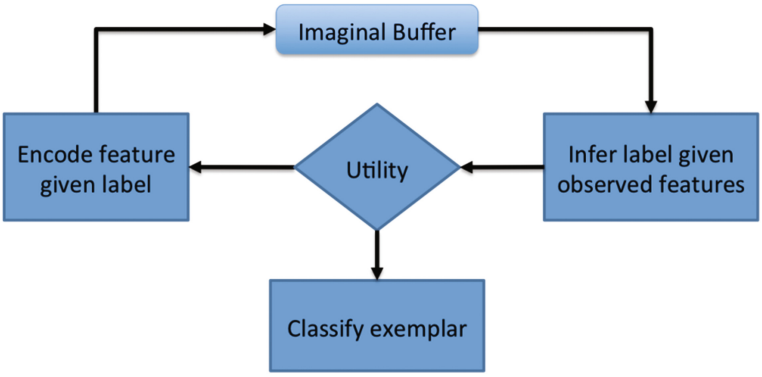


Fig. 1. The general structure of the categorization model

Figure 1 depicts the general structure of CAT. Three types of productions drive the classification process: feature encoding, class inference, and decision. Feature-encoding productions take a single feature (e.g., from a semantic map) and place it in the imaginal buffer used to construct the object representation. Configuration chunks in the imaginal buffer represent the associations among features and classes, having a slot for each feature encoded and the class label. Due to the serial firing of feature-encoding productions, stimulus information accumulates in the imaginal buffer.

Class-inference productions use the context provided by observed features in the imaginal buffer to retrieve class labels that have been associated with those features in the past. That is, observed features serve as a cue for finding a configuration chunk in declarative memory (i.e., the declarative module) that approximately matches current observations but also has a class label. Retrieving such a chunk amounts to an inference about what the class label should be (i.e., a class hypothesis), given the features observed thus far in the classification process.

The firing of a class-inference production initiates an exemplar retrieval process. Activation (see Eq. 1), a numerical quantity associated with each chunk in ACT-R Declarative Memory (DM), drives the retrieval process. Activation spreads from features in the imaginal buffer to configuration chunks in DM that share one or more of the observed features; chunks are activated by the degree to which they share features, plus a random noise component, ε. The configuration chunk with highest activation is retrieved.

$$A_i = \sum_{k=1}^{K} W_k S_{ki} + \varepsilon \tag{1}$$

A decision production terminates the cycle of feature-encoding and class-inference productions firing over time and assigns a class label to the stimulus. CAT's process of choosing between competing productions involves a numeric quantity associated with each ACT-R production called utility. Utility reflects a production's usefulness for achieving a goal - stimulus classification in this case. In cases where multiple productions' conditions match the context provided by buffers, the production with the

highest utility will fire. Therefore, conflict resolution is based on a production's utility value, including a random noise component (see Eq. 2), which promotes response variability and exploration of decision strategies. In CAT, the decision production competes with feature-encoding productions. Thus, the information accumulation process terminates when the utility of a decision production exceeds the utility of all feature-encoding productions.

$$Pr_i(i) = e^{U_i/\sqrt{2s}} / \sum_j e^{U_j/\sqrt{2s}} \tag{2}$$

Production utilities can be learned from experience by reinforcing productions that lead to successful task performance. RL involves the distribution of rewards in the production system. The reward received by a production is scaled based on the time between the distribution of a reward and when a production fired. Rewards are temporally discounted so that productions that fired immediately before reward distribution receive more than productions that fired in the more distant past. Temporal discounting is necessary because rewards are distributed to all productions that have fired since the last time rewards were distributed. The reward is then used to gradually adjust utility until it matches the average reward received (see Eq. 3).

$$U_i(n) = U_i(n-1) + \alpha[R_i(n) - U_i(n-1)] \tag{3}$$

where $U_i(n)$ is the utility of production i after its nth firing and $R_i(n)$ is the reward received by production i after its nth firing.

Implementing the classification process described above in an IBL process supports the bootstrapping of class knowledge from experience, and provides a hook for controlling the distribution of rewards. That is, IBL involves feedback about model performance, which can be used to distribute rewards too. We thus added three feedback productions to CAT: one distributes rewards for a correct classification; one distributes rewards for an incorrect classification and corrects the class slot in the imaginal buffer; and one harvests (stores) the configuration chunk in the imaginal buffer as an instance in DM.

Feedback productions fire after a class decision production fires, and create an instance in DM that associates observed features and (corrected) class labels. Simultaneously, the utility of a subset of feature-encoding and decision productions is adjusted. As the utilities of these competing productions evolve, the order and number of features encoded will vary.

CAT, however, will have no effective way to learn a classification task in which some features are important signs of one class but not so important for others because, as described above, productions encode features without regard to the current class hypothesis. We avoided this problem by including the class hypothesis as an additional constraint in the condition side of feature-encoding productions. This constrains the distribution of rewards in a more task-relevant manner by drawing a distinction between encoding a particular feature in the context of one class hypothesis and encoding that same feature in the context of another class hypothesis. Thus expectations, in the form of

class hypotheses, change the way CAT encodes the world and subsequently parses it into categories.

3 Testing and Revision

We tested CAT using the publicly available Cohn-Kanade facial expression dataset [6]. The dataset associates emotion labels for expressions from 123 participants with facial Action Units (AUs). We used 327 exemplars of 7 elicited expressions (see Fig. 2). Their associated AUs are based on Ekman's Facial Action Coding System [7, 8], which separates the observation of facial actions from inferences about emotional state.

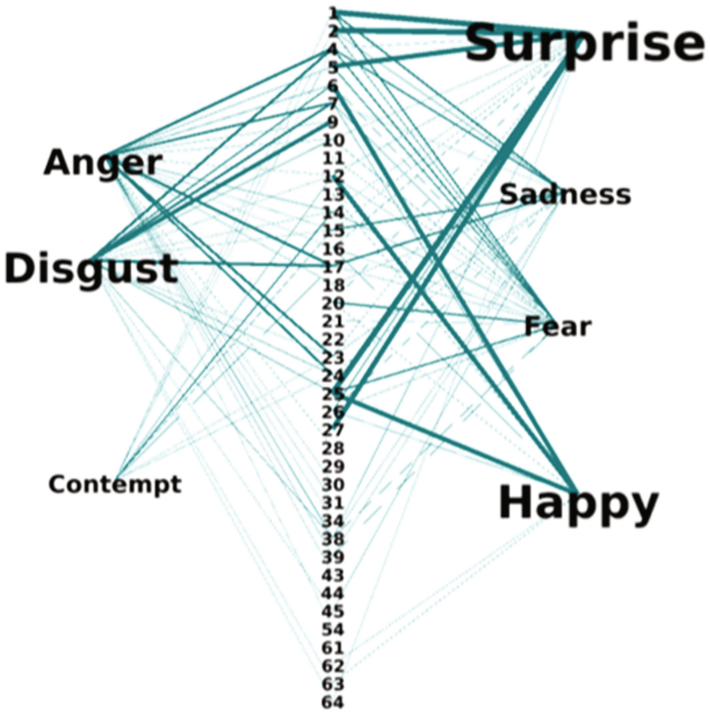


Fig. 2. The relationship between features and expressions

We removed 25 AUs that were constant across the exemplars we used, leaving a total of 39 AUs. Figure 2 shows the association of the remaining AUs (as numbers in the center) with emotion classes (as words on the sides). Font size of the words codes class frequency, whereas the weight of the line connecting emotion to AU codes co-occurrence frequency. Notice that the base rates of the classes differ; features differ among classes; the number of relevant features varies between classes; and classes share features.

3.1 Testing Procedure

Revisions to CAT were based on a simple testing procedure in which the 327 exemplars were randomized, and then presented to the model one by one. For each exemplar, CAT makes a classification decision and receives feedback about the exemplar's true class. This procedure was repeated for a total of 10,000 trials to examine learning over an extended period. Unless otherwise noted, global ACT-R parameters were left at their default settings.

3.2 Results and Incremental Refinement

The learning curves in Fig. 3 depict classification accuracy for the Canonical CAT model described above, plus four, cumulative revisions. The points plotted for each line represent the proportion of correct responses in a sliding window of 1000 trials. The proportions are right-aligned so that the mean of the first 1000 trials is plotted with an x-coordinate of 1000, the mean of the second window as 1001, and so on. The low amplitude oscillations from trial to trial reflect noise in the retrieval of chunks from DM used for classification and the firing of relevant productions in procedural memory used for feature encoding.

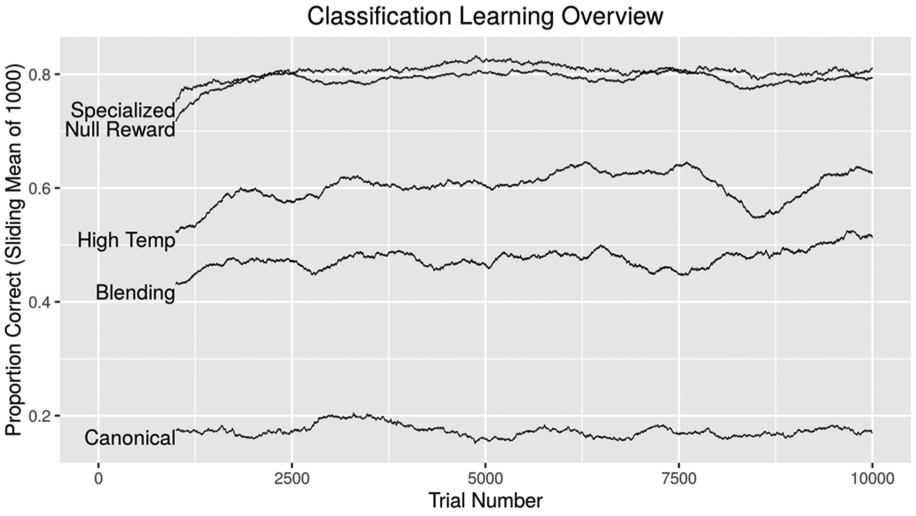


Fig. 3. A comparison of learning results for all models.

The Canonical CAT model reflects the basic IBL and RL processes in ACT-R (see Fig. 3). As can be seen, the Canonical model exhibits near chance-level classification performance and no long-term learning trend.

The Blending model uses the ACT-R blending buffer rather than its retrieval buffer to access DM. All other aspects of the Canonical model remained the same. The blending module provides a mechanism for constructing chunks from relevant past

experience, in contrast to the retrieval module, which returns the most relevant chunk. It was developed for tasks that require quantitative responses (e.g., magnitude estimation) to support a form of interpolation between slot values found in relevant chunks in DM [9]. Thus the slot-values V in the constructed chunk are synthesized from slot-values V_i in corresponding slots across multiple DM chunks in some situations by minimizing their dissimilarity, weighted by the retrieval probability P_i of each chunk (see Blending Eq. 4). That retrieval probability is similar to the utility selection (2), with activation playing the role of utility, scaled by a temperature parameter.

$$V = argmin \sum\nolimits_i P_i \cdot Sim^2(V, V_i) \tag{4}$$

In terms of categorization theory, the use of the blending mechanism changes CAT from being strictly exemplar-based to a model based on a mix of exemplars and prototypes. The Blending model exhibits a gradual learning trend and an overall classification accuracy near 50%. To increase the impact of observed features (context) on the class inferences made, we increased blending temperature in the High Temp model to broaden the range of generalization. This produced about a 10% gain in performance, along with improved overall performance of around 60% and a learning trend.

An examination of class confusions in the High Temp model indicated that overgeneralization might be hurting accuracy. To correct this, the Null Rewards model adds "guessing" productions to prevent rewards from being distributed in cases where the hypothesized class changes immediately before a class decision is made, a heuristic for uncertain classifications. The addition of guessing productions increases overall accuracy to about 80%.

The Specialized model replaces the decision production that terminates the encoding process with class-specific decision productions. This supports modeling the consequences of decisions, which may be important in some domains (e.g., identifying mines vs clutter as in [10]). Accordingly, the costs/benefits of different class decisions can be represented as rewards, which would lead to liberal/conservative class decisions. The Specialized model performs like the Null Rewards model, indicating that such a representation is viable.

Examining FS in the Specialized model, we see that the number of AUs encoded decreases with experience for almost all emotions (Fig. 4). Thus CAT gradually gains efficiency while maintaining or improving classification accuracy. It is currently unclear why exemplars of surprise, the most prevalent class, are processed differently than the other classes. The heat maps in Fig. 5 indicate some overlap in the features selected for each class by CAT (right panel) and those present in the Cohn-Kanade exemplars (left panel). Thus CAT appears to be reducing the dimensions of the classification problem (in most cases), while maintaining an overall accuracy near 80%.

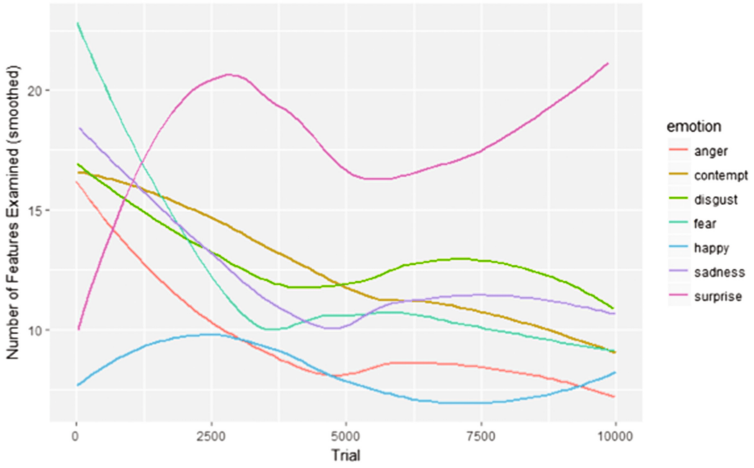


Fig. 4. Dimension reduction in the specialized model

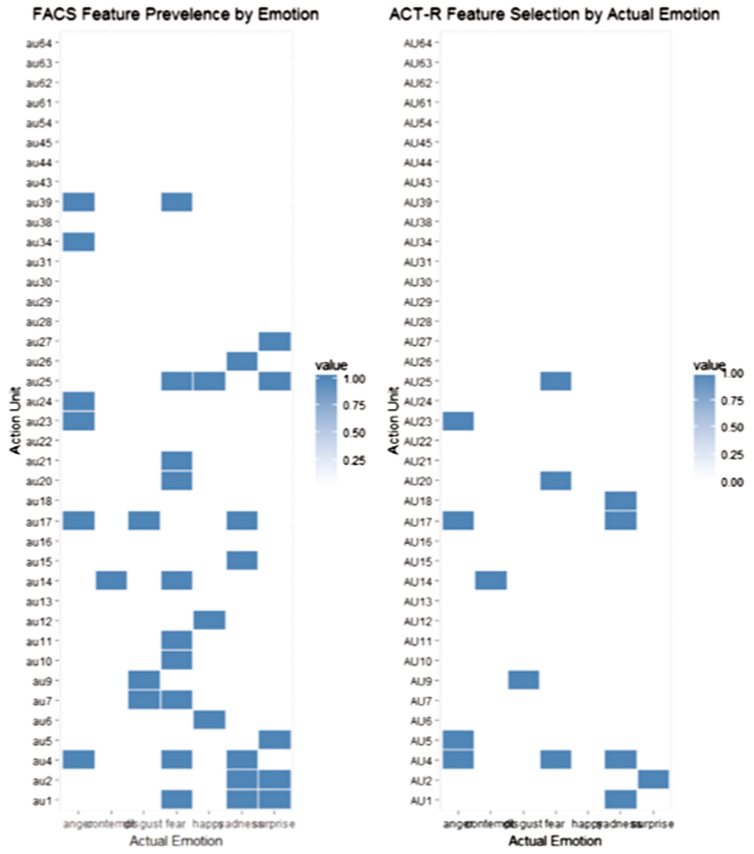


Fig. 5. Feature selection in the specialized model

4 Conclusions and Future Directions

Cognitive robotics refers to the application of psychological frameworks, theories or models of information processing to autonomous systems. The basic idea behind cognitive robotics is that we might build better autonomous systems by using insights from the amassed body of knowledge about human cognition. In turn, we might learn more about human cognition by seeing how cognitive models fare in autonomous systems facing real-world constraints.

ACT-R models that "listen" to the architecture and avoid clever, unrealistic model engineering are more likely to generalize across domains. Our categorization model relies only on ACT-R's subsymbolic learning mechanisms and simple symbolic representations. We showed that the model learned to correctly classify new stimuli while learning to select task-relevant features. By using ACT-R's blending mechanism and adjusting the reward strategy, we improved overall classification accuracy from a baseline near 15% to around 80%.

In the future we plan to compare CAT to more traditional ML approaches including deep learning approaches that simultaneously learn feature selection and classification. We also plan to analyze the model's generality by applying it to datasets from different domains.

Finally, we want to establish a feedback loop between the perceptual and cognitive systems. So far in our work, information flows from the perceptual system to the cognitive system allowing us to learn to classify stimuli based on the features observed. Information, such as confidence measures, flowing from the cognitive system could support both active perception strategies that try to confirm the presence of important features and human robot interaction that elicit help from humans to establish and name new categories.

Acknowledgements. This research was supported by OSD ASD (R&E) and by the Army Research Laboratory's Robotics Collaborative Technology Alliance.

References

1. Fields, M., Lennon, C., Liu, C., Martin, M.K.: Recognizing scenes by simulating implied social interaction networks. In: Liu, H., Kubota, N., Zhu, X., Dillmann, R., Zhou, D. (eds.) ICIRA 2015. LNCS, vol. 9246, pp. 360–371. Springer, Cham (2015). doi:10.1007/978-3-319-22873-0_32
2. Gonzalez, C., Lerch, F.J., Lebiere, C.: Instance-based learning in dynamic decision making. Cogn. Sci. **27**(4), 591–635 (2003)
3. Anderson, J.R., Betz, J.: A hybrid model of categorization. Psychon. Bull. Rev. **8**(4), 629–647 (2001)
4. Nosofsky, R.M., Palmeri, T.J.: An exemplar-based random walk model of speeded classification. Psychol. Rev. **104**(2), 266 (1997)
5. Lamberts, K.: Information-accumulation theory of speeded categorization. Psychol. Rev. **107**(2), 227 (2000)

6. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53. IEEE (2000)
7. Ekman, P.: Facial expressions. In: Handbook of Cognition and Emotion, vol. 16, pp. 301–320 (1999)
8. Cohn, J.F., Ambadar, Z., Ekman, P.: Observer-based measurement of facial expression with the facial action coding system. In: The Handbook of Emotion Elicitation and Assessment, pp. 203–221 (2007)
9. Lebiere, C.: The dynamics of cognitive arithmetic. Kognitionswissenschaft [J. Ger. Cogn. Sci. Soc.] Special issue on cognitive modelling and cognitive architectures, Wallach, D, Simon, H.A. (eds.), **8**(1), 5–19 (1999)
10. Lebiere, C., Staszewski, J.: Expert decision making in landmine detection. In: Proceedings of Human Factors and Ergonomics Society Conference, San Francisco (2010)

ENWalk: Learning Network Features for Spam Detection in Twitter

K.C. Santosh[1(✉)], Suman Kalyan Maity[2], and Arjun Mukherjee[1]

[1] University of Houston, Houston, USA
{skc, arjun}@uh.edu
[2] IIT Kharagpur, Kharagpur, India
sumankalyan.maity@cse.iitkgp.ernet.in

Abstract. Social medias are increasing their influence with the vast public information leading to their active use for marketing by the companies and organizations. Such marketing promotions are difficult to identify unlike the traditional medias like TV and newspaper. So, it is very much important to identify the promoters in the social media. Although, there are active ongoing researches, existing approaches are far from solving the problem. To identify such imposters, it is very much important to understand their strategies of social circle creation and dynamics of content posting. Are there any specific spammer types? How successful are each types? We analyze these questions in the light of social relationships in Twitter. Our analyses discover two types of spammers and their relationships with the dynamics of content posts. Our results discover novel dynamics of spamming which are intuitive and arguable. We propose *ENWalk*, a framework to detect the spammers by learning the feature representations of the users in the social media. We learn the feature representations using the random walks biased on the spam dynamics. Experimental results on large-scale twitter network and the corresponding tweets show the effectiveness of our approach that outperforms the existing approaches.

Keywords: Social network · Spam detection · Feature learning

1 Introduction

Social medias are increasing their influence tremendously. Twitter is one of the popular platforms where people post information in the form of tweets and share the tweets. Twitter is available from wide range of web-enabled services to all the people. So, the real time reflection of a society can be viewed in twitter. Celebrities, governments, politicians, businesses are active in twitter to provide their updates and to listen to the views of the people. Thus, the bidirectional flow of information is high. The openness of the online platforms and reliance on users facilitates the spammers to easily penetrate the platform and overwhelm the users with malicious intent and content. This work attempts to detect the spammers in social network using a case study of twitter.

Spammers in social networks constantly adapt to avoid the detection. Moreover, they follow reflexive reciprocity [6, 17] (users following back when they are followed by someone to show courtesy) to establish social influence and act normal. So, it is

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 90–101, 2017.
DOI: 10.1007/978-3-319-60240-0_11

becoming difficult for traditional spam detection methods to detect the spammers. Such spammers have widespread impacts. There are several reports of army of fake Twitter accounts[1] being used to troll[2] and promote political agendas[3]. Even US President Donald Trump has been accused of fake followers[4].

In this paper, we present ENWalk, a framework that uses the content information to bias a random walk of the network and obtain the latent feature embedding of the nodes in the network. ENWalk generates the biased random walks and uses them to maximize the likelihood of obtaining similar nodes in the neighborhood of the network. We study the twitter content dynamics that could be important to bias those random walks. We found that there are two types of spammers: *follow-flood* and *vigilant*. We found that success rate, activity window, fraudulence and mentioning behaviors can be used to compare the equivalence of users in the twitter. We calculate the network equivalence using these four behavioral features between pairs of nodes and try to bias the random walks with interaction proximity of the pair of nodes. Experimental results on 17 million user network from twitter show that the combination of behavioral features with the underlying network structure significantly outperform the existing state-of-the-art approaches for deception detection.

2 Related Work

There have been several works on spam detection in general, especially review spam [7], and opinion spam. However, in Twitter there are limited attempts. One of the earliest works was done by Benevenuto et al. [1]. They manually labeled and trained a traditional classifier using the features extracted from user contents and behaviors. Lee et al. leveraged profile-based features and deployed social honeypots to detect new social spammers [9]. Stringhini et al. also studied spam detection using honey profiles [14]. Ghosh et al. studied the problem of link farming in Twitter [4] and introduced a ranking methodology to penalize the link farmers. Abuse of online social networks was studied in [16]. Campaign spams was studied on [3, 10, 19].

Skip-gram model [12] has been popular to learn the features from a large corpus of data. It inspired to establish an analogy for networks by representing a network as a "document". Similar to document being an ordered sequence of words, we can create an ordered sequence of nodes from a network using sampling techniques. DeepWalk [13] learns d-dimensional feature representations by simulating uniform random walks. LINE [15] learns the d-dimensional features into two phases: d/2 BFS-style simulations and another d/2 2-hop distant nodes. Node2vec [5] creates the ordered sequence simulating the BFS and DFS approaches. All these feature learning approaches don't use the data associated with node which are important to learn the behaviors of the nodes.

[1] http://theatln.tc/2m8g3eA.

[2] http://bzfd.it/2m8rlja.

[3] http://bit.ly/2kJiMKu.

[4] http://bit.ly/1ViorHd, http://53eig.ht/2kzrhfL.

3 Dataset

For this work, we use the Twitter dataset used in [18]. It contains 17 million users having 467 million Twitter posts covering a seven month period from June 1 2009 to December 31 2009. To extract the network graph for those 17 million users, we extracted the follower-following topology of Twitter from [8] which contains all the entire twitter user profiles and their social relationships till July 2009. We pruned the users so that they have social relationship in [8] and tweets in [18] and are left with 4,405,698 users.

Twitter suspends the accounts involved in the malicious activity (https://support.twitter.com/articles/18311). To obtain the suspend status of accounts, we re-crawled the profile pages of all the 17 million users. This yielded a total of 100,758 accounts that had been suspended (the profile page redirects to the page https://twitter.com/account/suspended). We use this suspension signal as the primary signal for evaluating our models as the primary reason for account suspension is the involvement in the spam activity. However, there might be other reasons like inactivity. So, to ensure the suspended accounts are spammers, we further checked for malicious activities for those users. For this, we examined various URLs from the account's timeline and checked them against a list of blacklisted URLs. We use three blacklists: Google Safebrowsing (http://code.google.com/apis/safebrowsing/), URIBL (http://uribl.com/) and Joewein (http://www.joewein.net/). We found that 75% of suspended accounts posted at least one shortened URL blacklisted. We also looked for duplicate tweets enforced for promotion. After applying these additional criteria, our final data comprised of 86,652 spammers and 4,319,046 non-spammers, which was used for evaluating our model.

4 Spam Analysis

Characterizing the dominant spammer types is important as it is the first step in understanding the dynamics of spamming. We studied the follower-following network creation strategies of the spammers. We found that there are two main types of spamming based on the follow-following strategies: (1) *follow-flood* spammers and (2) *vigilant* spammers. So, the question arises why some spammers are more successful? In this section, we study the behavioral aspects of tweet dynamics of spammers. We later leverage them in model building.

4.1 Spammer Type

To analyze the strategies of follower-following, we calculated the number of followers (users that are following the current user) and the number of followings (users that the current user is following) for each spammer. Figure 1 shows the plot in log scale count. It shows that the follower and following count differ for each spammers. The users with more followers than followings tend to be more successful as they have been able to "earn" a lot of users who are following them. So, we define success rate as:

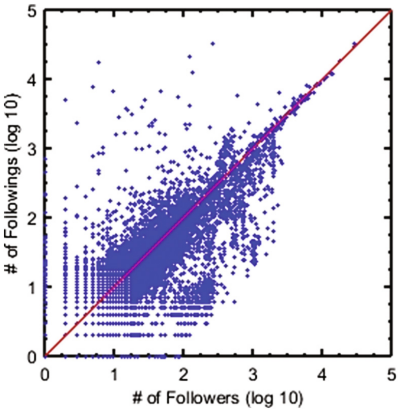


Fig. 1. Follower-Following Count of Spammers. Each Blue dot represents a spammer and the red line is the plot of y = x line. (Color figure online)

$$sr_u = \frac{\#\text{ of followers of } u}{\#\text{ of followings of } u} \tag{1}$$

Based on the network expansion success rate, we find that there are two dominant spamming strategies:

- *Follow-flood* Spammers: These are less successful spammers who just flood the network with friendship initiation so as to get followers who they can influence. We categorize the spammers with success rate (sr_u) less than 1 in this type.
- *Vigilant* Spammers: These are successful spammers who take a cautious approach of friendship creation and content posting. Spammers with success rate (sr_u) greater or equal to 1 are categorized as *vigilant*.

To learn the dynamics of each spammer type, we further analyzed the success rate of spammers with other behavioral aspects – activity window, usage of promotion words or blacklist words and hashtag mentioning.

4.2 Activity Window

We compute the activity window as the number of days a user is active in the twitter network. Since, we don't have the exact time when a user was suspended, we approximate the time of suspension as the date of the last tweet tweeted by the user. We found that the average activity window of a *vigilant* spammer is 138 days with a standard deviation of 19 days compared to the average of 35 days and standard deviation of 12 days for *follow-flood* spammers. Although, the basic strategy of any spammer is to inject itself into the network and emit the spam contents, the success rate also depends how long it can remain undetected in the network. So, *vigilant* spammers have a higher success rate.

4.3 Fraudulence

One of the primary reason to spam is to inject constant fraudulence information. So, we analyzed the fraudulence behavior of the two types of spammers. We labeled the tweets containing promotional, adult words or the blacklisted urls as fraud tweets. So, we compute fraudulence as:

$$fr_u = \frac{\#\text{ of fraud tweets of } u}{\text{total }\#\text{ of tweets of } u} \tag{2}$$

We found that the average fraudulence of *vigilant* spammers is 0.34 compared to 0.86 of *follow-flood* spammers. So, the follow-flood spammers are more involved in spam.

4.4 Mentioning Celebrities and Popular Hashtags

Mentioning the popular celebrities or hashtags empowers a tweet. So, one of the common strategies of spammers is to include the popular ones in their tweets. We studied mentioning phenomenon and found that *vigilant* spammers mention half the celebrities per tweets compared to the *follow-flood* spammers.

5 Learning Latent Features for Spam Detection

Having characterized the dynamics of spamming in Twitter, can we improve spam detection beyond the existing state-of-the-art approaches? To answer this we used our Twitter data to setup a latent feature learning problem in networks. Our analysis is general and can be used to any social network.

5.1 Overview

As discussed in the previous section, the dynamics of Twitter are interesting and can be leveraged to catch the spammers. So, we use the spam dynamics to formulate the latent feature learning in social networks. Let $G = (V, E, X)$ be a given network with vertices, edges and the social network data of users in the social network. We aim to learn a mapping function $f : V \rightarrow \mathbb{R}^d$ from nodes to a d-dimensional feature representations which can be used for prediction. The parameter d specifies the number of dimensions of the latent features such that the size of f is $|V| \times d$.

We present a novel sampling strategy that samples nodes in network exploiting the spam dynamics such that the *equivalent neighborhood* $EN(u) \subset V$ contains the node having similar tweeting behaviors with the node u. We generate $EN(u)$ for each nodes in the network and predict which nodes are the members of u's equivalent neighbors based on the learnt latent features f. The basic rationale is that we wish to learn latent feature representations for nodes that respect equivalent neighborhoods (which are based on the spamming dynamics) so that classification/ranking using the learned representation yields results that leverage the spamming dynamics.

5.2 The Optimization Problem

As our goal is to learn the latent features f that best describe the equivalent neighborhood $EN(u)$ of node u, we define the optimization problem as follows:

$$\max_{f} \sum_{u \in V} \log Pr(EN(u)|f(u)) \tag{3}$$

To solve the optimization problem, we extend the SkipGram architecture [5, 13, 15] which approximates the conditional probability using an independence assumption that the likelihood of observing an equivalent neighborhood node is independent of observing any other equivalent neighborhood given the latent features of the source node.

$$Pr(EN(u)|f(u)) = \prod_{v \in EN(u)} \Pr(v|f(u) \tag{4}$$

Since, the source node and the equivalent neighborhood node have symmetric equivalence, the conditional likelihood can be modeled as softmax unit parameterized by a dot product of their features.

$$\Pr(v|f(u)) = \frac{exp(f(v).f(u))}{\sum_{t \in V} exp(f(t).f(u))} \tag{5}$$

The optimization problem now becomes:

$$\max_{f} \sum_{u \in V} \left[-\log Z_u + \sum_{t \in EN(u)} f(t).f(u) \right] \tag{6}$$

For large networks, the partition function $Z_u = \sum_{t \in V} \exp(f(t).f(u))$ is expensive to compute. So, we use negative sampling [12] to approximate it. We use stochastic gradient descent over the model parameters defining the features f. Feature learning methods based on Skip-gram architecture are developed for natural language [11]. Since natural language texts are linear, the notion of a neighborhood can be naturally defined using a sliding window over consecutive words in sentences. Networks are not linear, and thus a richer notion of a neighborhood is needed. To mitigate this problem, we use multiple biased random walks each one in principle exploring a different neighborhood [5].

5.3 Equivalent Neighborhood Generation

The analyses of spam dynamics leads to an important inference that the nodes are similar if they have similar spam dynamics. So, we want to exploit those dynamics to generate the equivalent neighborhood $EN(u)$ for the node u. Nodes in a network are equivalent if they share similar behaviors. We use the random walk procedure which can be biased to generate the equivalent neighborhood.

We bias the random walks based on the four dynamics: common time of activity (ct_{tv}), success rate difference (sr_{tv}), fraudulence commonalities (fr_{tv}) and common mentioning in tweets (me_{tv}). We calculate each dynamics as follows:

$$ct_{tv} = \frac{\#\text{ of days with common activity}}{\#\text{ of days either } t \text{ or } v \text{ is active}} \tag{7}$$

$$sr_{tv} = 1 - \left| \max\left(1, \frac{\#\text{ of followers of } t}{\#\text{ of followings of } t}\right) - \max\left(1, \frac{\#\text{ of followers of } v}{\#\text{ of followings of } v}\right) \right| \tag{8}$$

$$fr_{tv} = 1 - \left| \frac{\#\text{ of fraud tweets of } t}{\#\text{ of tweets of } t} - \frac{\#\text{ of fraud tweets of } v}{\#\text{ of tweets of } v} \right| \tag{9}$$

$$me_{tv} = \frac{\text{common mentions between } t \text{ and } v}{\text{total mentions of } t \text{ and } v} \tag{10}$$

For all the above four features, a higher value represents a closer connection between the pair of nodes. For a source node u, we generate a random walk of fixed length k. The i^{th} node c_i of a random walk starting at node c_0 is generated with the distribution:

$$P(\mathrm{c}_i = t | \mathrm{c}_{i-1} = v) = \begin{cases} \mathcal{B}_{vt}, if(v,t) \in E \\ 0, otherwise \end{cases} \tag{11}$$

where $\mathcal{B}_{vt}$ is the normalized transition probability between nodes v and t. The transition probability are computed based on the spam dynamics so that the source node has equivalent spam dynamics with its neighborhood nodes.

Algorithm 1: ENWalk $(G, d, \lambda, l, k, [p, q, r, s])$

Input: graph $G(V, E, W, X)$
embedding dimensions $\boldsymbol{d}$
walks per node λ
walk length l
context size k
tweet parameters $\boldsymbol{p}, \boldsymbol{q}, \boldsymbol{r}, \boldsymbol{s}$
Output: matrix of latent features $\boldsymbol{F}$

```
(CT, SR, FR, ME) = Preprocess(G, p, q, r, s)
Initialize walks to empty
for i = 1 to λ do
    for each v_i ε V do
        Initialize walk to v_i
        for j = 1 to l do
            x = GetEquivalentNeighbor(G, CT, SR, FR, ME, walk[j], W)
            Append x to walk
        Append walk to walks
F = StochasticGradientDescent(k, d, walks)
```

We define four parameters which guide the random walk. Consider that a random walk just traversed edge (t, v) to now reside at node v. The walk now needs to decide on

the next step so it evaluates the transition probabilities on edges (v, x) leading from v. We set the transition probability to $\mathcal{B}_{vx} = \alpha_{pqrs}(t, v, x).w_{vx}$, where

$$\alpha_{pqrs}(t, v, x) = p.(ct_{tv} + ct_{vx}) + q.(sr_{tv} + sr_{vx}) + r.(fr_{tv} + fr_{vx}) + s.(me_{tv} + me_{vx}) \quad (12)$$

where the parameters p, q, r, s are used to prioritize the tweet dynamics. To select the next node, the random walk is biased towards the nodes which have similar tweet dynamics to both the current node and the previous node in the random walk.

5.4 Algorithm: ENWalk

Algorithm 1 details our entire scheme. We start with λ fixed length random walks at each node l times. To obtain each walk, we use *GetEquivalentNeighbor*, the random sampler that samples the node based on the transition probabilities computed in Eq. 12. It is worth noting that the tweet dynamics between the nodes (CT, SR, FR, ME) defined in Eqs. 7, 8, 9 and 10 respectively can be pre-computed. Once, we have random walks we can obtain d dimensional numeric features using the optimization function in Eq. 6 with a window size of k. The three phases preprocessing, random sampling and optimization are asynchronous so that ENWalk is scalable.

6 Experiment

We applied ENWalk to twitter dataset to evaluate its effectiveness. In this section, we discuss the baseline methods and compare with ENWalk for classification and ranking.

6.1 Baseline Methods

For classification, we compare our model with two graph embedding methods: Deepwalk and node2vec. We use PageRank and Markov Random Field (MRF) approaches for ranking of spam nodes. We did not use feature extraction techniques like [1] as they only use the node features without using the graph structure.

Table 1. Propagation Matrix for (S)pammer, (M)ixed, (N)on-Spammer

	S	M	N
S	0.80	0.40	0.025
M	0.15	0.50	0.125
N	0.05	0.10	0.850

Deepwalk [13]. It is the first approach to integrate the language modeling for network feature representation. It generates uniform random walks equivalent to sentences in the language model.

Node2vec [13]. It is another representation learning for nodes in the network. It extends the language model of random walks employing a flexible notion of neighborhood. It designs a biased random walk using BFS and DFS neighborhood discovery.

PageRank Models. PageRank is a popular ranking algorithm that exploits the link-based structure of a network graph to rank the nodes of the graph.

$$PR = (1 - \alpha) * M * PR + \alpha * p \tag{13}$$

where M is transition probability matrix, p represents the prior probability with which a random surfer surfs to a random page and α is damping factor. For variations of PageRank, we vary the values of M and p using trustworthiness of a user. Trustworthiness (f_{Trust}) is using a set of features (# of Blacklist URL, # of tweets, # of mentions, # of duplicate tweets, # of tweets containing adult/bad words, # of tweets containing violent words, # of tweets containing promotional words and the total time of activity for the user). We manually labeled f_{Trust} score of 800 users (400 non-suspended and 400 suspended). We gave a real-valued trustworthiness score between 0 and 1. A value closer to 0 means the user is most likely a spammer. We then obtain the weight of the features by learning linear regression model on the users.

- **Traditional PageRank:** We use the default PageRank settings for M and p.
- **Trust Induced and Trust Prior:** Transition matrix M is modified as $M_{uv} = \mathrm{M}_{uv} * f_{Trust}(v), \forall u, \forall v$ and $f_{Trust}(v)$ is used as prior probability.

Markov Random Field Models. Markov Random Fields are undirected graphs (and can be cyclic) that satisfy the three conditional independence properties (Pairwise, Local, and Global). For the inference, we use the Loopy Belief Propagation algorithm. Inspired by spam detection in [2], we define 3 hidden states {Spammer, Mixed, Non-Spammer} and the Propagation Matrix is used as in Table 1. Logically, spammers follow other spammers more (hence 0.8 probability) and non-spammers tend to follow other non- spammers. We also include the mixed state to include those users who are difficult to categorize spammers or non-spammers.

6.2 Node Classification

We obtained the feature representations from three different algorithms: ENWalk, node2vec and DeepWalk using the settings used in node2vec and DeepWalk. All the feature learnings are unsupervised. Similar to node2vec and DeepWalk, we used $d = 128, \lambda = 10, l = 80, k = 10$. We found that the parameters d, λ, l, k are sensitive in a similar style to node2vec and DeepWalk. We used each feature representation as an example for standard SVM classifier. We used 10-fold cross-validation using balanced data obtained from sub-sampling of the negative class. From the classification results in Table 2, ENWalk performs better. It has higher precision, recall, F1-score and accuracy due to the biased random walks.

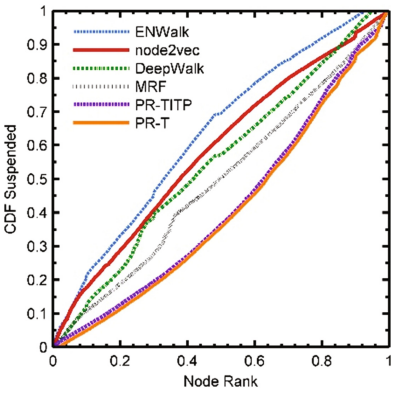


Fig. 2. Cumulative distribution function of suspended nodes

Table 3. Ranking results: area under CDF curve (AUC) and Precision@100(P@100)

Model	AUC	P@100
PR-T	0.4059	0.02
PR-TITP	0.4181	0.03
MRF	0.4944	0.02
DeepWalk	0.5502	0.05
node2vec	0.5836	0.05
ENWalk	**0.6335**	**0.12**

Table 2. Classification Results: Precision (P), Recall (R), F1-score (F) and Accuracy (A)

Model	P	R	F	A
DeepWalk	0.44	0.49	0.46	0.51
Node2vec	0.46	0.53	0.49	0.57
ENWalk	0.59	0.66	0.62	0.71

6.3 Node Ranking

We use two metrics to evaluate the ranking results: *Cumulative Distribution Function of Suspended Users* and *Precision@n*. We rank all the nodes in the graph and provide a node rank percentile. For each node rank percentile, we compute the number of suspended users in that percentile. We plot the cumulative distribution function for those suspended users. We also calculate the Area Under Curve (AUC) for the CDF. The higher the area the better the model. Precision@n of Suspended Users evaluates how many top n nodes suggested by a model are actually the suspended users. This is effective to screen the nodes that are probable being spammers.

To evaluate the ranking performance of ENWalk, we use Logistic Regression on the features obtained from the model. We compare our model with PageRank and Markov Random Field models. We present the CDF in Fig. 2. We can see that ENWalk outperforms all the baseline models. We also computed the AUC and precision@100 (Table 3). A higher AUC and precision@100 signifies the ability to profile the top spammers.

7 Conclusion

We studied the problem of identifying spammers in Twitter who are involved in malicious attacks. This is very much important as it has many practical applications in today's world where almost everyone is actively social online. This paper proposed a method of spam detection in Twitter that makes use of the online network structure and information shared. This data driven approach is important as there is a lot of data of social medias online these days. We demonstrated the helpfulness of biased random walks in learning node embedding that can be used for classification and ranking tasks.

Acknowledgements. This work is supported in part by NSF 1527364. We also thank anonymous reviewers for their helpful feedbacks.

References

1. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, Electronic Messaging, Anti-abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
2. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. In: Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, 8–11 July 2013
3. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, pp. 35–47 (2010)
4. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, K.P.: Understanding and combating link farming in the twitter social network. In: Proceedings of the 21st International Conference on World Wide Web, pp. 61–70 (2012)
5. 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 (2016)
6. Hu, X., Tang, J., Zhang, Y., Liu, H.: Social spammer detection in microblogging. IJCAI **2013**, 2633–2639 (2013)
7. K C, S., Mukherjee, A.: On the temporal dynamics of opinion spamming: case studies on yelp. In: 25th International World Wide Web Conference, WWW 2016, Montréal, Québec, Canada, 11–15 April 2016
8. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: The International World Wide Web Conference Committee (IW3C2), pp. 1–10 (2010)
9. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+machine learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–442 (2010)
10. Li, H., Mukherjee, A., Liu, B., Kornfield, R., Emery, S.: Detecting campaign promoters on twitter using markov random fields. In: 2014 IEEE International Conference on Data Mining, ICDM, Shenzhen, China, pp. 290–299, 14–17 December 2014
11. Mikolov, T., Chen, K., Corrado, G. and Dean, J.: Distributed representations of words and phrases and their compositionality. Nips, pp. 1–9 (2013)
12. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations (ICLR 2013), pp. 1–12 (2013)
13. 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 (2014)
14. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9 (2010)
15. 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 (2015)
16. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 243–258 (2011)

17. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010), pp. 261–270 (2010)
18. Yang, J., Leskovec, J.: Patterns of temporal variation in online media. WSDM **2011**, 177 (2011)
19. Zhang, X., Zhu, S., Liang, W.: Detecting spam and promoting campaigns in the Twitter social network. In: Proceedings - IEEE International Conference on Data Mining, ICDM, pp. 1194–1199 (2012)

Understanding Russian Information Operations Using Unsupervised Multilingual Topic Modeling

Peter A. Chew(✉) and Jessica G. Turnley

Galisteo Consulting Group, Inc., 4004 Carlisle Blvd NE Suite H, Albuquerque, NM 87107, USA
{pachew,jgturnley}@galisteoconsulting.com

Abstract. What does this or that population think about a given issue? Which topics 'go viral' and why? How does disinformation spread? How do populations view issues in light of national 'master narratives'? These are all questions which automated approaches to analyzing social media promise to help answer.

We have adapted a technique for multilingual topic modeling to look at *differences* between what is discussed in Russian versus English. This kills several birds with one stone. We turn the data's multilinguality from an impediment into a leverageable advantage. But most importantly, we play to unsupervised machine learning's strengths: its ability to detect large-scale trends, anomalies, similarities and differences, in a highly general way.

Applying this approach to different Twitter datasets, we were able to draw out several interesting and non-obvious insights about Russian cyberspace and how it differs from its English counterpart. We show how these insights reveal aspects of how master narratives are instantiated, and how sentiment plays out on a large scale, in Russian discourse relating to NATO.

Keywords: Information operations · Topic modeling · Multilingual · Russia

1 Introduction

In the current geopolitical climate, and with growing use of social media, there is widespread interest in questions such as: What does this or that population think about a given issue? Which topics go viral and why? How does disinformation spread? How do populations view issues in light of national or religious 'master narratives'? The vast open-source trove that is social media has not only pushed these questions high on the agenda; it also begs automated solutions to providing answers.

Most approaches, such as sentiment analysis [2, 3, 7], deception detection [6], topic modeling [4], or master narratives analysis [8], focus in narrowly on one or another of these questions. In this paper, we describe an approach that attempts to balance many of these issues simultaneously. That may sound like a grand claim, but to be clear, we are not trying to subsume all the achievements of previous, focused work. Instead, we outline a role we think is eminently suited to *unsupervised machine learning* techniques such as [1]: allowing top-down exploration of multilingual social media not just to find trending topics cross-linguistically, but also to aid understanding of *differences* between different important subsets of the data, for example the subsets of social media data

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 102–107, 2017.
DOI: 10.1007/978-3-319-60240-0_12

represented by English- versus Russian-language posts. Via examples, we show how exploring such differences can lead to significant, non-obvious insights about how different populations view the world. Compared to most existing work, our unsupervised approach essentially reallocates the labor between human and machine: the machine does what it is best at – finding patterns, similarities and differences, leaving the higher-order analysis to the human, but making the human's task easier by focusing the human's attention in first and foremost on the most material patterns and prominent in the data, and helping a human avoid getting 'lost in the weeds'.

This paper is structured as follows. In Sect. 2 we briefly describe STEMMER, our approach to multilingual topic modeling, explaining how it can be adapted to review differences between subsets of the data ('information spaces'). Section 3 presents the results of applying this approach to two Twitter datasets, each comprising over 100,000 English-, Russian- and Ukrainian-language posts. We show a number of examples of non-obvious, and potentially significant and actionable, insights that our approach allows us quickly to tease out of the data. Finally, we conclude in Sect. 4.

2 STEMMER: A Framework for Unsupervised Analytics

We examine unsupervised pattern recognition on multilingual social media, and do not pursue supervised or similar analyses in this paper. We call the unsupervised-analysis approach that we have tailored specifically to multilingual data STEMMER (System for Top-down Exploration of Mixed Multilingual Electronic Resources), described in detail in [5]. STEMMER, a modified version of standard Latent Semantic Analysis (LSA), differs from other topic modeling approaches in that it induces *cross-language* topics, as in Fig. 1 (ibid). LSA takes any collection of text and derives prominent topical patterns from that text deterministically: the same input always produces the same output, and the top n topics are guaranteed to be the n most prominent topics in the corpus. This property qualifies LSA eminently well as a 'top-down' approach, one that helps analysts not miss the forest for the trees. STEMMER adds a linear-algebra pre-processing step, optimized for LSA, shoe-horning all documents into a single multilingual space. Recommending STEMMER is that it has been empirically validated: by fine-tuning, we have brought its accuracy up to 94.8% [9].

Doc #	Source text
10433	... RT @AFP: UPDATE: 2,000 Russian soldiers land in \
570	Russia seizes control of Crimea! .... #Ukraine http://t.co/hDjGFvOfoc
8082	Casi 2,000 soldados rusos han aterrizado en las ultimas horas en Crimea…

Fig. 1. A topical pattern from multilingual Twitter data

In the work cited [5, 9] the focus is on using STEMMER to identify groups of topically similar documents, where similarity is determined ultimately by the words used in each document. This is in itself a useful function for analysts, because it helps analysts

sort the data into 'buckets' that make sense, and again not miss the forest for the trees. But STEMMER can also be used to look at *differences* between different areas of the 'forest'. Most digital data, including Twitter, has an abundance of pre-supplied or easily inferable metadata, e.g. geospatial coordinates, language, author, date and time, etc. This metadata supplies obvious ways to subdivide the data.

Guided by how the Russian government talks about 'information spaces', we choose here to look at gross differences by *language*. Framed this way, the problem becomes one which unsupervised learning, and STEMMER in particular, is ideally placed to handle. We can use STEMMER to detect the biggest topical *differences* between Russian and English. Technically, the approach to finding such high-level topical differences is straightforward, once we have the STEMMER framework in place. A key output of STEMMER is a document (Twitter post) by topic matrix. Each cell in that matrix encodes how 'strong' a given topic is in a given document. Since we know the metadata (its language) for each document a priori, we can subdivide the matrix into blocks by language, as shown in Fig. 2. Then, it is simple, intuitive, and justifiable in linear-algebraic terms, to calculate, for each topic, a 'strength' of that topic by language: to do this, we can calculate for each block the sum of the squared cell values (in linear algebra terms, the magnitude) in that block.

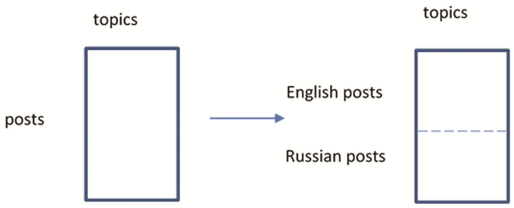


Fig. 2. Subdivision of document-by-topic matrix into blocks

The final step is to sort the topics by 'magnitude variance', as shown in Table 1. We can then explore the topic further by looking at what terms and documents most characterize that topic. This gives insight into *how*, topically, differences between one information space and another manifest themselves.

Table 1. Topics sorted by 'language magnitude variance'

Topic #	English magnitude	Russian magnitude	Magnitude variance
1	0.0653	0.1372	0.0719
2	0.0825	0.1261	0.0436
…	…	…	…

3 Results

We used the approach above to explore two separate datasets, each a sample of posts from the Twitter 'firehose' and within specified date ranges, shown in Table 2.

Table 2. Datasets analyzed

#	Dataset description	Date range	Seed words	# of posts by language		TOTAL
1	'Brilliant Jump' NATO exercises	May–June 2016	NATO, HATO	EN RU	127,220 79,514	**206,734**
2	NATO discussion in 2016 US election	10/13/2016-11/28/2016	NATO, HATO	EN RU	81,969 79,465	**161,434**

We used STEMMER to analyze the top 90 cross-linguistic topics for each dataset. We use a multilingual parallel corpus for the pre-processing step to 'multilingualize' each post, a process described in detail in [9]. We manually augmented that parallel corpus with the top 100 most frequent out-of-vocabulary words in each language, a process that takes an hour or two for someone with relevant language expertise. The ability of STEMMER to return multilingual topics is not absolutely dependent on this kind of manual intervention but is nonetheless improved with a relatively small investment of effort. 'Out-of-vocabulary words' tend to be proper names (e.g. Putin/Путин) or technical terms (e.g. генсек, Russian acronym for 'general secretary').

3.1 'Brilliant Jump' NATO Exercises Dataset

Language weighting plausibility: Topic #6, highly weighted towards English, included as top keywords: Корею (Korea), blast, Korea, Клинтон (Clinton), north, Clinton, речь (speech), speech. Twitter posts highly representative of Topic 6 were:

- Clinton To Blast Trump On North Korea, NATO In Foreign Policy Speech😂😂😂
- Reuters: Clinton to blast Trump on North Korea, NATO in foreign policy speech

From just the above it should be clear the common thread in this topic is Clinton and Trump's discussion of North Korea with respect to NATO. It is plausible that in the context of a dataset from May–June 2016, this topic is 'weighted toward' the English-language information space, for no other reason than that relatively few Russian speakers had much interest in the US Presidential election this early in the campaign.

Similarly, topic #5 was one of those most highly weighted towards Russian, we believe for a similar reason. Top keywords for this topic were: secretary, Poroshenko, Порошенко (Poroshenko), советником (advisor), генсек (secretary), ex, экс (ex), генсека (secretary), назначил (appointed). An English-language Twitter post most representative of this topic was 'Poroshenko appointed the ex-secretary of NATO as his advisor'. Poroshenko, the President of Ukraine, is relatively unknown to English speakers and it is thus plausible that few English-speaking Twitter users discuss him.

Result of interest: Topic #11 (Russian weighting .1152, English weighting .0992) was also among the topics with greatest magnitude variance, and included these top keywords: Russia, Poland, strike, global, глобального (global), удара (strike). A Twitter post highly weighted in Topic 11 was '#News. In the [Russian] Federation Council an announcement was made about a "global strike" by NATO on Russia/#Russia.'* (asterisks here and below denote our translation from original Russian).

With some background knowledge of NATO activities, we were aware that 'global strike' is a defensive/deterrent capability that NATO developed to strike back anywhere in the world within an hour against an emerging threat. Whether as a result of deliberate disinformation, a misconstruing of the true nature of 'global strike', attempts by Kremlin-controlled trolls to whip up fear, or something else, the Russian-language posts that are most representative of topic 11 seem to misinterpret 'global strike' as a direct threat to Russia, an intent of NATO to attack Russia. Further, with some background knowledge of Russian discourse and use of master narratives, we can say that this may be an instantiation of the 'Fortress Russia' master narrative that claims that Russia is under global threat by outsider adversaries [10].

To verify that this result was a real phenomenon in the data, we counted the posts in the dataset mentioning both 'global' and 'strike' (and inflected forms in Russian), by language. It turned out that only 78 English posts mentioned the two words, while 2,541 Russian posts did, confirming the reality of the phenomenon STEMMER automatically found as an emergent property of the data.

3.2 2016 Presidential Election Dataset – Results of Interest

Two topics that we thought were of most interest in this dataset were #25 (weighted towards Russian) and #22 (weighted towards English). Top keywords for topic #25 include 'Syria' and 'Lavrov'. A representative Twitter post highly weighted in this topic was 'Lavrov called NATO troops' bombing of Yugoslavia aggression'*.

It was initially unclear to us what the 1990s NATO bombing of Yugoslavia had to do with Syria (a top keyword for this topic) and 2016. However, a quick internet search led us to https://www.youtube.com/watch?v=kSXaAU-szqw, a short interview where Sergei Lavrov (Russian foreign minister) argues Russia's bombing of Aleppo, Syria, is no different from NATO's bombing of Yugoslavia. So here, STEMMER helped direct us to part of the Russian narrative of which we were unaware.

Finally, topic #22, weighted towards English, contained numerous posts similar to the following: 'While you were focused on the Walking Dead, NATO and U.S. marines prep for Russian war'. We were again unsure of the connection between NATO, Russia and the 'Walking Dead'. An internet search for 'walking dead NATO' turned up sites like thefreethoughtproject.com, russia.trendolizer.com, and thefringenews.com. When we opened some of these, unexpected pop-ups led us to suspect strongly the sites might be infected with malware and so we quickly curtailed further investigation. Were these sites themselves part of the Russian multi-pronged information warfare strategy? Possibly – and this too could be of interest to an analyst.

4 Conclusion

In this paper we outline a way to turn the multilinguality of social media content, usually an obstacle for analysis, into an asset; and at the same time a way also to take Russian strategic thinking as a starting-point for deconstructing Russian 'information warfare' so that analysts can better understand its vectors of attack and thus be in a better position

to develop countermeasures. The method is top-down, focusing first on the most important trends and patterns, and also on the most important *differences* between different natural subsets of the social media universe, for example Russian versus English content. The value of our approach is that it helps an analyst quickly see what is most important in hundreds of thousands of posts without being burdened by reading and translating many individual posts – essentially, it helps analysts see the forest for the trees. This technique has been empirically validated, although our focus in this paper has been not on the empirical validation which is discussed elsewhere, but on demonstrating the usefulness and plausibility of its results by example.

We believe the approach we have taken here points the way to best practices in data analytics of this type: keep the human fully engaged in the loop, and cleanly separate between areas of responsibility for human and computer: have the machine do what it is best at – finding patterns, and groups of things that are 'similar' or 'different' – but let the human effectively focus his energies on what humans do best: relating those patterns to subject-matter expertise, such as prior knowledge of country-specific dynamics. This, we believe, is the best way to ensure that neither the human nor the computer are allowed to be sidetracked by confirmation bias or other kinds of systemic bias. The human's attention should always be kept focused on what is most important *within a particular dataset*, and our approach is designed to do just that.

References

1. Duda, R.O., Hart, P.E., Stork, D.G.: Unsupervised learning and clustering. In: Pattern Classification, 2nd edn. Wiley, New York (2001). ISBN: 0-471-05669-3
2. Kim, S.-M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004), pp. 1367–1373 (2004)
3. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79–86 (2002)
4. Bader, B.W., Berry, M.W., Browne, M.: Discussion tracking in Enron email using PARAFAC. In: Berry, M.W., Castellanos, M. (eds.) Survey of Text Mining II, pp. 147–163. Springer, London (2008)
5. Chew, P.A.: 'Linguistics-Lite' topic extraction from multilingual social media data. In: Agarwal, N., Xu, K., Osgood, N. (eds.) SBP 2015. LNCS, vol. 9021, pp. 276–282. Springer, Cham (2015). doi:10.1007/978-3-319-16268-3_30
6. Tsikerdekis, M., Zeadally, S.: Online deception in social media. Commun. ACM **57**(9), 72–80 (2014)
7. Center for Computational Analysis of Social and Organizational Systems: Multilingual Twitter sentiment analysis (2016). http://www.casos.cs.cmu.edu/projects/projects/mltsa.php. Accessed 27 July 2016
8. Halverson, J., Corman, S., Goodall, H.: Master Narratives of Islamist Extremism. Macmillan, New York (2011)
9. Chew, P.: Multilingual retrieval and topic modeling using vector-space word alignment. Galisteo Consulting Group, Inc. Technical report GCG002, February 2016. doi:10.13140/RG.2.2.21482.11205
10. Bouveng, K.: The role of messianism in contemporary Russian identity and statecraft. Durham Theses, Durham University (2010). http://etheses.dur.ac.uk/438

Social Cyber Forensics Approach to Study Twitter's and Blogs' Influence on Propaganda Campaigns

Samer Al-khateeb(✉), Muhammad Nihal Hussain(✉), and Nitin Agarwal(✉)

Department of Information Science, University of Arkansas at Little Rock, Little Rock, AR 72204, USA
{sxalkhateeb,mnhussain,nxagarwal}@ualr.edu

Abstract. In today's information technology age our political discourse is shrinking to fit our smartphone screens. Online Deviant Groups (ODGs) use social media to coordinate cyber propaganda campaigns to achieve strategic and political goals, influence mass thinking, and steer behaviors. In this research, we study the ODGs who conducted cyber propaganda campaigns against NATO's Trident Juncture Exercise 2015 (TRJE 2015) and how they used Twitter and blogs to drive the campaigns. Using a blended Social Network Analysis (SNA) and Social Cyber Forensics (SCF) approaches, "anti-NATO" narratives were identified on blogs. The narratives intensified as the TRJE 2015 approached. The most influential narrative identified by the proposed methodology called for civil disobedience and direct actions against TRJE 2015 specifically and NATO in general. We use SCF analysis to extract metadata associated with propaganda-riddled websites. The metadata helps in the collection of social and communication network information. By applying SNA on the data, we identify influential users and powerful groups (or, focal structures) coordinating the propaganda campaigns. Data for this research (including blogs and metadata) is accessible through our in-house developed Blogtrackers tool.

Keywords: Cyber propaganda campaign · Misinformation · NATO · Trident Juncture Exercise · Narrative · Influence · Blogs · Twitter · Social media · Social Network Analysis · Cyber forensics · Blogtrackers · Social Cyber Forensics

1 Introduction

The inexpensive nature, ease of use, and the popularity of social media makes it a powerful tool that can be used to disseminate misinformation or coordinate cyber propaganda campaigns in order to influence mass thinking and steer behaviors or perspectives about an event. We investigate these phenomena in this research. Social media provides a rich source of information [1]. With millions of social network users around the globe, cyber forensic analysis of social media has profound applications [2]. Cyber forensic analysis of social media can help collect evidence that helps investigators develop a strong case [1]. Cyber Forensics (CF) is "the process of acquisition, authentication, analysis, and documentation of evidence extracted from and/or contained in a computer system, computer network, and digital media" [3]. With the use of metadata, extracted

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 108–113, 2017.
DOI: 10.1007/978-3-319-60240-0_13

using cyber forensics the relationship between deviant groups can be discovered. In this work, we identify and study the behavior of these ODGs and how they use social media to coordinate cyber propaganda campaigns using SNA and SCF techniques. We define ODGs as a collective that organizes a harmful activity using cyber space in which its result would affect cyber space, physical space or both, i.e., the "Cybernetic Space" [4]. We also develop methodologies to identify influential narratives in a cyber campaign. We use Maltego (available at: http://bit.ly/1Vm00JS) to conduct SCF analysis and enhance the collected data. We use SNA in combination with the metadata extracted from SCF analysis to have a comprehensive understanding of the propaganda campaign coordination. For conducting SNA we use NodeXL (available at: https://nodexl.codeplex.com) and Focal Structure Analysis (FSA) (available at: www.merjek.com). FSA helps discover an influential group of individuals in a large network. These individuals are connected and may not be the most influential individually, but by acting together they form a compelling power. We chose this approach because it was tested on many real world events including the Saudi Arabian women's right to drive campaign on Twitter (Oct26Driving), and the 2014 Ukraine Crisis when President Viktor Yanukovych rejected a deal for greater integration with the European Union [5]. For analyzing blog data, we use Blogtrackers (available at: http://blogtrackers.host.ualr.edu/).

This research has implications not only to the scientific community, but also for authorities as these ODGs pose non-negligible concerns for public safety and national security, e.g., one of the influential narratives in the data collected in this study called for civil disobedience, planned protests, or direct actions against the TRJE 2015 exercise. Therefore, we study: (1) Who are the important information actors in the campaign network? (2) What is the role of each social media channel in the propaganda campaign coordination? (3) What is the public opinion mostly concerned about? (4) Who are the coordinating network structure and influential groups (or, ODGs) in the campaign network, or in other words which set of nodes are most powerful in disseminating the message? (5) Can we identify influential narratives in the cyber campaign?

2 Literature Review

Digital forensics tools have been mainly used by law enforcement agencies for detecting and solving corporate fraud [6]. Cyber forensics tools can be traced back to the early 1980's when these tools were mainly used by government agencies, e.g., the Royal Canadian Mounted Police (RCMP) and the U.S Internal Revenue Service (IRS). With time, these tools got more sophisticated and in the mid of 1980's these tools were able to recognize file types as well as retrieve lost or deleted files, e.g., *XtreeGold* and *DiskEdit* by Norton. In 1990's these tools became more popular with more capabilities, e.g., recovering deleted files and fragments of deleted files using *Expert Witness* and Encase [7]. Nowadays, many tools are available to public to collect cyber forensics data and visualize it, e.g., Maltego. Blogs provide rich medium for individuals to frame an agenda and develop a discourse that could possibly influence the masses. Twitter, however due to the 140-character limit is primarily used as a dissemination medium. Typically, bloggers use Twitter to build an audience and as a vehicle to carry their

message to their audience. It is important to understand the disinformation dissemination network on Twitter but it is equally, if not more, important to understand the blog environment and specifically the blogger's influence, engagement with the audience, and motivations for agenda setting. Identifying influential individuals in blogosphere is a well-studied problem. Many studies have been conducted to identify influence of a blogger in a community [8]. A blog post having more in-links and comments indicates that the community is interested in it.

3 Methodology

The overall methodology of this study is depicted in Fig. 1. We identified six groups by searching their names on various social media platforms to identify their Twitter and blogging profiles. NATO's public affairs officers then verified these profiles. These six groups propagate their messages on social media inviting people to act against NATO and TRJE 2015 exercise. An initial set of twelve blog sites were identified that the groups use to develop narratives against the TRJE 2015 exercise. We were also able to identify Twitter handles used to steer the audience from Twitter to their blogs. We identified an initial set of 9 Twitter accounts used by the six groups. We used Twitter API through a tool called *NodeXL* to collect a network of *replies, mentions, tweets, friends*, and *followers* for all the nine Twitter accounts and whoever is connected to them with any one of the aforementioned relationships for the period 8/3/2014 to 9/12/2015. The dataset file we obtained contains 10,805 friends/followers, 68 replies, 654 tweets, 1,365 mentions, 9,129 total nodes, and 10,824 total edges. The twitter handles, blogs, and names of the groups studied in this research are publically available. However, in order to ensure their privacy, we do not disclose them here.

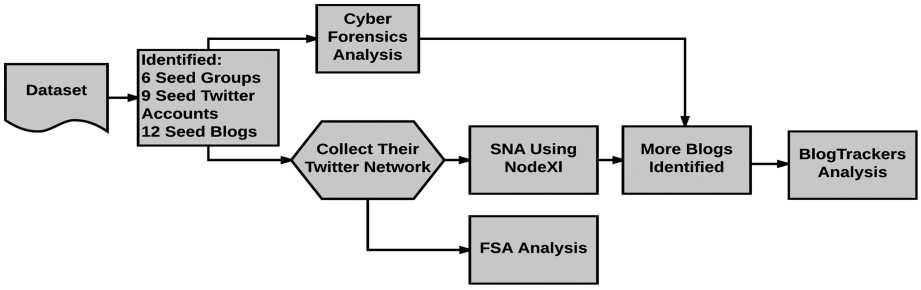


Fig. 1. Proposed methodology to analyze propaganda campaigns.

Metadata Extraction Using Maltego: Maltego is an open source information gathering and forensics application. Maltego can extract Google Analytics IDs from blog sites. Google Analytics is an online analytics service that allows a website owner to gather statistics about their website visitors such as their browser, operating system, and country among other metadata. Multiple sites can be managed under a single Google analytics account. The account has a unique identifying "UA" number, which is usually embedded in the website's html code [9]. Using this identifier other blog sites that are

managed under the same UA number can be identified. This method was reported in [9, 10]. So by using Maltego we can infer connections among blog sites and identify new sites that were previously undiscovered.

We used a seed set of 12 blog sites to discover other blogs that are connected to them using Maltego as explained earlier. We used Maltego in a snowball manner to discover other blog sites. We were able to identify additional 9 blogs that are connected to the initial seed blogs by the same Google analytics IDs. These newly identified websites have the same content published on different portals and sometimes in different languages. For example, a website written in English may also have another identical version but written in another language that is native to the region. Such blogs are also known as *bridge blogs* [11]. We went a step further to collect the IP addresses, website owner name, email address, phone numbers, and locations of all the websites. We obtained three clusters of websites based on their geolocation. These clusters are helpful to know the originality of the blog sites, which would help an analyst understand the propaganda that is being pushed by the specific blog site. Cluster 1 contains one website that is located in Russia, Cluster 2 has 8 websites located in USA, and Cluster 3 has 12 blog sites located in Spain, Cayman Islands, UK, and Germany. From initial 12 blog sites we grew to 21 blog sites, 6 locations, and 15 IP addresses. All the blog sites we identified during this study were crawled and their data was stored in a database that the Blogtrackers tool can access and analyze.

Applying SNA to Identify Influential Information Actors: After using Maltego to find other related blog sites used by the group to disseminate their propaganda, we applied SNA to find the most important nodes in the network by activity type. Using NodeXL we were able to find the most used *hashtags* during the time of the exercise. This helps in targeting the same audience if counter narratives were necessary to be pushed to the same audience. In addition to that, we found the most tweeted *URLs* in the graph. This gives an idea about the public opinion concerns. And finally we found the most used *domains*, which helps to know where the focus of analysis should be directed, or what other media platforms are used. For example, two of the top 10 hashtags that were used during the TRJE 2015 exercise were #YoConvoco (that translates to "I invite" using Google translation service) and #SinMordazas (that translates to "No Gags"). These two hashtags were referring to a campaign that is asking people for protests and civil resistance or civil disobedience. Also, investigating the top 10 URLs that were shared the most in the dataset reveals that these URLs were links to websites that are mobilizing people to raise objections on using taxpayers' money to fund military spending on wars.

Applying FSA to Identify Powerful Groups of Individuals Effecting Cyber Propaganda Campaign: We divided our network (9,129 nodes and 10,824 unique edges) into two types namely, *the social network*, derived from friends and follower's relations and *the communication network*, derived from replies and mentions relations. We ran the FSA algorithm on these two networks to discover the *most influential group of nodes*. Running FSA on the *social network* resulted in 1 focal structure with 7 nodes. These 7 nodes are in fact among the nine anti-NATO seed nodes we started with and

are very tightly knit (i.e., they exert mutually reciprocative relationships). This indicates a strong coordination structure among these 7 nodes, which is critical for conducting information campaigns. Running FSA on the *communication network* resulted in 3 focal structures with a total of 22 nodes. The same 7 accounts (out of the 9 seed accounts) found in the social network focal structures are distributed in these 3 focal structures. This gives those 7 accounts more power/influence than other nodes in the network because they are found in the focal structures of both networks, i.e., the communication and social network. The rest of the nodes (i.e., the additional 15 accounts) found in these 3 focal structures of the communication network are new nodes. These are important because they are either leaders or part of key groups conducting propaganda campaigns.

Using Blogtrackers to Analyze Blog Data: Using SCF analysis and SNA as explained in the previous sections, we were able to identify a total of 21 blog sites of interest. We trained web crawlers to collect data from these blogs and store the data in Blogtrackers database. We performed the following analysis: (1) we started exploring the collected dataset by generating the *traffic pattern* graph using Blogtrackers, for the period of August 2014 to December 2015. We observed a relatively higher activity in these blogs from September 2015 to December 2015, the period around the TRJE 2015, (2) then we generated a *keyword trends* graph for the keywords 'anti nato', 'trident juncture', 'nato'. The keyword trend for the 'anti nato' completely aligned with the traffic pattern graph indicating the posts actually had 'anti nato' keyword in it. We also observed that trend for 'anti nato' was consistently higher than 'nato' for this time period indicating there was more negative sentiment towards NATO in these blogs, (3) we ran the *sentiment* analysis in Blogtrackers for the same period and observed more negative sentiment than positive sentiment in the blogs, (4) we ran the *influential posts* analysis in Blogtrackers to identify posts with high influence. In other words, we want to identify what resonates with the community most, or which narratives are affecting the people most. The most influential post was an Italian blog post from the 'nobordersard' blog. Upon translation to English we found the post to be highly propaganda-riddled. The blogger used two of the conventional propaganda techniques [12] called "*Name Calling*" (associating a negative word to damage the reputation) and "*Plain Folks*" (presenting themselves as ordinary people or general public to gather support for their cause or ideology). The blog post used phrases like: NATO exercise was contributing to pollution and exploiting resources. It also categorizes this exercise as an act of militarization of territories to train for war. Furthermore, the blog was asking people to protest against the exercise.

4 Conclusion, Summary, and Future Directions

In this paper, we study the ODGs and their behavior in conducting deviant acts, especially disseminating propaganda against NATO and TRJE 2015. We further study how ODGs use social media in coordinating cyber propaganda campaigns. We conducted a *node-level* analysis, a *group-level* analysis, and *content* analysis. We collected Twitter network of the six deviant groups who had 9 twitter accounts and 12 blog sites. We analyzed this network to discover who are the top users in terms of activity, i.e., tweet, retweet, or mentions. We also discovered the most used hashtags, the most tweeted

URLs, and the most used domains. This served as node level analysis. Then we used SCF tool to discover other blog sites that are related to the seed blogs. This enabled us to discover how blogs are connected and if the same group owns multiple blogs. By applying FSA, we discovered the coordinating groups. This served as a *group level analysis*. We further analyzed the content of the blogs using Blogtrackers tool to discover the most prominent propaganda messages and the techniques these groups use to be effective in spreading their messages. This served as *content analysis*. The aforementioned methodologies constitute a tiny but promising sample from a spectrum of approaches to study cyber propaganda campaigns on social media.

Acknowledgements. This research is funded in part by NSF (IIS-1636933, IIS-1110868 and ACI-1429160), ONR (N000141010091, N000141410489, N0001415P1187, N000141612016, and N000141612412), AFRL, ARO (W911NF-16-1-0189), DARPA (W31P4Q-17-C-0059), and the Jerry L. Maulden/Entergy Fund at UA Little Rock.

References

1. Wright, B.: Social Media and the Changing Role of Investigators, Forensic Mag., December 2012
2. Mulazzani, M., Huber, M., Weippl, E.: Social network forensics: tapping the data pool of social networks. In: Eighth Annual IFIP WG, vol. 11 (2012)
3. Povar, Digambar, Bhadran, V.K.: Forensic data carving. In: Baggili, Ibrahim (ed.) ICDF2C 2010. LNICST, vol. 53, pp. 137–148. Springer, Heidelberg (2011). doi: 10.1007/978-3-642-19513-6_12
4. Al-khateeb, S., Agarwal, N.: Analyzing flash mobs in cybernetic space and the imminent security threats a collective action based theoretical perspective on emerging socotechnical behaviors. In: 2015 AAAI Spring Symposium Series (2015)
5. Sen, F., Wigand, R., Agarwal, N., Yuce, S., Kasprzyk, R.: Focal structures analysis: identifying influential sets of individuals in a social network. Soc. Netw. Anal. Min. **6**, 1–22 (2016)
6. Alherbawi, N., Shukur, Z., Sulaiman, R.: Systematic literature review on data carving in digital Forensic. In: Procedia Technology, vol. 11, pp. 86–92 (2013)
7. Oyeusi, K.: Computer Forensics. London Metropolitan University (2009)
8. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 207–218 (2008)
9. Alexander, L.: Open-Source Information Reveals Pro-Kremlin Web Campaign. Global Voices, 13 July 2015. https://globalvoices.org/2015/07/13/open-source-information-reveals-pro-kremlin-web-campaign/. Accessed 08 Oct 2015
10. Bazzell, M.: Open Source Intelligence Techniques: Resources for Searching and Analyzing Online Information, 4th edn. CCI Publishing (2014)
11. Etling, B., Kelly, J., Faris, R., Palfrey, J.: Mapping the Arabic blogosphere: politics, culture, and dissent. Berkman Cent. Res. Publ., vol. 6 (2009)
12. Standler, R.B.: Propaganda and How to Recognize it. RBS0, 02 September 2005

From Cyber Space Opinion Leaders and the Diffusion of Anti-vaccine Extremism to Physical Space Disease Outbreaks

Xiaoyi Yuan(✉) and Andrew Crooks(✉)

Computational Social Science Program, George Mason University, Fairfax, VA 22030, USA
{xyuan5,acrooks2}@gmu.edu

Abstract. Measles is one of the leading causes of death among young children. In many developed countries with high measles, mumps, and rubella (MMR) vaccine coverage, measles outbreaks still happen each year. Previous research has demonstrated that what underlies the paradox of high vaccination coverage and measles outbreaks is the ineffectiveness of "herd immunity", which has the false assumption that people are mixing randomly and there's equal distribution of vaccinated population. In reality, the unvaccinated population is often clustered instead of not equally distributed. Meanwhile, the Internet has been one of the dominant information sources to gain vaccination knowledge and thus has also been the locus of the "anti-vaccine movement". In this paper, we propose an agent-based model that explores sentiment diffusion and how this process creates anti-vaccination opinion clusters that leads to larger scale disease outbreaks. The model separates cyber space (where information diffuses) and physical space (where both information diffuses and diseases transmit). The results show that cyber space anti-vaccine opinion leaders have such an influence on anti-vaccine sentiments diffusion in the information network that even if the model starts with the majority of the population being pro-vaccine, the degree of disease outbreaks increases significantly.

Keywords: Agent-based modeling · Information networks · Infectious disease transmission

1 Introduction

Measles is a highly contagious disease and poses danger to communities around the world. Even though a safe and cost-effective vaccine has been available, we are still experiencing measles outbreaks every year in developed countries. Researchers and health practitioners have realized the potential danger of smaller clusters with high density of unvaccinated children (Lieu et al. 2015). In the US, the main cause of measles outbreaks is not the availability or affordability of vaccines but belief systems of individuals, which is seen with the advent of vaccine exemptions (May and Silverman 2003). The vaccine refusal rate is becoming higher in the last few years as non-medical exempt policies are being implemented (Wang et al. 2014). The fear with respect to the safety of the measles, mumps, and rubella (MMR) vaccine is the main argument behind

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 114–119, 2017.
DOI: 10.1007/978-3-319-60240-0_14

the "anti-vaccine movement." How anti-vaccine parents formed their opinions and how to prevent more parents from distrusting vaccination have become questions more important than ever. This paper explores the influence of anti-vaccine opinion leaders in cyber space and the consequences in disease outbreaks if the influence is carried over to physical space communities. An agent-based model (ABM) is developed to unveil mechanisms and causal relationships between opinion leaders, opinion clustering, and the degree of disease outbreaks.

2 Background

A large amount of research has been dedicated to exploring the driving force of anti-vaccine sentiment. With the advent of the Internet and social media platforms, parents search for information about MMR vaccine online (e.g. Kata 2012). Even a brief encounter of vaccine-critical websites could increase perceptions of risk of vaccinations and a decrease in the perception of vaccination benefits (e.g. Betsch et al. 2010). To better understand the multi-folded impact of online vaccine information, however, we should understand the "network effects" of vaccine information flow (Witteman and Zikmund-Fisher 2012). For instance, the fact that small groups of vocal "anti-vacciners" leveraged the power of social media to keep the ban on personal belief exempt bill of California from passing (DiResta and Lotan 2015). To date, little research has focused on the influence of online opinion leaders' gains on vaccination sentiments. Online opinion leaders are those who have disproportionally big advantage on disseminating their own voices and opinions. They are not necessarily celebrities in real life, but structurally, act as "hubs" in scale-free social media networks. It has been shown for example, that the retweet network of Twitter is a scale-free network (Tinati et al. 2012). This allows the influential users to be "opinion leaders" (Van Eck et al. 2011).

As the purpose of this paper is to connect opinion clustering to disease outbreaks, a prototype model must be mentioned, that of Salathé and Bonhoeffer (2008) who explored the effect of opinion clustering on disease outbreaks. The mechanism in their work was simple: 100 agents were created in a lattice network and each one was assigned a vaccination opinion (either support or oppose). After the opinion formation process, anti-vaccine agents were clustered together. Their model well demonstrated that a small clustering effect could increase the probability of a disease outbreak. The opinion formation process was probability-based, whereby an agent's choice for support/oppose was decided by two factors: the number of neighbors with opposite opinions (i.e. a dissimilarity index) and the parameter named "strength of opinion formation". Our model substitutes this probability based opinion formation process with one that simulates opinion diffusion in networks triggered by opinion leaders.

3 Methodology

The agent-based model was implemented in NetLogo 5.3.1. A brief model description is given below which loosely follows the 'ODD' (Overview, Design concepts, and Details) protocol (Grimm et al. 2010). A more comprehensive and in-depth description

of the model and source code can be found at https://www.openabm.org/model/5509/. Agent-based modeling allows exploring interactions among heterogeneous agents. By systematically changing initial conditions, the model simulates different scenarios (i.e. the different number of anti-vaccine opinion leaders and different level of anti-vaccine sentiment of the others). The main purpose of our model is to simulate how anti-vaccine extremism sentiment diffuse from anti-vaccine opinion leaders in a scale-free network and how this diffusion process creates anti-vaccine opinion clusters that gives rise to disease outbreaks.

3.1 Agents and Environment

The agents are heterogeneous individuals with different levels of anti-vaccine sentiment level, physical locations, connection status in the cyber network and other attributes in Boolean logic format – extremist or non-extremist, vaccinated or not vaccinated, susceptible or not susceptible, infected or not infected, recovered or not recovered. The environment of the model is two folded: physical space and cyber space. Agents are connected differently in either space. The physical space is represented by lattice. Each grid cell on the lattice is one physical location, which does not represent any real world geographical area but only an abstraction. This model does not consider different population densities and all the grids are occupied. The world is wrapped both horizontally and vertically. The neighborhood is a Moore neighborhood in that each individual has eight neighbors who are located in the most adjacent eight grids.

3.2 Process Overview and Model Scheduling

There are five steps processed one by one for each run (see the details in Sect. 3.1 in the ODD): Step (1) creates the scale-free cyber network, assign each one an anti-vaccine sentiment value (the distribution is a parameter), and identify a certain number (parameter) of the most connected individuals to be anti-vaccine opinion leaders. In step (2), those who are directly connected with opinion leaders in the cyber network and have an anti-vaccine sentiment value higher than a threshold (parameter) become extremists which allows extremism to spread. In our model, "Directly connected" means one-degree connection in the network. Followers of followers, for instance, do not get influenced by opinion leaders in the cyber network. This process is only executed once in the model for each run. In step (3), we spread extremism in physical space after the spread on cyber space is established, those who are extremists influence their local neighbors with the same diffusion mechanism as in step 2. Following this in step (4) we capture vaccination rates in the sense that extremists are those who are not willing to get vaccinated while the rest of the population gets vaccinated and is immune from the disease. Those who are not vaccinated are "susceptibles". Finally, in step (5) we model disease transmission for which we use a SIR (Susceptible-Infectious-Recovered) disease transmission model. Vaccinated individuals and those who are recovered are all treated as being immune from the disease. For each time step, not everyone who are susceptible and have infected neighbors will be infected as infection only happens under certain probability/rates. There is also a recovery rate for infected agents (this is discussed

further in Sect. 3.1 of the ODD). The model does not consider death and every infected person recovers. The model stops when nobody can be infected any more. "Time" is not counted until the model enters disease transmission part. Each time step, the model records of the number of infected agents, susceptible agents, and recovered agents to allow for further analysis.

3.3 Initialization

In the initial state of the model, contains 2601 (51 * 51) grid cells and individuals (i.e. agent) are created on each grid cell. Default values of the individuals include: the attribute "anti-vaccine-sentiment", which follows a normal distribution with a mean of 0 (overall the sentiment is neutral) and a standard deviation of 1 with an upper bound 1 and a lower bound –1. If it's positive, it means that this agent is opposed to vaccination and negative means supportive. All the agents are initialized as non-extremists and all their attributes related to disease transmission, "vaccinated?", "susceptible?", "infected?", "recovered?" are set as false.

In addition, the infection rates and recovery rates functions are also set up as the model initiated. For each agent, the infection rates are calculated based on the exponential function: $f(i) = 1 - \exp(-ßi)$. The function is from the above-mentioned prototype model by Salathé and Bonheoffer (2008). In all the simulations, $ß = 0.05$ and i represents the number of neighbors that are infected. Because the exponential distribution is fat-tailed, the infected probability will increase faster as the number of infected neighbors increases. The recovery rate is 0.1% for all agents. The start of the disease transmission is always from 2 randomly picked agents. Three parameters need to be specified to initialize the model: number of individuals connected in the network, number of opinion leaders, and the sentiment threshold.

4 Results

Before presenting the results, it needs to be noted that sensitivity testing of the model parameters and verification was performed to ensure the model was functioning as expected. To control the impact of other factors, in each experiment, the cyber network remains the same. Simulation experiments were carried out to compare the result of "experimental group" – experiments with the process of sentiment diffusion; and the "control group" – experiments without the diffusion process. To make it comparable, after the spread of sentiment in cyber network, there's a number of anti-vaccine extremists and the control group is constructed that use the number of anti-vaccine extremists from experimental group but picked randomly. For both experimental and control group, we randomly pick 2 agents as infected and spread the disease in the physical space. For each run, output the maximum number of people who are infected for each group. We experiment with two parameters: number of opinion leaders = [1, 2, 3, 4, 5, 10, 15, 20, 30] and sentiment threshold = [0.95, 0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, 0.5]. For each combination, 100 simulations were carried out for experimental and control group.

Our main finding, as shown in Fig. 1, is that even when the sentiment threshold is high, the clustering effect of disease transmission is still strong. Having a high sentiment threshold means that only a small proportion of the population is potentially extremists. For instance, when the parameter is 0.95, it indicates that only 5% of the model population has the possibly to be turned into anti-vaccine extremists. Even in this strict condition, the increase of the maximum number of infected is higher than 33%. This shows that when the majority population is pro-vaccine, there's still serious potential danger from small unvaccinated clusters. Additionally, what's surprising is that the increase in the maximum number of infected people is the largest when the number of anti-vaccine opinion leader is equal to one. When there's only one opinion leader in the network, the unvaccinated clusters are denser than those created under the influence of more than one opinion leaders. As the number of opinion leader increases, the model ends up having unvaccinated clusters that are more dispersed, which is structurally more like the randomly picked unvaccinated population (the correspond control group) and therefore, there's lower increase of maximum infected population.

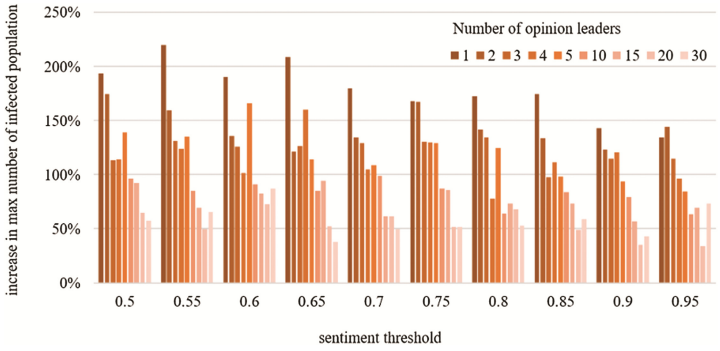


Fig. 1. Results of simulation experiments

5 Discussion

The model explores the two-folded influence of online anti-vaccine opinion leaders, i.e. the potential influence on cyber level opinion change and those being influenced by opinion leaders could become influencers themselves in their own physical community. Our major finding is that while the majority of the population is pro-vaccine, with very few online opinion leaders triggering the spread of sentiment of anti-vaccination, unvaccinated clusters in physical space can lead to a drastic increase in infection comparing to the scenario that unvaccinated population is not clustered. This model, however, is a simple model with assumptions that simplify real world scenarios but it is a foundation for our ongoing work that analyzes Twitter anti-vaccine textual data and the user retweet network. As for the model itself, one of the limits is that it only considers social influence. The phenomenon of opinion clustering can be both the result of social influence and homophily, which is often treated as being coupled together to create clusters in network theory (Shalizi and Thomas 2011). With this being said, this paper explores the

possibility of analyzing anti-vaccination opinion formation by taking social network attributes and social influence into consideration. Considering the growing trend in online health knowledge seeking and the increasing influence of anti-vaccine movement on various social network websites (e.g. Kata 2012), our model lays the foundation study the nexus of cyber and physical relationships and provides a heuristic tool to study how anti-vaccine opinion leaders potentially increase the severity of measles or other preventative disease outbreaks.

References

Betsch, C., Renkewitz, F., Betsch, T., Ulshöfer, C.: The influence of vaccine-critical websites on perceiving vaccination risks. J Health Psychol. **15**, 446–455 (2010)

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., Stefanidis, A.: Linking cyber and physical spaces through community detection and clustering in social media feeds. Comput. Environ. Urban Syst. **53**, 47–64 (2015). doi:10.1016/j.compenvurbsys.2014.11.002

DiResta, R., Lotan, G.: Anti-Vaxxers are using Twitter to manipulate a vaccine bill. https://www.wired.com/2015/06/antivaxxers-influencing-legislation/

Fox, J.P.: Herd immunity and measles. Clin. Infect. Dis. **5**, 463–466 (1983)

Grimm, V., Berger, U., DeAngelis, D.L., Polhill, J.G., Giske, J., Railsback, S.F.: The ODD protocol: a review and first update. Ecol. Model. **221**, 2760–2768 (2010)

Kata, A.: Anti-vaccine activists, Web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine **30**, 3778–3789 (2012)

Lieu, T.A., Ray, G.T., Klein, N.P., Chung, C., Kulldorff, M.: Geographic clusters in underimmunization and vaccine refusal. Pediatrics **135**, 280–289 (2015)

May, T., Silverman, R.D.: "Clustering of exemptions" as a collective action threat to herd immunity. Vaccine **21**, 1048–1051 (2003)

Salathé, M., Bonhoeffer, S.: The effect of opinion clustering on disease outbreaks. J. R. Soc. Interface **5**, 1505–1508 (2008)

Shalizi, C.R., Thomas, A.C.: Homophily and contagion are generically confounded in observational social network studies. Sociol. Methods Res. **40**, 211–239 (2011)

Tinati, R., Carr, L., Hall, W., Bentwood, J.: Scale free: Twitter's retweet network structure (2012)

Van Eck, P.S., Jager, W., Leeflang, P.S.H.: Opinion leaders' role in innovation diffusion: a simulation study. J. Prod. Innov. Manag. **28**, 187–203 (2011)

Wang, E., Clymer, J., Davis-Hayes, C., Buttenheim, A.: Nonmedical exemptions from school immunization requirements: a systematic review. Am. J. Public Health **104**, e62–e84 (2014)

Witteman, H.O., Zikmund-Fisher, B.J.: The defining characteristics of Web 2.0 and their potential influence in the online vaccination debate. Vaccine **30**, 3734–3740 (2012)

Event-Based Model Simulating the Change in DDoS Attack Trends After P/DIME Events

Adam Tse(✉) and Kathleen M. Carley

Institute for Software Research,
Carnegie Mellon University, Pittsburgh, PA 15213, USA
atse1@andrew.cmu.edu, kathleen.carley@cs.cmu.edu

Abstract. This paper describes the methods for creating an event-based simulation used to predict DDoS attacks against countries following international events. The model uses various parameters for an event and generates time series DDoS attack data for the two countries over one week. The simulation uses a weighted, tit-for-tat approach in determining retaliation. The model was evaluated using attack data of actual events provided by Arbor Networks consisting of a two-week interval plus a day, centered around the start of the events. The model was sufficient in predicting the change in frequency of DDoS attacks following hostile diplomatic events, but it was unsuccessful at simulating attacks following friendly, military, and economic events. Overall, the resulting simulation was a successful baseline for future work in the field.

Keywords: DDoS · Simulation · Modeling · Cyberwarfare · Cyber-policy · Cyber-attacks · International relations

1 Introduction

Cyberwarfare has become a difficult issue in international relations. Nations are suspected of facilitating cyber-attacks to steal intellectual property and attack urban infrastructure to improve their own economic status and there has been little success in holding countries responsible for cyber-attacks [1,9]. One prominent tactic of cyberwarfare is distributed denial of service (DDoS) attacks against infrastructure and state resources. DDoS attacks are a type of attack where a victim machine is made unavailable due to flooding of bandwidth or resources from a network of compromised machines. Few examples of state-sponsored attacks include a Russian attack on major Estonian webservers in 2007 and a series of Iranian DDoS attacks against banks in the United States (U.S.) in 2013 [2].

Previous research indicated that cyber-attacks are associated with social, political, and cultural conflicts [3]. Thus, a simulation was created using P/DIME (political/diplomatic, informational, military, and economic), a reputed methodology defined by the National Defense University for managing operations to attain the effect required to complete an objective, to classify international events and predict their effect on DDoS attack trends [8].

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 120–126, 2017.
DOI: 10.1007/978-3-319-60240-0_15

The paper is organized into the following sections: the relevant works, the methods of the simulation, the virtual experiment using five real events, the results and discussion, and limitations of the model, and a conclusion along with possible directions to expand the research.

2 Relevant Work

Though there has been a large amount of research on cyber-attacks, DDoS attacks, and attribution, most of those works have focused on the technical defenses and processes of DDoS attacks [5,6]. This paper is novel because it focuses on a "sociology of nations" perspective of DDoS attacks rather than a technical perspective. The most similar research includes a focus on game theory between institutions and other analysis on motivations of cyber-attacks [1,3,7,8].

3 Method

Users have the ability to control the P/DIME Category, whether an event was hostile or friendly (1 or 0), the severity of an event (1 to 3), the source and target country of the event, and a random noise coefficient represented as number between 0 and 10. Events tested in the simulation are defined by the previously mentioned properties minus the random noise coefficient. The schema was represented as E(Type, Hostility, Severity, Source, Target). The dependent variables for the simulation include the total number of DDoS attacks and the attacks per hour targeting both the source country of the event and the target country of the event.

The static variables were the ally matrix (matrix specifying ally countries; 0 or 1), the enemy matrix (matrix specifying enemy countries; 0 or 1), countries (U.S., China, and Russia), and country attributes (aggression, GDP [4], percentage of internet users [10], and average attacks per hour).

A high level diagram of each simulation iteration is shown on Fig. 1 and a more detailed description of the algorithm is provided in additional works in the Springer Journal.

4 Virtual Experiment

In order to test the simulation, the results of the simulation were compared with real DDoS attack data taken around the timeframe of the events. The data was provided by Arbor Networks from the Digital Attack Map website at http://www.digitalattackmap.com/.

The real DDoS attack data was compared with the simulated data for similarities using the Google CausalImpact R library. In each comparison, the pre period started a week before the event and the post period started 24 h after the event. If the scenario models exhibited behavior closer to the real data than the None scenario, then it was concluded that adding event-driven cyber-attacks improved the realism of the model. The events chosen included the following:

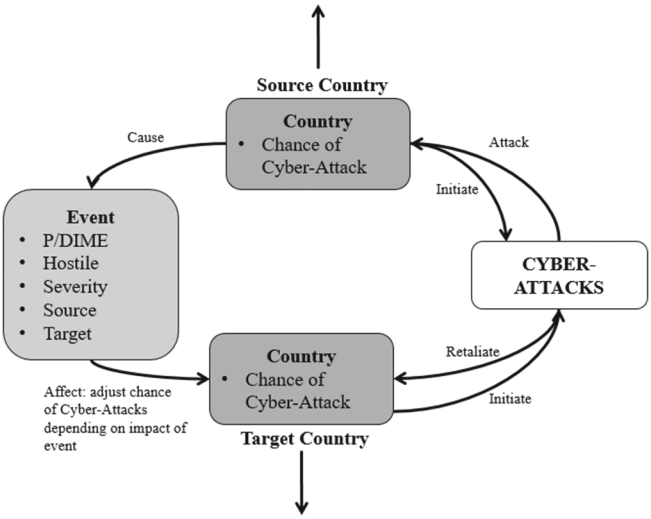


Fig. 1. System diagram of simulation algorithm

- China ignores court decision South China Sea 07/12/2016; E(D, 1, 2, China, U.S.)
- U.S. China Cybersecurity talk 06/14/2016; E(D, 0, 1, U.S., China)
- U.S. claims Russia hacked Democratic National Committee (DNC) and Hilary Clinton 10/14/2016; E(D, 1, 3, U.S., Russia)
- Vladimir Putin exits nuclear security pact 10/03/2016; E(M, 1, 3, U.S., Russia)
- Chen Feng's HNA Group bus $6.5 billion stake in Hilton 10/28/2016; E(E, 0, 2, China, U.S.)

These events were chosen because of their variance in P/DIME categories, hostility, and severity. Only events between U.S., China, and Russia were chosen because of the high number of cyber-attacks happening between the three countries. Additionally, in current events, the U.S., China, and Russia are currently the three strongest cyber powers in the world.

5 Results and Discussion

5.1 Overview

The complete results are presented in Table 1. A detailed analysis of each event and possible explanation of the results is below. Time series plots showing the impact of events over the week are in additional works in the Springer Journal. Overall, the event cases exhibited better performance than the no event case for all five events.

Table 1. Comparison of results of simulation with real world data

Event	Schema	Arbor	Simulation
South China Sea Dispute 07/12/2016	E(D, 1, 2, CN, US)	CN 6.3%, [−37%, 50%] P =.371	CN 1.8%, [0.76%, 2.7%] P =.001
		US 66%, [32%, 97%] P = .001	US 0.55%, [−0.28%, 1.3%] P = .084
China Cybersecurity Talk 06/14/2016	E(D, −1, 1, US, CN)	US 50%, [−0.51%, 103%] P = .027	US −4.4%, [−5.3%, −3.5%] P = .001
		CN −32%, [−77%, 14%] P = 0.082	CN 0.13%, [−0.82%, 1.1%] P = .404
Hacked DNC 10/14/2016	E(D, 1, 3, US, RU)	US 50%, [9.8%, 90%] P = .007	US 3.3%, [2.5%, 4.1%] P = .001
		RU 150%, [6.7%, 293%] P = .018	RU 0.18%, [−0.77%, 1.1%] P = .365
Nuclear Security Pact 10/03/2016	E(M, 1, 3, US, RU)	US −32%, [−74%, 7.7%] P = .059	US 61%, [60%, 62%] P = .001
		RU −57%, [−144%, 29%] P = .121	RU 148%, [147%, 149%] P = .001
HNA Group Stake Hilton 10/28/2016	E(E, −1, 2, CN, US)	CN −70%, [−166%, 23%] P = .067	CN 77%, [77%, 78%] P = .001
		US −20%, [−56%, 12%] P = .127	US 145%, [144%, 146%] P = .001

The simulated effect of the South China Sea Dispute Court Decision and Hacking of the DNC were very accurate. Both countries suffered an increase in DDoS attacks. Additionally, the effect of the Hacking of the DNC had a higher degree than the effect of the South China Sea Dispute which was consistent with the higher severity score of the Hacking of the DNC. The only inconsistency was the degree of increase in DDoS attacks. The inconsistencies with the South China Sea Dispute Court may be attributed to interference from a previous event where the U.S. had sent the U.S. Navy to hinder Chinese claims on the sea.

Though DDoS attacks against China after the U.S./China Cybersecurity Talk dropped, the assumption that friendly events cause a decrease in all DDoS attack trends was false as indicated by the increase in DDoS attacks against the U.S. This can be explained by the hidden agendas and intents behind events. Though China was stating publicly that they would try to stop cyber-attacks, sanctioned attacks could have still been going on as a result of the event.

In regards to the Failed Nuclear Security Pact and the Group Stake on Hilton, the simulation predicted a significant increase in DDoS attacks, however the Arbor data showed a significant, large decrease in DDoS attacks. The decline in DDoS attacks after the Failed Nuclear Security Pact can be a result of limitations

in data or it may indicate that both parties wanted to maintain relations after the argument. The decrease in DDoS attacks after the HNA Group Stake might be because economic events between the private industry may not have an effect on DDoS attacks and other events may be influencing the decrease. Additional study on these types of events may be needed to see if the correlations are consistent.

Overall, the results were very good considering the little amount of study done in the field. Diplomatic events were predicted fairly accurately and only little adjustments will be needed to improve the model. However, friendly, military, and economic events exhibited the opposite effect of what was predicted. More analysis of these events may be needed to create a more accurate model.

5.2 Limitations

The first issue with the methods mentioned above was the Causal Impact analysis. At the start of the experiment, it was assumed that an international event would have an effect lasting one week. Google's Causal Impact Library takes a pre-period and a post-period as parameters where the change between pre-period and post-period determines whether a significant event happened between the two periods. Because of the low significance in effect for some of the real data, the experiment indicated that events may have shorter effect times. As a result, further studies must be made on different post-periods to determine how long would an event impact the frequency of DDoS attacks.

Additionally, the virtual experiment was limited in depth. The experiment only consisted of one event of each type category which was not a significant enough sample given the many input parameters for the simulation. As a result, it is difficult to prove if these results can be generalized for all events. This can be seen by the outcome of friendly, military, and economic events. The results were different from the model. However, it was not clear if this was a result of an incorrect assumption in the model or an anomaly event. On the contrary, diplomatic events seemed accurate because it had a consistent effect between the two events. Future study must be made specifically on each event type and input variable to see if the results are accurate.

However, even without the validation and virtual experiment issues, the model has quite a few limitations because of the simple assumptions it makes. First, the simulation assumes events are isolated. In the real world, events take place concurrently affecting each other and it is difficult to determine causation because of the sheer amount of noise. Additionally, the simulation does not take into account groups within nations. It assumes events affect whole nations when they may only affect groups within the nation such as civilians, companies, non-profits, or governments. Lastly, it assumes events have a visible effect on DDoS attack trends one day after the event. This is probably dependent on the event itself and the experiment may need to be repeated where effects begin in different periods after the event occurred.

6 Conclusion

Overall, the simulation is on the right track to predicting the impact of DDoS attacks. Though it may need tuning on the degree of increase, it correctly predicted the increase of DDoS attacks for both hostile diplomatic events. It also predicted that there would be an increase in DDoS attacks against one country and a decrease in DDoS attacks for another country on diplomatic, friendly events. Some features of the simulation that require tuning are its estimation of friendly, military, and economic events, the time it takes for the effect to occur and the duration of the effect, and the degree of effect for hostile diplomatic events.

For future work, an in-depth analysis should be done on each parameter of the simulation to determine the exact correlation of the parameter if one is found. The most logical one to examine is hostile diplomatic events which seemed to have fairly accurate results. Afterwards, events with unknown effects such as Military and Economic events should be analyzed to determine how they can be modelled. Overall, this study has proved that modelling and simulating the impact of international events on DDoS attacks is feasible.

If the model was tuned accurately, the research should be pushed to being able to predict cyber-incidents rather than just DDoS attacks. Other possible directions would be to be able to simulate a sequence of events on cyber-attack trends and predict the effect of events on multiple countries over time. These additions would make it possible to recreate cyber-attacks worldwide and judge the vulnerability of countries and help attribute state-sponsored attacks.

References

1. Chaturvedi, A.R., Gupta, M., Mehta, S.R., Yue, W.T.: Agent-based simulation approach to information warfare in the seas environment. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, 10-pp. IEEE (2000)
2. Clarke, R.A., Knake, R.K.: Cyber War. HarperCollins, New York (2011)
3. Gandhi, R., Sharma, A., Mahoney, W., Sousan, W., Zhu, Q., Laplante, P.: Dimensions of cyber-attacks: cultural, social, economic, and political. IEEE Technol. Soc. Mag. **30**(1), 28–38 (2011)
4. World Bank Group. (n.d.). gdp (current US$) (2016). http://data.worldbank.org/indicator/NY.GDP.MKTP.CD
5. Kotenko, I., Alexeev, A., Man'kov, E.: Formal framework for modeling and simulation of DDOS attacks based on teamwork of hackers-agents. In: IEEE/WIC International Conference on Intelligent Agent Technology (IAT 2003), pp. 507–510. IEEE (2003)
6. Kotenko, I., Ulanov, A.: Simulation of internet DDoS attacks and defense. In: Katsikas, S.K., López, J., Backes, M., Gritzalis, S., Preneel, B. (eds.) ISC 2006. LNCS, vol. 4176, pp. 327–342. Springer, Heidelberg (2006). doi:10.1007/11836810_24
7. Kumar, S., Benigni, M., Carley, K.M.: The impact of US cyber policies on cyber-attacks trend. In: 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pp. 181–186. IEEE (2016)

8. Starr, S.H.: Toward a preliminary theory of cyberpower. Cyberpower and national security, pp. 43–88 (2009)
9. Tereshchenko, N.: US foreign policy challenges: cyber terrorism and critical infrastructure, e. International Relations 12 (2013)
10. International Telecommunication Union: Individuals using the internet 2005 to 2014, key ICT indicators for developed and developing countries and the world (totals and percentage rates) (2015). http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2014/ITU_Key_2005-2014_ICT_data.xls

Using a Real-Time Cybersecurity Exercise Case Study to Understand Temporal Characteristics of Cyberattacks

Aunshul Rege[1(✉)], Zoran Obradovic[2], Nima Asadi[2], Edward Parker[1], Nicholas Masceri[1], Brian Singer[1], and Rohan Pandit[1]

[1] Department of Criminal Justice, Temple University, Philadelphia, USA
rege@temple.edu
[2] Computer and Information Sciences Department, Temple University, Philadelphia, USA

Abstract. Anticipatory cyber defense requires understanding of how cyber adversaries make decisions and adapt as cyberattacks unfold. This paper uses a dataset of qualitative observations conducted at a force on force ("paintball") exercise held at the 2015 North American International Cyber Summit (NAICS). By creating time series representations of the observed data, a broad range of data mining tools can be utilized to discover valuable verifiable knowledge about adversarial behavior. Two types of such analysis discussed in this work include clustering, which aims to find out what stages show similar temporal patterns, and peak detection for adaptation analysis. Collectively, this mixed methods approach contributes to understanding how adversaries progress through cyberattacks and adapt to any disruptions they encounter.

Keywords: Adaptive human behaviour · Dynamic decision making · Temporal analysis · Time series data · Clustering · Field research

1 Introduction

Today's information networks and integrated systems are highly networked, thereby increasing the attack surface, resulting in greater cyberattacks [2]. Yet, conventional cyberattack management is reactionary and does not capture Advanced Persistent Threats (APTs), which increasingly target critical infrastructures and consistently circumvent traditional security measures, resulting in large and costly damages [1]. It is therefore essential that commercial and government organizations develop defenses which are able to respond rapidly to, or even foresee, new attack strategies and tactics [2]. While many important contributions in anticipatory/proactive cybersecurity have been made, they are technical in nature and downplay the relevance of the human agents behind the cyberattacks, and their decision-making processes and adaptation strategies [2].

This paper employs quantitative data science methods of time series analysis to assess the observed adversarial behavior at a force on force ("paintball") exercise. Collectively, this mixed method contributes to understanding

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 127–132, 2017.
DOI: 10.1007/978-3-319-60240-0_16

how adversaries progress through cyberattacks and adapt to any disruptions they encounter. This paper is structured as follows. Section 2 outlines the mixed methodology of observations and time series analysis. Next, the computational results are discussed. Finally, this paper discusses relevant findings and possible implications for adversarial movement and adaptability.

2 Methodology

In the Criminological discipline, crime scripts provide a systematic understanding of the crime commission processes [3]. The applications of crime scripts to cyberattacks as they unfold remains understudied. In the technical domain, crime scripts appear as intrusion chain models that capture the step-by-step process of cyberattacks. While there are many models of adversarial intrusion chains, we use the 12-step cyber intrusion chain model in [1], as it offers detailed attack stages that allow for thorough data analysis.

The Merit Network and the Michigan Cyber Range provide a virtual platform called Alphaville, which is used for cybersecurity training exercises. Alphaville emulates a typical city and consists of five locations: a school, a library, a city hall, a small business, and a power company, each of which contains servers and firewalls with intentional vulnerabilities. During the 2015 North American International Cyber Summit (NAICS), the researchers observed a five-hour force on force "paintball" exercise, where teams battled to claim Alphaville's network by controlling critical servers. Researchers observed one of the teams participating, which consisted of four members (henceforward referred to as Subjects S1, S2, S3, and S4).

Temporal analysis aims to extract and characterize the trends, patterns, and variations within a process over time using time series data. In order to create the time series, the timestamped observations of the team's actions and their durations were utilized. In this work, each time point in the generated time series represents a one minute time span. For each time point, the value of each time series is the accumulated number of minutes spent by the entire team on its corresponding intrusion stage. After creating the time series representation of the data, we performed temporal analyses of the intrusion process through data mining methods, namely, clustering and peak detection. We performed clustering of the time series in order to achieve a verifiable measurement of co-activation and co-dependence of intrusion stages. Clustering allows similar time series (measured by comparing the amplitude of the time series, which is the total amount of time in minutes allocated to each intrusion stage during each minute of the exercise practice) to be placed in groups. A high similarity between the time series of the intrusion stages A and B is an indication that whenever intrusion stage A was performed within a time point, the possibility of performing stage B during that time point was higher than any other intrusion stage.

In this work, we use Agglomerative hierarchical clustering [4]. The reason behind choosing this clustering model is its power in providing the order and

similarity hierarchy of the clusters, and the fact that no a priori information about the number of clusters to be made is required.

To understand the team's adaptation measures when facing disruptions, the time series were then analyzed for detecting local peaks after these disruptions occurred. We employed a Peak-Valley detection algorithm [5] to detect the 'adaptation stages' by finding peak values that were above the global mean (average of the mean of all time series amplitudes), which were separated from those stages that the red team spent minimal time on (peak values below global mean).

3 Results

3.1 Observed Duration of Adversarial Intrusion Chain Stages

The observed data summarized at Fig. 1 suggest that the team spent approximately 49% (140 min) of the exercise time on entering the system, establishing foothold, and moving laterally to gain further control over systems. This was followed closely by Reconnaissance (stages 2, 3, 4, and 5), which took up roughly 44% of the exercise time or 125 min. The researchers did not find other intrusion stages during observations and so these stages are excluded from further analysis.

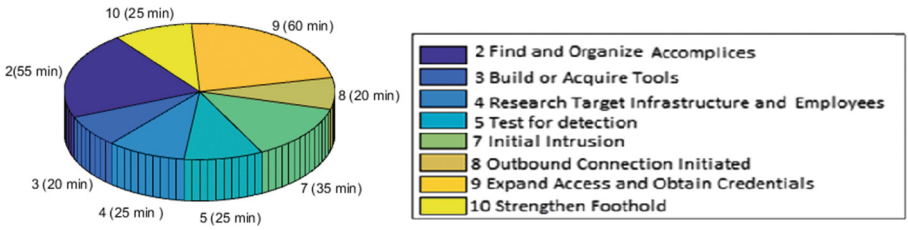


Fig. 1. Total time spent by the red team on each intrusion stage through the entire exercise

3.2 Time Series Generation and Clustering

Figure 2 shows the time series created for each intrusion stage. The clustering results are provided in Fig. 4, and an example of temporal pattern similarities is provided in Fig. 3. In Fig. 4, the vertical axis corresponds to the Euclidean distance of time series pairs. The clustering threshold, which determines the stages that are grouped together, was selected at the middle of the largest distance, which results in the red threshold line in Fig. 4. The results indicate the temporal similarities among intrusion stages; for instance, the occurrence of intrusion stage 3 (a peak in its time series), is more likely to be accompanied by the occurrence of the stages 4, 5, and 7 than any other intrusion stages.

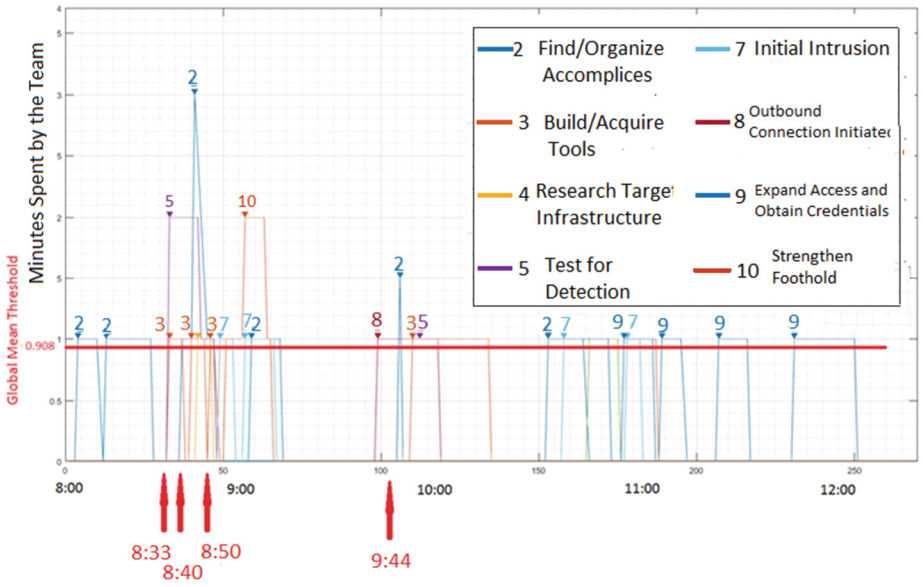


Fig. 2. Time series representation of the observational data. The arrows at the bottom show the disruptions corresponding to Table 1. (Color figure online)

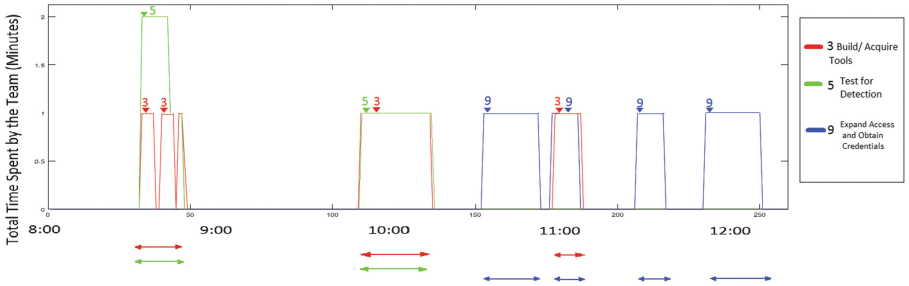


Fig. 3. An example of the similarities among time series; the peak/valley patterns happen more concurrently between intrusion stage pairs 3 and 5 compared to pairs 3 and 9, or pairs 5 and 9.

3.3 Analysis of the Adaptation Process

The local peaks of the time series and the global mean depicted by the horizontal red threshold line (global mean of 0.908 min total engagement per one minute interval) can be observed in Fig. 2. We can observe that within the 10 min time frame after a disruptive event, the amount of time allocated to certain intrusion stages was above this threshold at multiple intervals, indicating that the red team focused more on these stages in response to that disruption. For instance, in Fig. 2, we observe a spike in stage 2 (Find/Organize Accomplices) after the 8:40 access failure disruption (detailed in Table 1). Possible explanations for

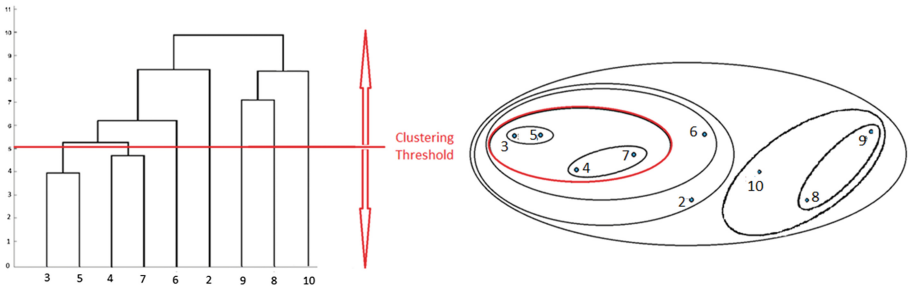


Fig. 4. Hierarchical clustering of the time series where each number corresponds to the intrusion stage number (Color figure online)

Table 1. Possible explanation for time spent on certain stages post disruptions

Time	Player	Hurdle	Disruption details (S)	Spiked stage (Mins Spent)	Stage sequence	Possible explanations for spike in the stage
8:33	S2	L	S3 Kills S2 attack chain	5(2), 3(1)	Con current	To test the targeted system's intrusion detection measure (spike in stage 5), the team was deciding which tools to use (spike in stage 3)
8:40	S2	L	S3: why do I keep losing my shell?	2(3), 3(1)	2, 3	Team member lost access, so may have sought help from other members (spike in stage 2) about which tools to use (spike in stage 3)
8:50	S3	S	S3 has a failed login attempt	7(1), 3(1)	7, 3	Team was possibly in stage 7 (spike in stage 7), moving laterally to strengthen foothold, but to gain access, may have tried different tools (spike in stage 3)
9:44	S2	L	S2 tries to get into the system	2(1.5), 3(1), 5(1)	2, (3, 5 concurrent)	Team member may be unsuccessfully trying to gauge target's defense measures (spike in stage 5) and hence may have sought help from other team members (spike in stage 2) about which tools to use (spike in stage 3)

disruptions and responses are provided in Tables 1, but these cannot be conclusive as they are based solely on observations, and as such, cannot account for the team's decision-making processes and dynamics.

4 Conclusion

There are some unavoidable limitations to this research such as generalizability and the fact that the case study is not representative of real cyberattacks.

However, the authors make the case that this paper is exploratory, methodologically unique, and based on one of the most reputable force on force ("paintball") exercises in the United States.

The time series analysis offers some interesting findings about the adversarial intrusion chains:

"Dispersed Spikes May Indicate Nonsequential Progression of Intrusion Stages". The greatest cumulative spike occurred for stage 2, but these spikes occurred at different times (Table 1, 8.40 and 9.44). This suggests that adversaries exhibit complex back and forth movement when they face disruptions.

"Parallel Stages and Stage 3 (Build/Acquire Tools)". After each disruption the team focused on multiple stages at either the same time (concurrent) or with a slight temporal lead (Table 1), suggesting that stages occur in parallel rather than in sequence. Also after each disruption, Stage 3 always occurred in parallel with other stages (Table 1), which suggests that building/acquiring tools may be a relevant stage during most adaptations.

Accessing Systems is Key across Multiple Stages. Most disruptions (Table 1: 8.40, 8.50, and 9.44) were related to difficulties in gaining or maintaining access to target systems, which was an issue at multiple stages.

Acknowledgements. This material is supported by the National Science Foundation (NSF) CAREER Award No. 1446574 and partially by NSF CPS Award No. 1453040. The authors thank the Merit Network and the Michigan Cyber Range for allowing data collection at their 2015 NAICS event.

References

1. Cloppert, M.: Attacking the cyber kill chain. http://digital-forensics.sans.org/blog/2009/10/14/security-intelligence-attacking-the-kill-chain. Accessed 2 Feb 2014
2. 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
3. Leclerc, B.: Crime scripts. In: Wortley, R., Townsley, M. (eds.) Environmental Criminology and Crime Analysis. Routledge (2016)
4. Rokach, L., Maimon, O.: Clustering methods. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 321–352. Springer, New York (2005)
5. Schneider, R.: Survey of Peaks/Valleys identification in Time Series. University of Zurich, Department of Informatics, Switzerland (2011)

Hybrid Modeling of Cyber Adversary Behavior

Amy Sliva(✉), Sean Guarino, Peter Weyhrauch, Peter Galvin, Daniel Mitchell, Joseph Campolongo, and Jason Taylor

Charles River Analytics, Cambridge, MA, USA
{asliva,sguarino,pweyhrauch,pgalvin,dmitchell,
jcampolongo,jtaylor}@cra.com

Abstract. Cyber adversaries continue to become more proficient and sophisticated, increasing the vulnerability of the network systems that pervade all aspects of our lives. While there are many approaches to modeling network behavior and identifying anomalous and potentially malicious traffic, most of these approaches detect attacks once they have already occurred, enabling reaction only after the damage has been done. In traditional security studies, mitigating attacks has been a focus of many research and planning efforts, leading to a rich field of adversarial modeling to represent and predict what an adversary might do. In this paper, we present an analogous approach to modeling cyber adversaries to gain a deeper understanding of the behavioral dynamics underlying cyber attacks and enable predictive analytics and proactive defensive planning. We present a hybrid modeling approach that combines aspects of cognitive modeling, decision-theory, and reactive planning to capture different facets of adversary decision making and behavior.

Keywords: Cyber defense · Adversary modeling · Cognitive models · Decision theory · Predictive analytics · Cyber simulation

1 Introduction

Over the last decade, the rapid increase in the number of networked devices, from desktop workstations to large-scale servers to ad hoc mobile devices, has vastly expanded the possible cyber attack surface. As modern cyber adversaries become more proficient and sophisticated, these network systems are increasingly vulnerable to cyber attacks. Despite significant investment in addressing cyber security, cyber attacks still remain a major threat to personal information, the global economy, and national security. In 2013, 7% of US organizations lost $1 million or more due to cybercrime, and 19% of entities had claimed losses between $50,000 and $1 million. Domestically, it is estimated that cyber-attacks cost $300 billion per year and cost $445 billion worldwide [1].

Many existing defensive mechanisms are based on analysis of network traffic or host-based observations, identifying anomalous behavior, looking for known patterns of alerts from monitoring appliances (e.g., intrusion detection systems), or finding malware application signatures. The biggest challenges currently facing cyber defenders are that these defenses tend to be reactive and static, addressing attacks after the damage has already been done and failing to adapt to evolving strategies of advanced adversaries.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 133–138, 2017.
DOI: 10.1007/978-3-319-60240-0_17

These reactionary postures mean that defenders are often one step behind the adversaries. However, what if it were possible to get ahead of the adversaries and proactively deploy defenses that can deter or derail their attack strategies? To gain this advantage, we must first develop analytic methods that provide insight into the cyber adversaries themselves. Rather than viewing cyber attacks simply as a sequence of network traffic or a malicious application, we can look at the human component of an attack, recognizing that at the other end of this attack is a human adversary with specific goals and characteristics that influence their decisions. In this paper, we propose a novel hybrid modeling approach to understanding the cognitive biases, decision-making processes, motivations, and behaviors of cyber adversaries that can be used to simulate cyber attacks to predict and proactively address vulnerabilities.

2 Hybrid Cyber Adversary Models

To understand and analyze behaviors of cyber adversaries, we need a multi-faceted modeling approach that enables us to capture the multiple dimensions and characteristics of human behavior. We have developed a hybrid modeling methodology that combines decision theoretic models, cognitive models, and a reactive agent framework informed by sociocultural context to enable representation of a realistic decision process that manages multiple competing goals and the tradeoffs and biases associated with risks and rewards when planning an attack. This flexible hybrid modeling approach is based on the AgentWorks™ computational modeling framework [2, 3], which provides a mechanism for merging disparate modeling formalisms into coherent, executable agents. Under this paradigm, different types of decision making and behaviors are captured by different mechanisms.

2.1 Decision Theoretic Models

When executing an attack, a cyber adversary inherently assumes some risk, such as the risk of getting caught or the risk of failure. Risk-based models derived from decision theory have been effective at modeling cyber adversary behavior [4], and capturing their decision-making process where they must balance the inherent risk of an attack against the potential reward if successful.

Decision theory posits that an agent will attempt to maximize their utility in any given situation, calculating the potential payoff given the likelihood of success. Using this insight, the authors developed a risk-based formalism for modeling cyber adversaries, defining a mathematical representation of the risk assumed by adversaries according to their characteristics [4]. In our approach, attacker risk is represented as a function of two attributes: (1) attack complexity; and (2) security features. The amount of risk an adversary assumes increases with the complexity of an attack and the security features. However, the rate of this increase in risk depends on two characteristics of an attacker —(1) skill level; and (2) access to resources—represented as parameters in the risk-based models that determine the risk an attacker assumes for an attack. For example, a low-skilled attacker may find a complex attack too risky since there is a high chance of

failure; however, a highly skilled adversary may find the same attack less risky and be more likely to choose this option.

We augmented these basic risk models with concepts from utility curves in micro-economics, assuming an increase in risk is not uniform, but has a marginal growth in the impact of each additional attack step or security feature. We quantify these factors in Eq. (1), where A and S are the observable attack complexity and security features, $1 < \alpha, \beta$ are parameters representing adversary skill level and resources, and d and e are risk constants based on the particular network under study [4].

$$R(A,S) = \frac{d * \frac{1}{\alpha}^{A} \beta^{S}}{\frac{1}{\alpha_{min}}^{A_{max}} \beta_{max}^{S_{max}}} - e \tag{1}$$

The risk model above are based only on characteristics of skill level and access to resources, which may be inferable based on the types of attack patterns seen. However, we identified an additional characteristic—adversary motivation or will—that has a large impact on risk-reward. To incorporate will into the risk function, we considered will as an additive endogenous factor that shifts the risk curve, capturing changes in how the attacker perceives or weighs the risk of a particular situation (e.g., as will increases, the perception of risk decreases—or, said another way, an attacker becomes less risk averse and is willing to assume a greater amount of risk to achieve their goal).

2.2 Cognitive Decision-Making Models

The way in which an adversary makes judgments about the risk-reward tradeoff not only depends on external characteristics, such as skill, but is also influenced by their cognitive and cultural biases. For example, Chinese culture tends to be skeptical of secret information, assuming these sources to be potentially deceptive, making adversaries less likely to value certain types of information in cyber espionage operations. Similarly, most humans will be susceptible to classic biases, such as confirmation bias, which may impact their willingness to take certain risks.

These biases are indicative of the vagueness and imprecision inherent in human decision making. To model this aspect of cyber adversaries, we can use the fuzzy logic component of AgentWorks to assess the tradeoffs of different attacks based on bias. Our fuzzy logic component constructs a fuzzy set of possible states based on the attack options. Once these sets have been generated, they are input into a computational rule-base that uses mathematically-based Boolean functions (e.g., minimum for <and>; maximum for <or>) to combine members from the sets. The rules are weighted based on the importance of the rule to a particular adversary. This weighting also includes aspects of cultural and cognitive biases, enabling us to model the impact on an adversary's perception of their options. Combining these rules, the fuzzy logic component derives a value for each potential goal and action, which can then be used to weight the decision-theoretic risk functions described in the previous section.

2.3 Grammatical Representation of Cyber Attack Vectors

In addition to modeling the decision making of cyber adversaries, to understand how they might exploit vulnerabilities on a network we also need a representation of the actual actions that an adversary can take in a given circumstance. To represent adversary actions, we use a formalism adapted from the sociolinguistic theory of Systemic Functional Grammars (SFGs) [11]. SFGs can model the goals and actions of cyber adversaries, capturing the decomposition of an attack into the actions and the conditions necessary to successfully carry out each goal. SFGs are powerful due to their rich knowledge representation and reasoning capabilities. They are designed to account for contextual information—such as the state of the network, the goals of an attacker, and external factors, such as economic conditions—making them particularly well-suited to modeling complex attack structures for cyber adversaries.

As a sociolinguistic theory, SFGs are widely used in seminal natural language processing (NLP) systems, such as Winograd's language understanding system [12] and the Penman language generation system [13]. We adapted this approach to the cyber domain, representing the sequence of functional choices that can be made by cyber adversaries to achieve their goals. There is a large body of work on using structured approaches to represent cyber attacks. Attack graphs [14, 15] can be used to describe vulnerabilities in a network and support detection of attacks along known vectors. However, these methods tend to be developed from a network rather than a behavioral perspective, requiring redesign each time there is a change. Conversely, SFGs are attack-centric. They focus on the goals and techniques of the attack itself, representing general attack patterns and their constraints. Therefore, SFGs do not need to be redesigned for different scenarios or networks.

The SFG structure has two layers, or strata: the grammatical stratum and the contextual stratum. In language, the grammatical stratum consists of an ontology of grammatical functions where each node may be associated with structural constraints. When applied to a cyber grammar, a node can represent *information gathering* and have the structural constraint that it must appear before a node that *exfiltrates data*. Just as there are millions of possible sentences in the English language, there are millions of possible cyber attacks that can occur in different scenarios. The contextual stratum consists of an ontology of intents and contexts where each node may map to one or more nodes in the grammatical ontology. This mapping from the contextual to the grammatical stratum allows us to generate elements of the grammatical stratum best suited to the current context.

2.4 Reactive Agent Framework for Realistic Goal Prioritization

Finally, our hybrid modeling approach draws from reactive planning [16] to represent the way an adversary might prioritize the goals and actions described in the SFGs in accordance with their biases and risk-reward decision making. Reactive planning models dynamic, adaptive decision making in reaction to beliefs about the state of the world. We have adapted Hap [17], a believable agent architecture developed to drive reactive, realistic agents in simulation environments.

The Hap framework is designed to support highly parallelized behavior and manage the exchange between potentially competing goals. Hap uses information from the probabilistic assessment of the belief state to factor in possible effects of its actions on the current state of the world and reprioritize its goals. Agent behavior is represented by decomposing each goal into subgoals and specific actions designed to accomplish that goal. Using the SFG representation, the Hap model will identify the grammatical stratum behavioral structure consistent with the current context, including the goals of the adversary.

At the top level of the behavior hierarchy are broad sets of activities, such as intelligence gathering and attack methods. At the lowest level, goals are specific actions that manipulate an environment to execute a particular activity. Goals can be either sequential, meaning one item must be completed before another can begin, or parallel, meaning the items can be executed at the same time. Each goal has prerequisites that must be completed before the goal or action can be executed.

3 Discussion and Conclusions

The hybrid modeling approach presented here enables cyber defenders to create rich, realistic models of cyber adversaries. Using this approach, we have constructed models for several behavioral templates, representing a hostile nation state, a hacktivist group, and a "script kiddie" hacker. The modular nature of this methodology allows cyber defenders to design new adversary profiles by recomposing elements of existing models. Adversary models will provide cyber defenders with a deeper understanding of possible vulnerabilities and what types of defensive postures may be most successful by clearly illustrating which vulnerabilities are likely to be targeted and the responses of adversaries to a variety of defensive postures.

As we continue to refine and mature this modeling approach, we are cognizant of a number of significant challenges that lay ahead. For example, the profile above was developed solely through human research and expertise. We plan to explore various automated machine learning approaches that can augment part of this process, particularly in developing the mathematical decision-theoretic models of risk-reward tradeoffs. In addition, validation and verification of models of human behavior are always challenging tasks, and perhaps more challenging for cyber adversaries due to the anonymity of cyberspace. We plan to verify this modeling approach through high-fidelity simulations of cyber attack scenarios, as well as use these methods to make forecasts of future attacks that can be validated against open source attack reporting. Finally, we plan to refine the components of our hybrid modeling approach, enabling more complex representations of biases for risk-reward and situation-assessment, such as more advanced probabilistic models or soft-logic representations. Further, we are exploring integration with simulation engines, such as the OneSAF framework, for using these models in for predictive analytics to develop proactive defenses.

Acknowledgements. This material is based upon work supported by the Communications-Electronics, Research, Development and Engineering Center (CERDEC) under Contract No.

W56KGU-15-C-0053 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 US 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 CERDEC, ODNI, IARPA, AFRL, or the US Government."

References

1. Bremmer, I.: These 5 Facts Explain the Threat of Cyber Warfare, TIME, 19 June 2015
2. Rosenberg, B., Furtak, M., Guarino, S., Harper, K., Metzger, M., Neal Reilly, S., Niehaus, J., Weyhrauch, P.: Easing behavior authoring of intelligent entities for training. In: Conference on Behavior Representation in Modeling and Simulation (BRIMS) (2011)
3. Furtak, M.: Introducing AgentWorks. In: 14th Intelligent Agents Sub-IPT (2009)
4. Li, S., Rickert, R., Sliva, A.: Risk-based models of attacker behavior in cybersecurity. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 523–532. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37210-0_57
5. Pfeffer, A.: Probabilistic relational models for situational awareness. In: AIAA Infotech@Aerospace (2010)
6. Friedman, N., Getoor, L., Koller, D., Pfeffer, A.: Learning probabilistic relational models. In: Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-1999) (1999)
7. Murphy, K.: Dynamic Bayesian networks: representation, inference, and learning, U.C. Berkeley (2002)
8. Pfeffer, A., Tai, T.: Asynchronous dynamic Bayesian networks. In: Uncertainty in Artificial Intelligence (2005)
9. Hongeng, S., Nevatia, R.: Large-scale event detection using semi-hidden Markov models. In: International Conference on Computer Vision, vol. 2, pp. 1455–1462 (2003)
10. Schrodt, P.A.: Forecasting conflict in the Balkans using hidden Markov models. In: Trappl, R. (ed.) Programming for Peace. Springer, Dordrecht (2006)
11. Halliday, M.A.: On Language and Linguistics, vol. 3. Continuum, New York (2003)
12. Winograd, T.: Understanding natural language. Cogn. Psychol. **3**, 1–191 (1972)
13. Mann, W.C., Matthiessen, C.: Nigel: a systemic grammar for text generation, USC/Information Sciences Institute (1983)
14. Phillips, C., Swiler, L.P.: A graph-based system for network-vulnerability analysis. In: Proceedings of the 1998 Workshop on New Security Paradigms, pp. 71–79 (1998)
15. Ammann, P., Wijesekera, D., Kaushik, S.: Scalable, graph-based network vulnerability analysis. In: Proceedings of the 9th ACM Conference on Computer and Communications Security, pp. 217–224 (2002)
16. Firby, J.R.: Adaptive execution in complex dynamic worlds, Yale University, Department of Computer Science (1989)
17. Loyall, A.B.: Believable Agents: Building Interactive Personalities. Carnegie Mellon University, Pittsburgh (1997)

Cyber-FIT: An Agent-Based Modelling Approach to Simulating Cyber Warfare

Geoffrey B. Dobson(✉) and Kathleen M. Carley

Carnegie Mellon University, Pittsburgh, PA, USA
{gdobson,kathleen.carley}@cs.cmu.edu

Abstract. This paper introduces the Cyber - Forces Interactions Terrain (FIT) Simulation Framework. This framework provides an apparatus with which to carry out virtual experiments involving cyber warfare engagements. Our agent-based modelling approach is a first attempt at providing the necessary components with which military planners can reason about cyber force projections on varying terrains and against various adversarial forces. We simulate and then predict the results of cyber warfare at the level historically desired by military planners: vulnerabilities, asset degradation, and mission capability rate.

Keywords: Cyber warfare · Agent-based modelling · Simulation · Military

1 Introduction

The U.S. Department of Defense (DoD) published its Cyber Strategy [3] in 2015, laying out strategic goals and objectives to defend the cyberspace assets that the nation and its allies depend on. The report calls out the need to "establish an enterprise-wide cyber modeling and simulation capability", and to "assess the capacity of the projected Cyber Mission Force to achieve its mission objectives when confronted with multiple contingencies". In this paper, we introduce the Cyber-FIT (Forces, Interactions, Terrain) Framework, which is designed to model and simulate cyber mission forces defending assigned terrain that is confronting multiple contingencies.

Modeling cyber warfare has proven to be very difficult. There are a multitude of variables, many of which are either dependent on the specific situation encountered, or difficult to measure. At the highest level, we can construct a modeling and simulation world, which can allow us to reason about cyber interactions amongst agents. The agents being: "forces" and "terrain", depicted in Fig. 1. By assigning characteristics to the forces, interactions, and terrain, we can observe projected outcomes of cyber engagements.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 139–148, 2017.
DOI: 10.1007/978-3-319-60240-0_18

Fig. 1. Cyber-FIT simulation framework visualization

2 Background

Ormrod, Turnbull and O'Sullivan [7] defined a data representation of cyber attack to model multiple domains common amongst military units. This work improves our understanding of the consequences of cyber warfare. Hamilton [8] described "executable architectures" that can be used to simulate distributed denial of service attacks against a simulated working network architecture. There are a number of simulation tools that work in this manner, but lack the ability to model the interaction of those architectures, attacks, and cyber forces simultaneously. Fischer, Masi, Shortle and Chen [6] presented an Optimal Splitting Technique for Rare Events to simulate the effects on network traffic from a worm based cyber attack. This is an example of modeling terrain damage from specific well known attack behavior. Cayirci and Ghergherehchi [5] created a model that defined human behavior responses to cyber attacks that can be used to design training scenarios. Santhi, Yan and Eidenbenz [4] created CyberSim and simulated a one million node network's response to malware propagation. The attack exploited a specific known vulnerability present in many real systems. For cyber warfare simulation to be realistic, empirically observed computer vulnerabilities must be present in the model. Similarly, military planners must use realistic cyber warfare simulation in order to achieve victory in the newest domain of war.

All of these approaches focus on some aspect of cyber warfare, but none in this field, that we are aware of, exist at a higher level, where we can integrate the behaviors of the systems as a whole. Our approach aims to define the low level interactions, in order to reason about the interplay between humans, technology, and the environment they exist in. We define two classes of agents, terrains and forces, and the interactions that define their behavior. Our primary objective, Cyber-FIT 1.0, is to attempt to answer specific questions about how cyber force packages might perform in realistic missions, thereby defining an expandable framework.

3 The Cyber-FIT Simulation Framework

3.1 Model Definition

The CYBER-FIT framework is an attempt to provide a holistic approach to conducting experiments about the interaction of cyber terrain and forces. It is an agent-based modeling tool built using NetLogo. NetLogo provides a useful interface with which the operator can set parameters, execute the simulation, and then view dependent variables over time. Figure 2 displays the NetLogo interface that controls the model.

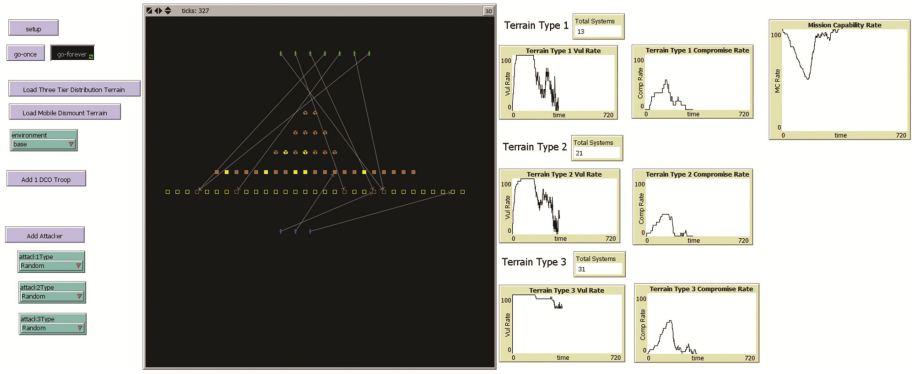


Fig. 2. Cyber-FIT 1.0 NetLogo user interface

3.2 Terrain

Terrain is defined as the computer systems that military units depend on to execute their assigned mission. We use three terrain types, as defined in Table 1.

Table 1. Terrain types

Terrain type	Summary description	Percentage of sampled vulnerabilities
1	Networking systems such as routers and switches	14%
2	Server systems such as web servers, domain controllers, file servers, and intrusion prevention systems	28%
3	User systems such as personal computers, devices, and tablets	58%

The different terrain types will become vulnerable at different rates. The vulnerability rates were computed by taking the known number of vulnerabilities on each of the terrain types from a sample of systems from MITRE's common vulnerability and exposures database, an industry standard for defining, assigning and tracking vulnerabilities [1, 2]. The vulnerability rates are associated with a probability based on the relative number of known vulnerabilities, also shown in Table 1.

The different terrain type vulnerability rates will also be affected by the environment that they are deployed in. The current model defines three environment types that represent common military areas of responsibility. The environments are "base", "tactical", and "industrial". Table 2 provides a description of the three environments currently modeled that will affect terrain characteristics.

Table 2. Terrain environments

Environment	Summary description
Base	The Base environment refers to a long term fixed military installation
Tactical	The Tactical environment refers to a temporary military installation stood up for the purpose of an overseas conflict
Industrial	The Industrial environment refers to a non-military facility that controls an energy production operation the military depends on

The different environments will affect how quickly systems become vulnerable, by terrain type. Based on interviews with vulnerability experts, the terrain types were scored relative to each other, to determine within which environment vulnerabilities appear at higher or lower rates. Table 3 defines the relative vulnerability rate across the three environments and details the probability that the system in that given environment will become vulnerable at any time. This information is incorporated into the code that determines if a given terrain is vulnerable at any given time. That is, in a cell labeled "High", the probability of a system moving from non-vulnerable to vulnerable is equal to the relative share of common vulnerabilities and exposures (CVEs) as defined by MITRE [1, 2]. In a cell labeled "Medium", the probability is reduced 50%. In a cell labeled "Low", the probability is reduced 50% again.

Table 3. Relative vulnerability rates by terrain type across environments

Terrain type	Base	Tactical	Industrial
Type 1 (Networking)	Low	Medium	High
Type 2 (Servers)	Low	High	Medium
Type 3 (Users)	High	Medium	Low

3.3 Forces

Forces are defined as the military members that are deployed to the military scenario. The current version of Cyber-FIT only supports defensive and offensive cyber forces, but future versions will support all force types. The defensive forces are deployed with the purpose of protecting the assigned cyber terrain. The model currently allows the operator to add any number of defensive forces, up to sixteen. The defensive forces will remove vulnerabilities that exist on the terrain at any given hour (each time tick in NetLogo). The defensive forces select vulnerable systems randomly, according to a schedule. At all hours, the forces defend Terrain Type 3, every third hour they defend Terrain Type 2, and every sixth hour they defend Terrain Type 1. This models the real-world constraint that servers and networking equipment can only be defended at certain times, e.g., when they are being patched. The offensive forces will attack the systems based on what type of attack is being launched. The model currently supports three attack types that offensive forces can launch, as defined in Table 4.

Table 4. Offensive force attacks

Attack	Target terrain
Random	All Types
Routing protocol attack	Type 1 (Networking Systems)
Denial of service	Type 2 (Server Systems)
Phishing	Type 3 (User Systems)

3.4 Interactions

Interactions are defined as any instance when a force is actively accessing cyber terrain. In the real world this could be performing operations and maintenance, coding malware, applying patches, etc. In the current version of Cyber-FIT, two types of interactions are modeled: offensive actions and defensive actions, which are limited to offensive and defensive forces, respectively. The defensive forces will perform operations and maintenance activities, and apply patches at every hour to a randomly selected vulnerable system. That system will become non-vulnerable following this interaction. The offensive forces will attack randomly selected systems of the type associated with the attack selected, at every hour.

In order for a system to become compromised, it must be vulnerable at the time that it was attacked (an offensive interaction by offensive force). If vulnerable, then the system has a 5% chance of becoming compromised. Currently all systems are modeled to have a 5% compromise rate, given that the offensive force has access and the system is vulnerable.

3.5 Model Outputs

The model currently outputs seven dependent variables: vulnerability rate per terrain type, compromise rate per terrain type, and overall mission capability rate. Table 5 describes each dependent variable.

Table 5. Dependent variable descriptions

DV	Description
Mission capability rate	Average Percentage of systems (all types) available
Vulnerability rate	Average Percentage of systems vulnerable (by type)
Compromise rate	Average Percentage of systems compromised (by type)

4 Virtual Experiments

We conducted three virtual experiments using the current model, seeking to answer questions a planner might have. For each experiment we provide the virtual experiment motivation, the results of the experiment, and discussion.

4.1 How Many Forces Should We Deploy to Minimize the Effect of a Routing Protocol Attack (RPA) in an Industrial Environment?

In this experiment, we are considering a specific attack (RPA), in a specific environment (base). We'll vary the number of forces from one through fifteen and examine the decrease on Type 1 system (networking) compromise rate. We're specifically searching for the number of forces, where, when adding one more troop, the projected compromise rate is within one standard deviation of the current projected force package effectiveness. We expect that as the number of forces increases, decrease in compromise rate will level off. Results are shown in Fig. 3 and Table 6.

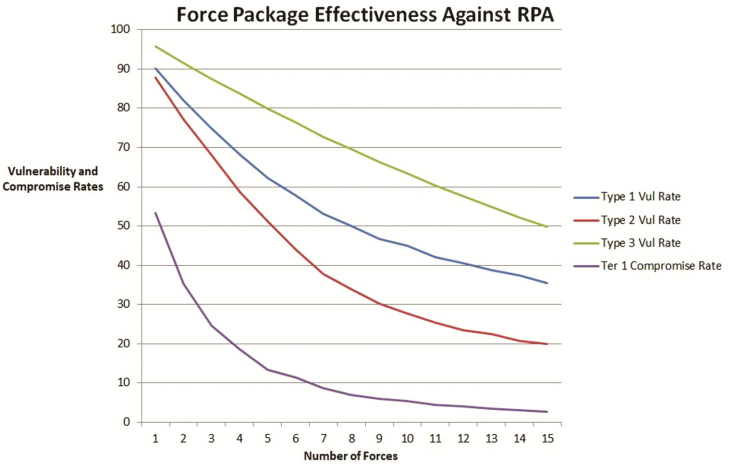


Fig. 3. Projected force package effectiveness against RPA

Table 6. Projected compromise rate and standard deviation of Type 1 systems

Forces	Compromise rate	Standard deviation	Forces	Compromise rate	Standard deviation
1	53.51	2.68	9	6.06	0.52
2	35.33	3.20	10	5.36	0.72
3	24.68	2.44	11	4.37	0.77
4	18.60	1.79	12	4.06	0.82
5	13.74	1.88	13	3.39	0.37
6	11.34	0.96	14	3.13	0.54
7	8.70	0.99	15	2.73	0.37
8	6.96	0.74			

As shown in Fig. 3, we can expect a substantial increase in effectiveness moving from one troop to five. After five troops, the projected performance improvement tapers off. We still see improvements on the projected compromise rate of Terrain Type 1, our primary concern in this simulated mission, but it will be decreasing as

we continue to add forces. To find the point when adding troops will make no difference at all, we search for the point where the increase in effectiveness is within one standard deviation of the current projected average Type 1 compromise rate. This is laid out in Table 7. This point is found, at forces = 11. At that point, the projected compromise rate is 4.64 with a standard deviation of 0.77. The projected compromise rate, when adding one more troop to the mission, is 4.06, within one standard deviation of the previous projection.

This shows the importance of weighing the cost of adding more resources with the effectiveness of those resources. In this scenario, what do these numbers represent? We have a simulated mission on terrain that includes 21 Type 1 systems. So, if the average compromise rate, at forces = 5, is 13.74, then we can expect, on average, 2.89 systems are always compromised when facing a routing protocol attack. At forces = 6, we can expect, on average, 2.38 systems are always compromised when facing a routing protocol attack. So, somewhere between two and three systems will go down. Perhaps this is acceptable risk? Also, once the attack is recognized, will five forces be enough to make an emergency change, repair the compromised terrain, and block the attack? This might be the case, which means that the planner should actually choose to deploy five forces, rather than eleven, due to acceptable level of risk, external constraints, and knowledge of mission resources.

4.2 What Will Be the Expected Effect on Cyber Terrain if the Adversary Switches from a Fifteen Day Routing Protocol Attack, to a Denial of Service Attack in a Base Environment with Six Troops Deployed?

In this experiment, we are considering the difference in how the forces and terrain will perform against two different types of attacks. Military deception has been around for as long as human warfare. This occurs quite frequently in the cyber domain. Offensive forces will start one attack, in order to focus resources on specific terrain, only to then switch the attack on different terrain. This is the attack vector we are modeling in this experiment. The adversarial force will begin with an RPA, and then switch to DOS attack halfway through the deployment time frame. Figure 4 shows the change in compromise rate of Type 1 and Type 2 systems, of one run of the virtual experiment. Table 7 shows the average compromise rate of the Type 1 and Type 2 systems, after all virtual experiment runs.

Table 7. Average compromise rate of Type 1 and Type 2 systems

Summary of simulations	
Number of forces	6
Environment	Base
Terrain architecture	Three Tier Distribution
Compromise rate of Type 1 systems	1.24
Compromise rate of Type 2 systems	0.89

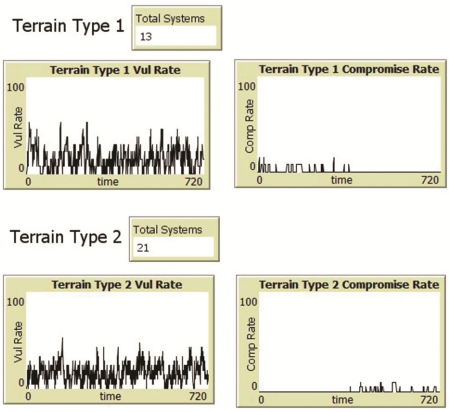


Fig. 4. Visualization of simulation results

The importance of visualization is displayed in Fig. 4. The Cyber-FIT interface displays real-time feedback to the user showing exactly what is occurring on the terrain at every time interval. This aids planners and researchers by allowing them to carry out test runs and ensure what they have conceived, conceptually, matches what the model is providing. In Fig. 4, we can see that in the given circumstances, the terrain will hold up quite well against both attacks. The terrain and number of forces deployed, in the base environment will handle a DOS attack better than an RPA. This means that planners and enterprise architects can address this difference. If the difference isn't acceptable, leadership could send additional resources to the Type 1 systems in the way of additional forces or a better maintenance schedule, to decrease the expected compromise rate.

4.3 What Number of Forces Maximizes Expected Cyber Terrain Mission Capability Rate Against Random Attacks in a Tactical Environment?

In this experiment, we are considering a tactical deployment and attempting to determine which number of forces maximizes the mission capability rate when the adversary is launching random attacks against the cyber terrain. When military planners are considering what resources to send to battle, they will attempt to package forces and equipment that will perform at a high level. Since resources are limited, a challenging part of their job is deciding which number of forces will maximize the likelihood that each unit will accomplish its mission. For this experiment, we are modeling a situation where the planners are considering a deployment of cyber terrain which will likely be attacked in multiple ways. So, we selected random cyber attacks for the adversary. Then, we simulated cyber battles against the terrain, each time increasing the number of forces. Figure 5 shows the results of the simulations.

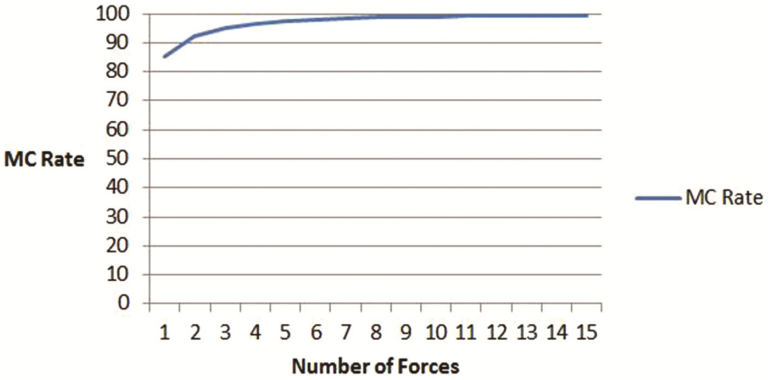


Fig. 5. Projected mission capability rate as forces increase

Figure 5 shows that the projected mission capability rate will increase sharply as forces are added. A force package of six troops should provide a mission capability rate above 98.0%. A force package of ten troops should eclipse a 99.0% mission capability rate. The highest number of troops deployed for this set of experiments was 15, resulting in an average mission capability rate of 99.55%. This information would prove valuable for determining the number of troops to deploy to this type of mission.

5 Discussion

The Cyber-FIT simulation framework, in current form, presents a successful proof of concept. The three elements of the model (forces, interactions, and terrain) are all conceptual at this time. Forces differ in vulnerability patching routines, and attack targets. Further development of forces could include: skill level, specialty, and experience. Terrains differ in types of systems present, vulnerability state, and environmental deployment. Further development of terrain could include: increasing types of systems, realistic lists of vulnerabilities, cost, and access control.

There are nearly limitless potential extensions to this work. For example, in future work we plan to explore various improved definitions of mission capability rate. To define that, we'll model various units that depend on different parts of the terrain for mission success. Mission capability rate will be defined as the ability to provide working systems, when demanded, to various units. Another example would be adding different types of adversary complexities. Hacktivist organizations, organized crime rings, and nation states would all have different adversarial capabilities and limitations. Then the simulation could predict performance of the forces and terrain against different classes of adversaries

6 Conclusion

We introduced the Cyber-FIT simulation framework, an agent-based cyber warfare simulation framework. We showed that the framework can enable virtual experiments that answer questions about military cyber force projections. Three virtual experiments were conducted, each testing specific questions currently being considered by military planners all over the world. In the first experiment, we found that adding any number over 11 troops does not improve terrain performance. In the second virtual experiment, we found that the terrain would handle a denial of service attack better than a routing protocol attack. In the third virtual experiment we found that a force package of ten troops would provide a cyber terrain mission capability rate above 99%.

The Cyber-FIT simulation framework will be further developed by adding empirical data. This will provide more realistic virtual experiments. Future work will focus on presenting simulations to Department of Defense experts interested in specific questions that cannot be addressed in real world scenarios due to limitations of time and resources. Our long term goal is to continually add modules that can take disparate model results as input to our model.

References

1. MITRE Common Vulnerabilities and Exposures. http://cve.mitre.org/
2. MITRE CVE Details. http://www.cvedetails.com/
3. Department of Defense, The DoD Cyber Strategy. DoD, Washington D.C. (2015)
4. Santhi, N., Yan, G., Eidenbenz, S.: CyberSim: geographic, temporal, and organizational dynamics of malware propagation. In: Proceedings of the 2010 Winter Simulation Conference, pp. 2876–2887 (2010)
5. Cayirci, E., Chergherehchi, R.: Modeling cyber attacks and their effects on decision process. In: Proceedings of the 2011 Winter Simulation Conference, pp. 2632–2641 (2011)
6. Fischer, M.J., Masi, D.M.B., Shortle, J.F., Chen, C.H.: Simulating non-stationary congestion systems using splitting with applications to cyber security. In: Proceedings of the 2010 Winter Simulation Conference, pp. 2865–2875 (2010)
7. Omrud, D., Turnbull, B., O'Sullivan, K.O.: System of systems cyber effects simulation ontology. In: Proceedings of the 2015 Winter Simulation Conference, pp. 2475–2486 (2015)
8. Hamilton Jr., J.A.: DoDAF-based information assurance architectures. CrossTalk **19**, 4–7 (2006)

Information, Systems, and Network Sciences

Large-Scale Sleep Condition Analysis Using Selfies from Social Media

Xuefeng Peng[1,2(✉)], Jiebo Luo[1,2(✉)], Catherine Glenn[1,2], Jingyao Zhan[1,2], and Yuhan Liu[1,2]

[1] Department of Computer Science, University of Rochester, Rochester, NY 14627, USA
xpeng4@u.rochester.edu,
{jiebo.luo,catherine.glenn}@rochester.edu
[2] Department of Psychology, University of Rochester, Rochester, NY 14627, USA

Abstract. Sleep condition is closely related to an individual's health. Poor sleep conditions such as sleep disorder and sleep deprivation affect one's daily performance, and may also cause many chronic diseases. Many efforts have been devoted to monitoring people's sleep conditions. However, traditional methodologies require sophisticated equipment and consume a significant amount of time. In this paper, we attempt to develop a novel way to predict individual's sleep condition via scrutinizing facial cues as doctors would. Rather than measuring the sleep condition directly, we measure the sleep-deprived fatigue which indirectly reflects the sleep condition. Our method can predict a sleep-deprived fatigue rate based on a selfie provided by a subject. This rate is used to indicate the sleep condition. To gain deeper insights of human sleep conditions, we collected around 100,000 faces from selfies posted on Twitter and Instagram, and identified their age, gender, and race using automatic algorithms. Next, we investigated the sleep condition distributions with respect to age, gender, and race. Our study suggests among the age groups, fatigue percentage of the 0–20 youth and adolescent group is the highest, implying that poor sleep condition is more prevalent in this age group. For gender, the fatigue percentage of females is higher than that of males, implying that more females are suffering from sleep issues than males. Among ethnic groups, the fatigue percentage in Caucasian is the highest followed by Asian and African American.

Keywords: Sleep condition prediction · Fatigue analysis · Social media · Selfies

1 Introduction

In modern societies, with growing pressures from life and work, sleep condition has become increasingly a big concern to people. Individuals who encounter sleep disorders such as insomnia, sleep deprivation, sleep apnea, and so on may not only appear less healthy and attractive [6], but also suffer from poor physical and mental performances during the daytime [3]. To prevent and minimize the impairments caused by sleep disorders, many researchers endeavor in studying human sleep conditions. Traditionally, there are two popular types of methods. The first type is self-report based; one example of such type is the Pittsburgh Sleep Quality Index (PSQI) [8]. Another type is electronic device based; this

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 151–161, 2017.
DOI: 10.1007/978-3-319-60240-0_19

type of methods measures individual sleep quality via digital PSG recording and scoring [11]. The former relies on individual's self-report, which may carry significant biases, while the latter requires complicated medical equipment and time. Such drawbacks prevent gaining deeper insights about human sleep conditions at a large scale.

Therefore, we are particularly interested in (1) finding a new way to predict sleep condition in a relatively easier and faster fashion, and (2) applying this new way to analyze the sleep conditions of the massive number of selfies on the social media.

With respect to our first goal, we understand that measuring the sleep condition directly could be difficult, thus we have decided to approach it indirectly by looking at sleep-deprived fatigue. It has been clinically identified that fatigue is one of the most common symptoms of poor sleep condition. A research paper has quantitatively associated sleep-deprived fatigue rate with human facial cues [20]. According to [20], sleep deprivation caused fatigue is heavily correlated with eight facial cues, including hanging eyelids, red eyes, dark circles under eyes, pale skin, droopy corner mouth, swollen eyes, glazed eyes, and wrinkles/lines around eyes. Figure 1, which is reprinted from [20], shows the correlations between the perceived fatigue rate and those facial cues. The rate from 0 to 100 indicates the degree of the facial cue from 'not at all' to 'very'.

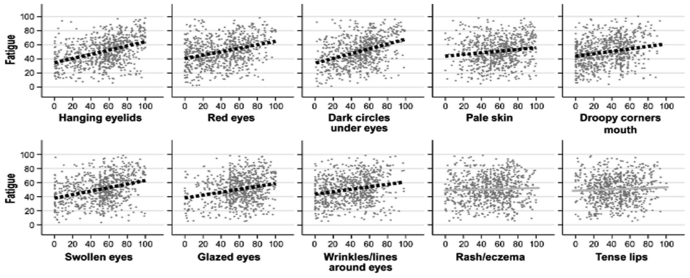


Fig. 1. Relationships between fatigue rates and the eight facial cues rates, reprinted from [20]. The first eight charts demonstrate that sleep-deprived fatigue correlates with the eight facial cues, whereas the last two show that sleep-deprived fatigue does not correlate with Rash/eczema and Tense lips.

Based on the correlation coefficients given in Fig. 1, we constructed eight separate models (explained in Sect. 2) to measure the eight facial cues for each individual. Next, with the measurement results, the overall fatigue rate can be quantified and treated as an indicator of this individual's sleep condition. Using these models along with our fatigue predicting criterion (explained in Sect. 2.5), given a selfie, we can predict the perceived fatigue rate from the face within milliseconds, and with that fatigue rate we can preliminarily assess this individual's sleep condition.

We applied our prediction method on around 100,000 faces obtained from user timelines on Twitter and Instagram. For each face, we first utilized the *Face*++ API[1], which is the state-of-the-art and reliable open-access face engine [4], to identify its demographic information and facial landmarks. We then located the interest areas that

[1] http://www.faceplusplus.com.

can indicate the facial cues through the landmarks. Those interest areas are sent to our models to produce an overall fatigue rate.

Our study found among the age groups (10-year per interval), fatigue percentage of 10–20 adolescent group is the highest, implying that poor sleep condition is more prevalent in this age group. This echoes the finding in [14] that sleep problems are common among adolescents. For gender, the fatigue percentage of females is higher than that of males, implying that more females are suffering from sleep issues than males. This result is consistent with the finding in [2]. Finally, among racial groups, the fatigue percentage in Caucasian group is the highest followed by Asian and African American.

Specifically, our main contributions include: (1) we develop a new method to predict sleep-deprived fatigue rate for a given face, (2) we apply our methodology at a large scale to study the population sleep condition using selfies from social media, and discover the correlations between sleep-deprived fatigue and demographic information.

2 Model Construction

Briefly, for each given face, we need to check eight facial cues that are highly correlated with sleep-deprived fatigue shown in Fig. 1. To compute the fatigue rate according to the linear regression estimators given by Fig. 1, the rate for each facial cue is also necessary. Therefore, we cooperated with the YR^{2}Lab in the Clinical and Social Psychology Department at the University, which provides the ratings of the eight facial cues for each training face. Next, we trained eight independent regression models where each of them will predict one corresponding facial cue rate. Subsequently, the combined linear regression estimator (Eq. 1 in Sect. 2.3) is used to generate the overall fatigue rate.

2.1 Training Data Collection and Rating

Our most ideal training dataset would be a dataset that contains enough unique faces, some of which appear more fatigued than others, and some of which do not look fatigued at all. Moreover, for each unique face in the dataset, a few more facial images of that same face are also needed as references for the rating process (explained in Sect. 2.2). Considering the expectations above, we have chosen the COLORFERET database[2] as our training dataset. The COLORFERET database is sponsored by NIST[3]. For each face, more than five high-quality facial imageries are provided. We picked 964 faces as our

Table 1. Age, gender, and race distribution of training set. Note that the average prediction confidences for gender and race are 95.2% and 90.5% respectively.

Age			Gender		Race		
0–20	20–40	4060	Female	Male	Asian	African American	Caucasian
140	607	178	306	658	148	152	664

[2] https://www.nist.gov/itl/iad/image-group/color-feret-database.
[3] https://www.nist.gov.

training dataset. Following Table 1 shows the distribution of age, gender, and race. Figure 2 shows the overall fatigue rate distribution.

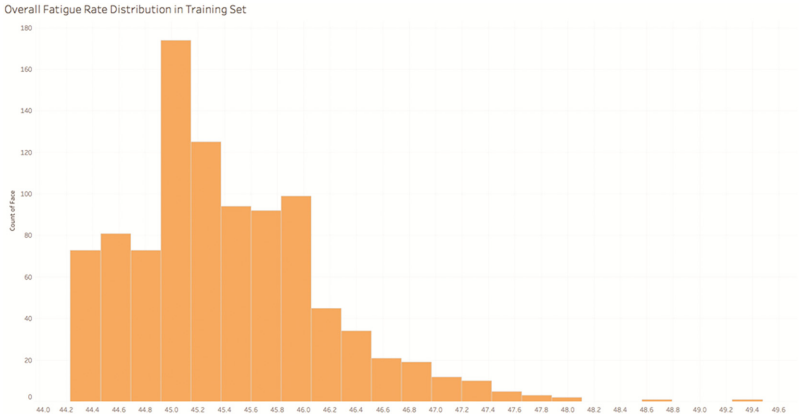


Fig. 2. Overall fatigue rate distribution for faces in training set.

For each training face, there are eight facial cues, to make the rating as objective as possible, we have invited three YR^{2}Lab[4] members with prior clinical research experiences to help us rate the training faces. All three raters have received an inter-judge agreement that describes the details about which area to scrutinize while rating a facial cue. Integers from 0 to 4 were used to indicate the rate of each facial cue with 0 means 'not at all' whereas 4 means 'very'. In addition, we applied three techniques to ensure the objectiveness of the ratings. First, each of the three raters was asked to rate eight facial cues for all 964 training faces, and the final rating for a facial cue of a face is calculated as the mean of the ratings given by the three raters. Second, sometimes the rater may be influenced by the previous facial image while rating the current one. Such influence could be significant if the display order is identical to all three raters. Therefore, we randomized the display order of the training faces for each rater to minimize such influence. Third, while displaying a face, the to-be-rated facial image is displayed, along with four or more images of the same face as references.

2.2 Feature Extraction for Facial Images

Six areas of interest are used to identify the eight facial cues of any given face. Left and right eye areas are examined to identify hanging eyelids, red eyes, swollen eyes, glazed eyes, and wrinkles/lines around eyes. The left and right eye bottom areas identify dark circle under eyes. The cheek area is for pale skin, and finally the mouth area is for droopy corner mouth.

For each training face, we first called the *Face++* API to obtain the facial landmarks of that face. We then cropped the areas of interest from the original facial image

[4] http://www.yr2lab.com/.

accordingly. The feature extraction algorithm we employed is Dense SIFT [1]. This algorithm will generate a feature descriptor for each cropped image. Note that for eye-related areas of interest, we concatenated the feature descriptors of the left and right eyes as a single feature descriptor, which we name eye feature descriptor; and we created the eye bottom feature descriptor in the same fashion.

2.3 Model Training and Prediction

We have eight facial cues to rate for any given face, and the rate of each facial cue depends on one corresponding feature descriptor. Therefore, we built eight separate models to predict the rate for each facial cue.

We fit our data by an ensemble of regression leaners. Prior to the training process, we standardized each feature descriptor. Next, we trained each model with the corresponding feature descriptors and the ground truth facial cue rates.

We performed Bayesian Optimization [19] to optimize the hyper-parameters of our models. The optimization aims to locate the combination of (1) ensemble-aggregation method, (2) number of ensemble learning cycles, (3) learning rate for shrinkage, and (4) minimal number of leaf node observations that produces the least 5-fold cross validated RMSE. The following Table 2 shows the best hyper-parameters for each model found within 30 iterations, along with the RMSE rates.

Table 2. Parameters and performance of models. P1, P2, P3, and P4 represents method, learnCycle, learnRate, and minLeafSize respectively. X_i denotes the model for facial cue $\boldsymbol{x_i}$.

mdl	P1	P2	P3	P4	RMSE	mdl	P1	P2	P3	P4	RMSE
X1	LSBoost	499	0.011	122	1.825	X5	Bag	340	-	27	1.747
X2	LSBoost	50	0.090	146	1.651	X6	Bag	195	-	1	2.056
X3	LSBoost	86	0.065	67	1.983	X7	Bag	118	-	59	1.895
X4	Bag	494	-	481	1.745	X8	LSBoost	499	0.017	42	2.033

Those eight models together constitute of our final model, which we call composite model, and it produced a 1.16% 5-fold cross validated SMAPE[5] on overall fatigue rates.

We located the coordinates of each linear regression estimator (black dot line) in the charts given by Fig. 1 to derive its mathematical expression. Since the ranges for the fatigue rate and each facial cue rate are equivalent, we therefore can derive a combined linear regression estimator to compute the overall fatigue rate by averaging those eight models as

$$y = \mathbf{0.037}x_1 + \mathbf{0.03}x_2 + \mathbf{0.041}x_3 + \mathbf{0.014}x_4 + \mathbf{0.022}x_5 + \mathbf{0.033}x_6 + \mathbf{0.027}x_7 + \mathbf{0.024}x_8 + \mathbf{44.41} \quad (1)$$

Eight variables $x_1, x_2, x_3, x_4, x_5, x_6, x_7$, and x_8 represent the rate of hanging eyelid, red eye, dark circle, pale skin, droopy corner mouth, swollen eye, glazed eye, and wrinkles, respectively. The following is the procedure outline for predicting the overall fatigue

[5] $\text{SMAPE} = \frac{2}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{|F_t| + |A_t|}$.

rate for any given face: (1) obtain the facial landmarks, (2) crop the interest areas and run corresponding models to obtain the rate of each facial cue, (3) compute the overall fatigue rate through Eq. 1.

3 Selfie Collection, Processing and Prediction

There are several keywords that people frequently tag when they post selfies on social media. They are "#selfie", "#me", "#happy", "#fun", "#smile", "#nomakeup", "#friends", and "#family". We used those tags to search the photo posts on Twitter and Instagram. For each post, we acquired a uid (user Id), and such uid can be utilized to backtrack the uid owner's timeline posts, which are with keywords.

It is possible that photos in a user's timeline contain not only his/her own faces, but also the faces of his/her friends, family members, and even strangers. Therefore, we grouped the faces into distinct face sets. Subsequently, we identified age, gender and race, located the facial landmarks of every single face in the face sets, then the areas of interest were extracted and the feature descriptors were generated. Finally, the overall fatigue rate of each face is predicted. The algorithms we have used for face detecting, facial landmarks locating, and face grouping are provided by *Face*++ API.

4 Main Result

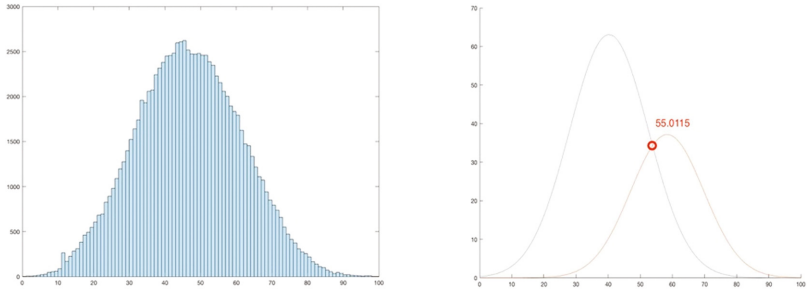
The demographic distributions of the social media selfies we have collected are summarized in the following Table 3.

Table 3. Demographic distribution. The average prediction confidence for gender and race are with 94.6% and 86.2%, respectively.

Age (10-year per interval)							Gender		Race		
0–1	1–2	2–3	3–4	4–5	5–6	6–7	F	M	A	AA	C
7624	23666	34251	23814	5678	2587	1096	58129	39033	17394	6168	73600

To see the overall fatigue rate differences more clearly, we normalized the overall fatigue rates into [0,100]. To pick the threshold that classifies faces into the fatigue set and non-fatigue set, we fitted our data into two Gaussian distributions. The right chart in Fig. 3 shows the two distributions. We have chosen their intersection 55.0115 as our classification threshold (circled in red in Fig. 3).

Using this threshold, we found that among all the unique faces we have collected, 29.09% faces are classified into the fatigue class, and the remaining 70.91% faces are classified into the non-fatigue class.



(a) (b)

Fig. 3. The overall Fatigue rate distribution in (a), and the fitted Gaussian distributions in (b). (Color figure online)

4.1 Age, Gender, and Race

In terms of age, more people in the youth and middle-age groups tend to appear fatigued and thus suffer from worse sleep conditions, one possible causation could be higher rates of anxiety and stress among these groups. The fatigue percentage of 0–20 age group is the highest, as the fatigue proportion difference between 0–10 and 10–20 groups is statistically insignificant. We employed Marascuilo procedure [12] to conduct the multiple proportions test, and the test result suggests that the proportion differences between 0–20 age group and other groups are statistically significant. An explanation is that many adolescents in this age group appear fatigue mainly due to their school workload or their frequent stay-up-late behaviors. A worldwide research on adolescent sleep [14] reveals that sleep problems are common among adolescents and many countries have reported high incidences of sleep disturbances in these age groups; and this outcome further confirms this notion using a large-scale data set that is obtained unobtrusively. We reported the critical value, critical range, and 95% confidence interval in Table 4 (Fig. 4).

Table 4. The details of the statistical significance test on fatigue proportion difference among age groups. According to [12], the difference is significant if the critical value exceeds the critical range. Note we only report significant differences. 0 stands for 0–10, 1 stands for 10–20, so forth.

Age	Critical val.	Critical range	95% CI (%)	Age	Critical val.	Critical range	95% CI (%)
0–3	0.021	0.020	1.05–3.24	1–4	0.037	0.023	2.49–5.00
0–4	0.027	0.026	1.22–4.17	1–5	0.055	0.031	3.82–7.26
0–5	0.045	0.034	2.61–6.37	1–6	0.076	0.046	4.97–10.19
0–6	0.065	0.048	3.81–9.25	2–5	0.036	0.030	1.94–5.31
1–2	0.019	0.013	1.19–2.63	2–6	0.057	0.046	3.08–8.26
1–3	0.032	0.014	2.40–3.99				

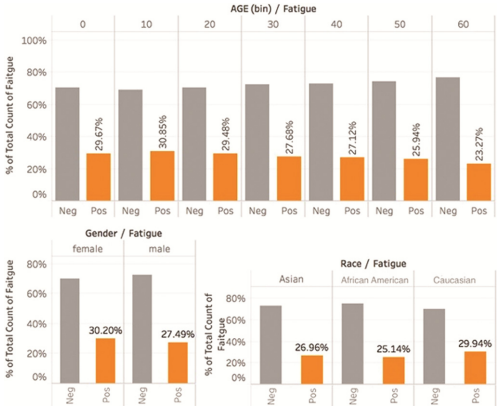


Fig. 4. Age, gender and race proportions on fatigue.

With respect to gender, the proportion test indicates that the fatigue percentage of female is 2.1% to 3.3% higher than that of male with 95% confidence. This result echoes one of the findings in the 2005 Sleep in American poll summary [2] that females are more likely than males to have difficulty falling and staying asleep and thus more likely to experience more daytime fatigue. Among racial groups, our result suggests that the fatigue percentage in Caucasian group is the highest followed by Asian and African American. The Marascuilo procedure suggests that the proportion differences among those racial groups are statistically significant with 95% confidence interval of differences between Asian and African American, Asian and Caucasian, African American and Caucasian are 0.5% to 3.0%, −3.1% to −1.6%, and −5% to −3.1%, respectively.

5 Related Work

Our work builds upon previous research on sleep condition, fatigue studies, and computer vision. In sleep condition and fatigue studies, several studies have pointed out that fatigue is one of the most common symptoms of sleep deprivation. Furthermore, researchers also have found that the sleep-deprived fatigue rate is correlated with eight facial cues [20], and our study is based on the assumption that sleep-deprived fatigue is mainly reflected by those eight facial cues, and the fatigue rate can imply sleep condition. In terms of computer vision research, our work is related to face detection [16], gender, race and age identification [5], facial landmarks location [15], face grouping [21], and using visual social media to monitor mental health [10].

6 Limitations and Future Work

Our work is built on the assumptions that the sleep-deprived fatigue can indicate one's sleep condition, and that sleep-deprived fatigue is associated with the eight facial cues. However, it is possible that the fatigue appearance is caused by some other chronic

diseases [13] or a day of extremely heavy labor. Note that the use of social media selfies as the sensory data partially mitigates the above factors because people in those conditions are unlikely to post selfies. In addition, some people may naturally appear more fatigue than others thus using fatigue to infer their sleep conditions could be biased. Therefore, we plan to develop more robust techniques to establish the face appearance *baseline* for a given individual before inferring his/her sleep condition. In addition, the overall fatigue rates of individuals who wear make-up can be underestimated. That is why we included "#nomakeup" as one of our keywords while retrieving selfies. Lastly, due to the restriction in accessing the original data in [20], Eq. (1) is solely derived by averaging those eight linear estimators presented in Fig. 1, which may lead to inaccurate measurement of the impact each facial cue constitutes for the overall fatigue rate. Nonetheless, we believe the trends and distributions in our study will remain consistent and valid, especially given that a large number of facial images are analyzed.

Our current research primarily focuses on studying the sleep condition distribution regarding to the demographic information. In the future, we may consider ways to identify the occupations of the face owners to investigate the distribution among different occupations. A recent study [9] has proposed a way to identify if a Twitter user is a college student or not. This method may help us unveil more interesting fatigue patterns of the student population. Moreover, we are also interested in examining the fatigue distribution over different geographical areas.

7 Conclusions

We have developed a new method to gauge a face's overall fatigue rate and use this rate to predict the face owner's sleep condition. This leads to a data-driven methodology to include a massive number of faces on social media to obtain the fatigue distributions with respect to age, gender, and race. Our main findings are largely consistent with those reported by using conventional small-scale empirical studies. This is extremely encouraging as it validates the effectiveness of the data-driven approach to study public health at large scales. Some of our findings are beyond those reported in the literature, e.g., those related to the interplays of age, gender and race, pointing to the potential to discover factors that the empirical studies have overlooked. Moreover, analyzing social media posts and pictures offers the potential to provide a method for mass screening for individuals at risk for a range of poor health conditions. Social media has been used to pick up signals when individuals use language that could be related to risky behaviors [17]. Tracking pictures could work in the same way to detect risk and allow for early intervention. Our work is in the same vein as [17, 18] and such social media-driven methods are expected to find more successes in computational psychology.

Acknowledgement. We thank the support of New York State through the Goergen Institute for Data Science, and our corporate research sponsors Xerox and VisualDX.

References

1. Vedaldi, A., Fulkerson, B.: Vlfeat: an open and portable library of computer vision algorithms. In: Proceedings of the 18th ACM International Conference on Multimedia (MM 2010), New York, NY, USA, pp. 1469–1472 (2010)
2. Anon: 2015 Sleep in America Poll. Sleep Health **1**, 2 (2015)
3. Griffith, C., Mahadevan, S.: Sleep deprivation effect on human performance: a meta-analysis approach (PSAM-0010). In: Proceedings of International Conference on Probabilistic Safety Assessment & Management (PSAM), pp. 1488–1496
4. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 386–391 (2013)
5. Han, H., Jain, A.K.: Age, Gender and Race Estimation from Unconstrained Face Images. Michigan State University, Technical Report (2014)
6. Axelsson, J., Sundelin, T., Ingre, M., Van Someren, E.J.W., Olsson, A., Lekander, M.: Beauty sleep: experimental study on the perceived health and attractiveness of sleep deprived people. BMJ **341**, c6614 (2010)
7. Desforges, J.F., Prinz, P.N., Vitiello, M.V., Raskind, M.A., Thorpy, M.J.: Sleep disorders and aging. New England J. Med. **323**(8), 520–526 (1990)
8. Lack, L., Wright, H.: Pittsburgh sleep quality index. In: Encyclopedia of Quality of Life and Well-Being Research, pp. 4814–4816 (2014)
9. He, L., Murphy, L., Luo, J.: Using social media to promote STEM education: matching college students with role models. In: Berendt, B., Bringmann, B., Fromont, É., Garriga, G., Miettinen, P., Tatti, N., Tresp, V. (eds.) ECML PKDD 2016. LNCS, vol. 9853, pp. 79–95. Springer, Cham (2016). doi:10.1007/978-3-319-46131-1_17
10. Manikonda, L., De Choudhury, M.: Modeling and understanding visual attributes of mental health disclosures in social media. In: Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) (to appear, 2017)
11. Tripathi, M.: Technical notes for digital polysomnography recording in sleep medicine practice. Ann. Indian Acad. Neurol. **11**(2), 129 (2008)
12. Marascuilo, L.A.: Large-sample multiple comparisons. Psychol. Bull. **65**(5), 280–290 (1966)
13. Swain, M.G.: Fatigue in chronic disease. Clin. Sci. **99**(1), 1 (2000)
14. Gradisar, M., Gardner, G., Dohnt, H.: Recent worldwide sleep patterns and problems during adolescence: a review and meta-analysis of age, region, and sleep. Sleep Med. **12**(2), 110–118 (2011)
15. Wang, N., Gao, X., Tao, D., Li, X.: Facial feature point detection: a comprehensive survey. CoRR (2014)
16. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Computer Vision and Pattern Recognition (2001)
17. Pang, R., Baretto, A., Kautz, H., Luo, J.: Monitoring adolescent alcohol use via multimodal data analysis in social multimedia. In: Proceedings of IEEE Big Data Conference on Special Session on Intelligent Mining (2015)
18. Abdullah, S., Murnane, E.L., Costa, J.M.R., Choudhury, T.: Collective smile: Measuring societal happiness from geolocated images. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (2015)
19. Snoek, J., Larochelle, H., Adams, R.P.: Practical bayesian optimization of machine learning algorithms. Adv. Neural. Inf. Process. Syst. **25**, 2960–2968 (2012)
20. Sundelin, T., Lekander, M., Kecklund, G., Van Someren, E.J.W., Olsson, A., Axelsson, J.: Cues of fatigue: effects of sleep deprivation on facial appearance. Sleep, January 2013

21. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: a literature survey. ACM Comput. Surv. **35**(4), 2003 (2003)
22. Wu, Y., Yuan, J., You, Q., Luo, J.: The effect of pets on happiness: a data-driven approach via large-scale social media. In: Proceedings of IEEE Big Data Conference (2016)

Modeling the Co-evolution of Culture, Signs and Network Structure

Peter Revay(✉) and Claudio Cioffi-Revilla

Center for Social Complexity, George Mason University, Fairfax, VA, USA
{pfroncek,ccioffi}@gmu.edu

Abstract. We use agent-based simulation to investigate the interplay between the acquisition and transmission of cultural traits, the dynamics of social network structure and the emergence of meaningful signs. We assume agents in our model must cooperate to thrive in their environment and to be successful in doing so, they must synchronize their otherwise selectively-neutral strategies. We further assume that maintenance of social ties is costly, and that agents cannot directly identify the strategies of their counterparts. We show that when cooperation is biased by the possession of arbitrary observable markers, evolutionary dynamics lead to small-world social structures with communities defined by shared culture and the establishment of markers as signs of community membership.

Keywords: Agent-based modeling · Evolution · Culture · Social networks

1 Introduction

In this study we investigate the relationship between evolution of culture and the evolution of social network structure. Specifically, we explore the interplay between selectively neutral cultural traits, selectively neutral external markers (or 'tags' as they are referred to elsewhere in ABM literature) and the maintenance of social ties in a context which requires cooperation and coordination of efforts. Humans are cultural organisms: they can acquire and transmit shared sets of values, knowledge and behaviors [1]. Often these attitudes and behaviors are functionally equivalent to their alternatives. For example, distances can be measured in metric or imperial units, and the findings can be communicated effectively, as long as the actors involved share knowledge regarding which system is being used and a familiarity of that system. The only problem would arise, if some of the actors were unaware that one system was being used rather than the other, or if they were unfamiliar with it. More complex tasks, such as large construction projects, often require the collaboration of multiple individuals and their success depends largely on the assumption that all parties involved possess the same set of domain knowledge. Another aspect of culture are sign systems. Languages are the most complex sign systems, but other more rudimentary examples are commonly encountered as well, such as flags representing

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 162–171, 2017.
DOI: 10.1007/978-3-319-60240-0_20

nations. Signs can be used to mark the possession of cultural traits, which might be difficult to ascertain directly. For example, an accent might be a sign of the speaker's place or class of origin. Finally, culture has a local character. While generalizing in their nature, cultural systems vary across physical and social space, forming into more or less defined clusters.

We argue that these phenomena are intertwined and have co-evolved over time through the mechanisms of indirectly biased interaction and transmission. Specifically, assuming conditions under which it is (a) necessary to coordinate efforts of multiple actors to solve complex problems, (b) costly to maintain meaningful social connections necessary for cooperation, (c) disproportionately difficult for actors to ascertain possession of cultural traits in others directly and (d) possible for actors to direct their behavior based on the possession of observable markers, we hypothesize that over time such populations will form distinct cultural clusters and meaningful cultural signs will emerge.

To test this hypothesis we build an evolutionary agent-based model that implements the above conditions. In the model, agents each possess a selectively-neutral variant of a cultural trait representing functionally equivalent approaches to solving problems in an abstract domain, and a tag which represents its observable characteristics. The agents are periodically faced with hazards, which can only be averted by successfully coordinating their efforts with another agent. The agents cannot deduce the specific variants of the trait possessed by others, however they may periodically choose to abandon partners or find new ones.

Numerous studies have explored evolutionary dynamics of culture in an agent-based simulation environment [2] and many have approached this from the perspective of cooperation [3–6]. Similarly, the co-evolution of network structure and cooperation has also been explored [7]. Most of these studies use the Prisoner's Dilemma to mode the interplay between individual benefits of defection and group benefits of cooperation, although some have used other games, e.g. the ultimatum game [8]. Here we assume some form of cooperation is necessary for the agents' survival, but we also assume that cooperative behavior is only successful when agents match their strategies, all n of which are equally adept at solving the task at hand. The problem thus becomes an $n \times n$ coordination game with n Nash equilibria. Furthermore, in most previous studies agents' strategies are hidden to their interaction partners, and there have been several models where the connection between hidden traits and tags has been investigated [9–11]. It is common that the tags are "pre-fabricated" signs, in that agents either recognize them as indicators of group membership [9,11] or are able to learn a pre-existing relationship between the tag and another trait [10]. We take a different approach and attribute the tags randomly and observe whether any signifying quality emerges from the dynamics of the system. It has been hypothesized that indirectly biased transmission—the selection of traits affected by individuals' selection of models on the basis of unrelated attributes—is an important force in cultural evolution and in the emergence of symbolic culture [1]. The role of indirect bias and tags has been studied with mathematical models [1], even on small groups of live subjects [12]. Here we add to the larger

field of socioculture by analyzing its effects through an agent-based simulation on a large social network. Although we focus on the selection of collaborators rather than behavioral models, the indirect bias exerted on cultural transmission remains.

2 Methodology

We initialize agents on nodes of a random network. An agent can interact with others that connect to it via an edge. Each agent possesses one of 10 possible tags and one of 10 possible variants of a cultural trait. The tag and trait variables are categorical. While the tags are visible to the other agents, the trait variants are initially unknown to them. Moreover, each agent possesses a tag preference matrix. For each tag T an agents stores two values. The first, τ_T^+, represents the salience of positive experiences related to tag T, while τ_T^- represents the salience of negative experiences. This salience is expressed in form of a real-valued number from the interval $(-\infty, \infty)$. In each round every agent randomly selects one of its neighbors, and the two interact. The interaction consists of the two agents playing a coordination game: they compare their trait variants and if they match, they both receive one point towards their score; if the variants do not match, both agents deduct one point from their score. After each interaction the agents update their tag preference matrix. Assuming that the interaction was positive, the salience τ_T^+ for the neighbor's tag T is modified as follows:

$$\tau_T^+ = \ln\left[\sum_i^n t_i^{-d}\right] \approx \ln\left[\frac{1}{n-d} t_n^{-d}\right]$$

Here t_i is the time elapsed since the i-th relevant experience, n is the total number of such experiences and d is the rate of decay. An equivalent update applies to negative experiences. The quantity is very sensitive to recent experiences, while the importance of older experiences progressively decays with time. This representation is based on the ACT-R model of agent cognition [13]. Due to the computational infeasibility of the above relationship for large n, we implement a widely-used approximation [13]. In line with convention, we use $d = \frac{1}{2}$ [14]. Because we assume that maintaining social relationships is costly the agents pay a constant link maintenance cost c per round per link. After each round of interaction the agents can choose to cut ties with certain agents or create new ones. Each agent possesses a positive threshold value τ^+ and a negative threshold value τ^-. First it identifies any neighbors with tags whose negative salience exceeds τ^- and it cuts ties with them. It then searches the remaining population of agents and identifies all agents with tags whose positive salience exceeds τ^+. If there are any such agents, it randomly samples a subset of them and attempts to create a new link. The link will attach only if the preference is mutual and if the updated node degree does not exceed the average degree of the initial network. To prevent the network from reaching a degenerative state, we force the agents to maintain at least one connection at all times.

After a predetermined number of steps the population is evaluated and a subset of parents is selected. The parents reproduce to create offspring before they are removed from the population, so that only the offspring remain. Selection is based on agents' scores accumulated throughout the current generation. For each node two parents are selected via a tournament of size 3, chosen randomly from the set of the node's neighbors and the current occupant of the node. The parents recombine their phenotypes to create a single offspring via uniform crossover, i.e. each part of the phenotype is inherited from one of the parents with 50% probability. The inherited phenotype includes the tag, the trait variant, as well as the tag preference matrix and the salience thresholds. Mutation is applied to each part of the phenotype with 1% probability; random resetting is used for categorical variables, while Gaussian perturbations are used for real-valued variables. The population size is kept constant. Because the tag preference matrix is inherited in its final form, as it was molded throughout parents' lives, the model implements a Lamarckian version of evolution in which acquired characteristics are passed on to subsequent generations.

To control for the effects of biased interaction we introduce two other model configurations. In the second configuration, we eliminate the plasticity of the tag representation matrix. Thus, agents preferences for the individual tags are set at birth and they do not change throughout their lives as a result of interaction. This implementation is closer to a purely "genetic" model of bias.

Finally we devise a "baseline" configuration, one which does not include any tag-related bias. In this case the agents do not possess the tag preference matrix, they do, however, display some ability to maximize their interaction utility. The agents only remember their last interaction that took place. If that interaction was negative, the agent will cut ties with the partner in question. If the last interaction was positive, the agent, emboldened by its recent success, will create an additional connection to a randomly selected agent.

3 Results

Figure 1 shows the differences in evolutionary dynamics using the three different model configurations with a single choice of parameters. Here, we initialize a population of 1000 agents on a random network with average degree $\langle k \rangle = 8$, link maintenance cost $c = 0.16$ and 10 rounds of interaction per generation, for 100 generations. Figures 1(a), (b) illustrate the distribution of tags and trait variants over time. We observe that in the case with no bias the trait distribution is by far the most skewed, in fact drifting away towards a single variant. The tag distribution is fairly similar in all cases, remaining only moderately skewed. This is to be expected in the unbiased case, as there is no selective pressure exerted on the tags. In the biased configurations, this requires further investigation.

We then introduce a measure of tag "entropy". Based on Shannon's Entropy [15] it measures the (un)predictability of specific states. In this case we measure how well the tag "alphabet" encodes different trait variants. We write:

$$E_j = -\sum_i p_i \ln p_i \quad \text{and} \quad \overline{E} = \frac{1}{N}\sum_j \frac{E_j}{n_j}.$$

Here i iterates over the set of trait variants, p_i is the probability of encountering the i-th variant in the current tag, j iterates over the set of tags, E_j is the entropy of the j-th tag, N is the cardinality of the tag set and finally, we use the size of each tag sub-population, n_j, to normalize and obtain the average metric entropy, $\overline{E}$. Figure 1(c) shows tag entropy results for different model configurations. While the Lamarckian configuration is able to evolve into a state with low entropy, where each tag more or less faithfully encodes for a specific cultural trait variants, the unbiased configuration shows even lower values of entropy. This is surprising at first glance, as the tags are an afterthought in the unbiased model with no real significance. However, it is important to note, that this is a result of the drift in trait variants. It is then easy for every tag to encode a specific variant, as there is only a single one present in the population.

We thus further analyze the interplay of tags and cultural traits by measuring the modularity of the agent networks in terms of both attributes. Network modularity measures how neatly the network decomposes into communities defined on the possession of a shared attribute [16] and is defined as:

$$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w).$$

Here m is the edge count of the network, v, w are nodes, A refers to the adjacency matrix, k_v and k_w are the degrees of the nodes, c_v and c_w are the attributes of the nodes, and δ refers to the Kronecker delta function. Figures 1(d), (e) show that the the biased versions of the model produce communities that are well defined by both their tags and the trait variants that they share. Such communities do not appear in the unbiased configuration.

To better elucidate the situation we visually tracked the changes in the social networks of the agents over the course of the simulations. Figure 2 shows illustrative snapshots of the network evolution. We observe that in the beginning, the networks become very sparse in both cases and social interaction thus becomes minimal. This changes as the simulation progresses, and the differences between the two model versions becomes clear. In the unbiased model a single giant component emerges; meanwhile the Lamarckian bias creates a network organized into clear communities marked by possession of distinct tags and trait variants. Moreover, each tag locally strongly correlates with a single trait variant. The presence of clearly defined clusters suggests the small-worldness of these networks. To confirm this we compute the average clustering coefficients and the average path lengths of the networks. Watts and Strogatz [17] assign "small-world" properties to networks if they have significantly lower path lengths yet comparable clustering coefficients as a regular lattice of the same size and average degree as the target network. Figures 1(f), (g) show these measures, and we observe that while all three configurations show significantly shorter path lengths than regular lattices, the Lamarckian configuration demonstrates significantly

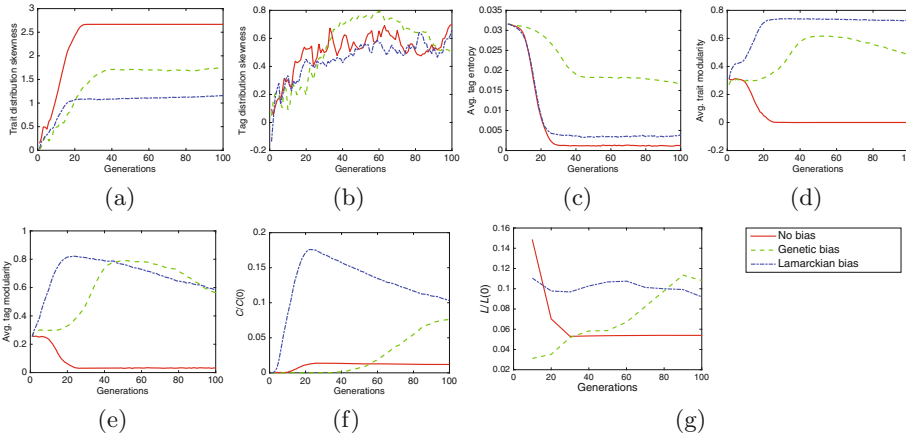


Fig. 1. Model statistics for $\langle k \rangle = 8, c = 0.16$ and $H/V = 10$.

higher clustering coefficients than the other two. This indicates a small-world quality of the networks under the Lamarckian model, while the other versions produce networks that are more random in nature.

To ascertain how sensitive these outcomes are to the choice of parameters we performed a partial parameter sweep. Namely, we tested the sensitivity with respect to the average degree $\langle k \rangle$ of the initial network, the link maintenance cost c, and the ratio between horizontal interaction frequency and vertical transmission frequency H/V. Figure 3 shows the dependence of some of the measured quantities on the cost and the ratio between interaction and transmission events frequency[1]. We observe that the Lamarckian model demonstrates higher clustering coefficient values than the other configurations across a major part of the parameter space, especially in cases with very low maintenance costs. Similarly, the Lamarckian model shows consistently shorter path lengths, while the other configurations display sensitivity to increasing the generation length (i.e. the genetic bias configuration) or increasing the maintenance cost (i.e. the unbiased model). Moreover, both of the biased versions display consistently higher rates of tag modularity than the unbiased configuration. Once again, this is especially pronounced in low maintenance cost regimes. A similar narrative applies to trait modularity, although it is worth noting, that in high maintenance cost regions the Lamarckian model now performs worse than the unbiased model.

4 Discussion

We have shown that bias can have a significant effect on social network structures in the long run, especially if bias acquired throughout agents' lifetimes

[1] We have not observed significant effects of changing the average degree, and therefore do not show figures including this parameter.

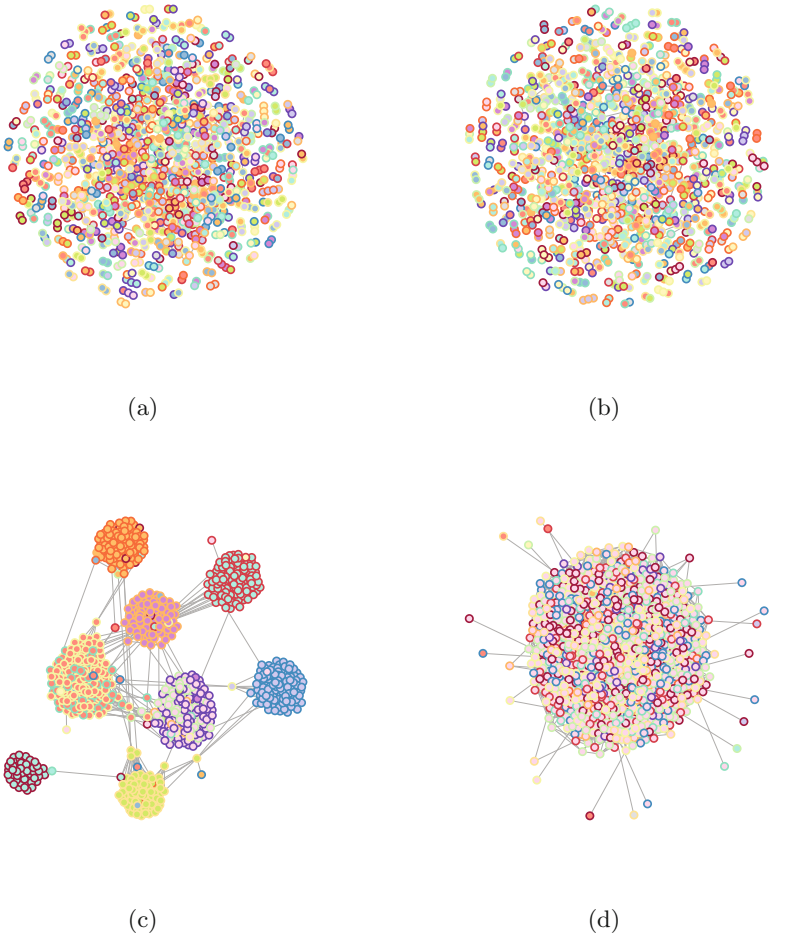


Fig. 2. Visualizations of the agent network in the Lamarckian bias (left) and un-biased (right) configurations after 5, 20 and 100 generations. The border color of a node determines its external marker, while the inner color determines its cultural trait variant. $\langle k \rangle = 8, c = 0.16, H/V = 10$. (Color figure online)

is inherited by their offspring and the cost of social tie maintenance is low. In the beginning, the system dynamics are similar in all of the observed cases. Because trait variants and tags are assigned randomly, and the placement of agents on nodes is also random, the probability of any agent having a successful interaction is fairly low at the onset. Tie maintenance is costly and therefore in the biased configurations evolution selects for individuals who possess strict preference thresholds, and thus can maximize their utility by minimizing the frequency of social interactions. The same outcome can be observed in the unbiased configuration where agents simply prune their neighborhoods after every failed interaction. This explains why we witness a dramatic decrease in

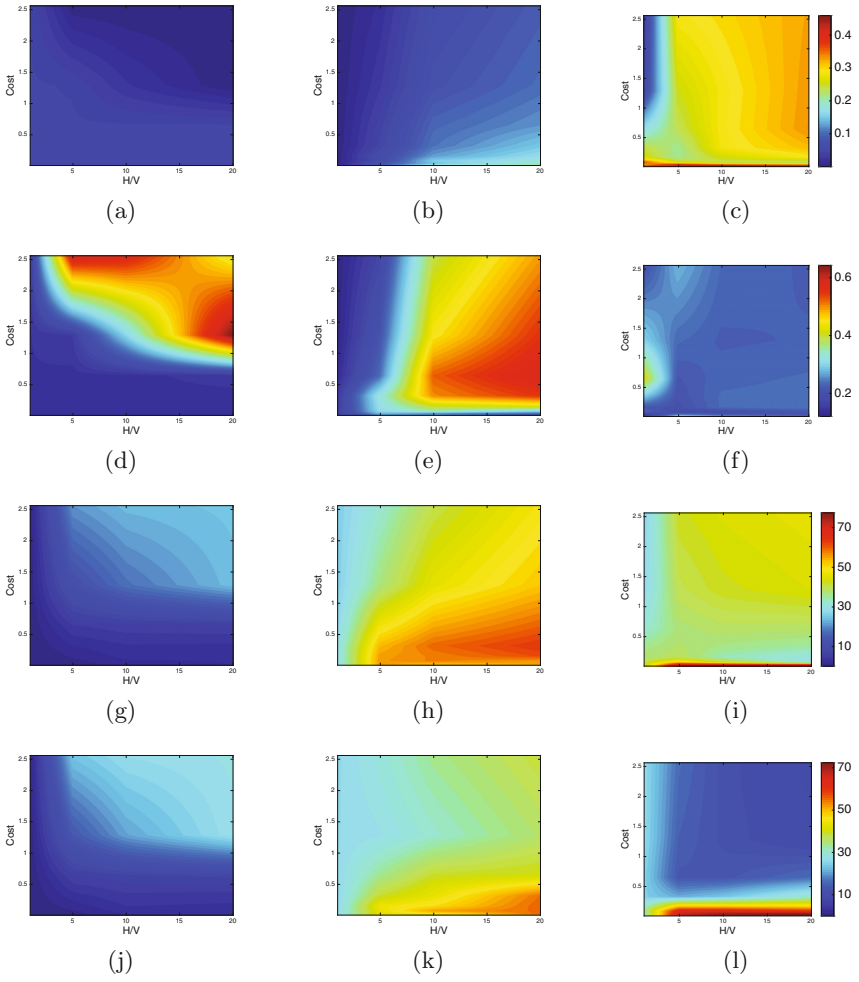


Fig. 3. Model statistics as a function of maintenance cost and interaction/transmission ratio for the unbiased (left), genetic bias (centre), and Lamarckian bias (right). Parts (a)–(c) show the average local clustering coefficients at the end of runs compared to regular lattices. Parts (d)–(f) show average path lengths at the end of runs compared to regular lattices. Parts (g)–(i) show the sum of tag modularity values at the end of each generation. Parts (j)–(l) show the sum of trait modularity values at the end of each generation. $\langle k \rangle = 32$.

network density in the beginning stages. After this phase, the dynamics diverge. In the biased versions, sub-populations of agents who are located on fairly small connected components of the network eventually converge on a single trait variant and a single tag. We argue that this is due to drift effects, which become exacerbated in small populations, and most importantly due to kin selection [18]. Because offspring are always born in their parents' vicinity, agents find

themselves in neighborhoods which are increasingly homogeneous with respect to agent attributes. If, by chance, the dominant phenotype possesses a high-scoring preference for its own tag, the dynamics are only reinforced. Furthermore, the integrity of the tag group is unharmed by growth, as new links are only made to the already established preferred tags. Because these processes are able to emerge in parallel within the many disjoint network components, this results in the formation of multiple clusters defined by a dominant tag-variant pair. Due to this effect of bias, the tags can become imbued by meaning, transforming into cultural signs representing possession of specific variants. It is worth noting that due to the local nature of network coalescence the meaning is also local: the same variant can be represented by different tags in different clusters, and different variants can be represented with identical tags in separate communities.

The situation is quite different in the unbiased configuration. The lack of distinct tag groups is self-explanatory, however, the absence of trait clustering deserves more attention. We argue that the main driver of the network structure in this case is drift. Without bias, agents attach to others randomly. If a certain trait variant owns a slight edge in frequency, agents possessing that variant will have a better chance of meeting their counterparts. As this trait group grows so does its advantage, attracting and infecting smaller components more easily.

We have also observed that the biased populations were able to evolve into more distinct clusters when the cost of tie maintenance was low. When costs were high the "small-worldness" of the biased networks has diminished and trait variant distributions displayed higher takeover rates. We argue that this is partly because as costs increase the effect of drift surpasses the effect of kin selection. As the maintenance cost becomes high relative to interaction payoffs, mistakes—nearly ubiquitous at the onset—become cruelly punished. This leads to a highly uniform distribution of fitness. Moreover, the high costs lead the agents to hold higher standards, and thus to keep smaller neighborhoods. Both of these facts contribute to higher rates of drift and lower rates of modularity and clustering.

The importance of cultural markers throughout has been documented throughout history. But cultural dimensions are likely important in modern networks of trust as well. Does culture factor in transactions between strangers on web services such as AirBnB or Craigslist? How does large-scale collaboration on open-source software projects emerge? We believe that our model could be applied to such real-world scenarios and validated with the help of empirical data.

While the simulations analyzed for this study have provided insight into the effect of biased interaction on the evolution of network and community structure there are still questions worthy of further investigation. It is natural to ask how exactly and to what degree kin selection affects the dynamics. It is therefore worthwhile to consider designing an experiment in which one would to isolate the dynamics of kin selection and analyze them precisely. Furthermore, we did not attempt to adjudicate the question whether the proposed strategies are evolutionarily viable and robust. In future research we will focus on simulating mixed populations, and identify conditions under which bias-driven strategies can permeate populations from random mutations and resist invasion.

References

1. Boyd, R., Richerson, P.J.: Culture and the Evolutionary Process. University of Chicago Press, Chicago (1985)
2. Bianchi, F., Squazzoni, F.: Agent-based modeling in sociology. WIREs Comput. Stat. **7**, 28306 (2015). doi:10.1002/wics.1356
3. Axelrod, R.: An Evolutionary approach to norms. Am. Polit. Sci. Rev. **80**(4), 1095–1111 (1986)
4. Miller, J.H.: The co-evolution of automata in the repeated prisoner's dilemma. J. Econ. Behav. Organ. **29**(1), 87–112 (1998)
5. Macy, M., Skvoretz, J.: The evolution of trust and cooperation between strangers: a computational model. Am. Sociol. Rev. **63**, 38–660 (1998)
6. Bowles, S., Gintis, H.: The evolution of strong reciprocity: cooperation in heterogeneous populations. Theor. Popul. Biol. **65**, 17–28 (2004)
7. Santos, F.C., Pacheco, J.M., Lenaerts, T.: Cooperation prevails when individuals adjust their social ties. PLoS Comput. Biol. **2**, e140 (2006)
8. Shutters, S.T.: Strong reciprocity, social structure, and the evolution of fair allocations in a simulated ultimatum game. Comput. Math. Organ. Theor. **15**(2), 64–77 (2009)
9. Hales, D.: Cooperation without memory or space: tags, groups and the prisoner's dilemma. In: Moss, S., Davidsson, P. (eds.) MABS 2000. LNCS, vol. 1979, pp. 157–166. Springer, Heidelberg (2000). doi:10.1007/3-540-44561-7_12
10. Janssen, M.: Evolution of cooperation when feedback to reputation scores is voluntary. J. Artif. Soc. Soc. Simulat. **9**(1), 17 (2005)
11. Hammond, R.A., Axelrod, R.: The evolution of ethnocentrism. J. Conflict. Resolut. **50**, 26–936 (2006)
12. Efferson, C., Lalive, R., Fehr, E.: The coevolution of cultural groups and ingroup favoritism. Science **321**, 1844–1849 (2008)
13. Anderson, J.R., Lebiere, C.: The Atomic Components of Thought. LEA Publishers, Mahwah (1998)
14. Petrov, A.A.: Computationally efficient approximation of the base-level learning equation in ACT-R. In: Fun, D., Del Missier, F., Stocco, A. (eds.) Proceedings of the Seventh International Cconference on Cognitive Modeling. Edizioni Goliardiche, Trieste, Italy (2006)
15. Shannon, C.E.: A mathematical theory of communication. AT&T Tech. J. **27**(3), 379423 (1948)
16. Newman, M.E.J.: Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**(23), 85778696 (2006)
17. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'Small-World' networks. Nature **393**, 440–442 (1998)
18. Hamilton, W.D.: The evolution of altruistic behavior. Am. Nat. **97**(896), 354356 (1963)

Simulating Population Behavior: Transportation Mode, Green Technology, and Climate Change

Nasrin Khansari[1(✉)], John B. Waldt[1], Barry G. Silverman[1], Willian W. Braham[2], Karen Shen[1], and Jae Min Lee[2]

[1] Electrical and Systems Engineering Department, University of Pennsylvania, Philadelphia, USA
{khansari,jbrooke,basil,shenk}@seas.upenn.edu
[2] School of Design, University of Pennsylvania, Philadelphia, USA
{brahamw,jaemlee}@design.upenn.edu

Abstract. This paper presents a decision tool intended to help achieve the goal of reduction in Green House Gas (GHG) emissions in the greater Philadelphia region by the year 2050. The goal is to explore and build a pre-prototype to evaluate the value of the role for agents, alternative data sources (Census, energy reports, surveys, etc.), GIS modeling, and various social science theories of human behavior. Section 2 explains our initial research on an Agent Based Model (ABM) built upon the Theory of Planned Behavior (TPB) and the Discrete Decision Choice model (DDC) to model consumer technology adoption. The users can utilize the proposed ABM to investigate the role of attitude, social networks, and economics upon consumer choice of vehicle and transportation mode. Finally, we conclude with results on agent decisions for which transit mode to use and whether to adopt greener technologies.

Keywords: Agent Based Models (ABM) · Decision-making process · Climate change · Energy use in transportation · Technology adoption

1 Introduction

The purpose of this research is to build a tool to help the greater Philadelphia region establish carbon emission reduction goals by 2050. Following a standard complex adaptive systems approach, we propose to research, design, construct, and validate an agent based model (ABM) as this tool. In an earlier paper, we discussed the types of policy issues such a tool should help decision makers to evaluate, and we return to that topic at the end of this article in the wrap up [1]. We posit that individual people and their micro-decision making are going to determine the macro-behaviors that emerge in terms of technology adoption and usage to impact the GHG problem. So this paper focuses on the agent model rather than policy choices.

ABM has emerged as a powerful analytical and computational method for studying complex adaptive systems and understanding of micro processes and their emergent consequences at the macro level. This new method offers a flexible architecture that allows for a detailed representation of complex agent systems, including the behavior

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 172–181, 2017.
DOI: 10.1007/978-3-319-60240-0_21

of agents, their social interactions and the physical and economic environments surrounding them. Agents represent discrete decision-makers as individual people and aggregates of individuals. These agents are simulated by autonomous entities with individual characteristics and independent internal decision making processes. Modeling this behavior provides better understanding and predicting of real world agents' decision making processes. In sum, ABM is an experimentation tool to study and demonstrate diffusion patterns resulting from simple decision rules followed by different agents in the system [2–8].

This paper describes our progress to date in prototyping and studying how ABM can work to accomplish the goals of the tool for supporting Delaware Valley Regional Planning Commission (DVRPC). We are just at the beginning of this research, and we are still experimenting with alternative ABM ideas and formulations, some of which we report here. We then present the baseline model of DVRPC transportation and forecast CO_2 emissions to 2050 in the case of business-as usual. This reveals how much CO_2 needs to be reduced to achieve 80% target. We then turn to some experiments with modeling of agents and present a couple of approaches we have investigated to date. Lastly, we conclude with lessons learned and ideas for the desired tool.

2 Theoretical Background

The Theory of Planned Behavior (TPB) states that "human behavior is the result of the intention to perform the behavior. In turn, the intention itself is driven by the individual's attitude toward the behavior, subjective norms, i.e., perceptions about social expectations and pressure, and perceived behavioral control (PBC), i.e., the individual's perception of her ability to actually perform the behavior. Thus, as a general rule, the more favorable the attitude and subjective norm, and the greater the perceived control, the stronger should be the person's intention to perform the behavior in question". In sum, considering control as an economic attribute, then human actions are led by three variables including attitudinal, social, and economic variables [9].

However, TPB is usually considered as a static model of behavior. Although TPB considers the effect of attitude, social norms, and control in intention, and in turn the effect of intention in actual behavior, it ignores the change of these factors over time and the related change of intention accordingly. Therefore, to compensate this ignorance of changes, we should consider evolving agent variables while integrating TPB in the ABM [5].

To investigate the process of individuals' behavior change over time, we use Dynamic Discrete Choice (DDC) model. In practice, DDC models are among the most sophisticated approaches for analyzing consumer choice. Benefiting from a time component, DDC considers intertemporal tradeoffs. DDC models can be utilized to study the impact of individual decision-making processes on system outcomes due to considering individual as the unit of analysis in these models [9].

In practice, social networks play an important role in leveraging individuals' awareness level. In fact, networks provide individuals with information about the state-of-the-art technologies including energy technologies, the cost and benefits of such

technologies and the tendencies to adapt new technologies and accordingly new pattern of behaviors. Previous studies show the importance of peer influences on individuals' behavior, e.g., adoption of hybrid-electric vehicles. In the TPB theory, decision making process is highly affected by economic aspects. However, along with economic aspects, the awareness regarding affordability or lack of affordability to adopt a technology is also highly important in the decision making process. In practice, payback can be considered as a key factor in adopting a new technology as does perceived adoption obstacles [5].

2.1 The Agent Based Model

Agent choices to change mode or adopt new technology are made by changing the option that scores highest in all three components. This is Eq. 1 which shows the three terms of the TPB model. In the current prototype, the attitude component (Eq. 2) is set based on a function of political party archetype and awareness level, the social component based on the percentage of users in the agent's network (Eq. 3), and the economic factor based on the upfront costs, payback period, and obstacles to adoption (Eqs. 4 and 5). In our model; we have four groups including active, aware (sympathetic towards the environmental movement), unaware, and negative aware. There are negative information centers and information centers on the agent landscape that influence the nearby agents who might happen to come in contact. Agents chose whether to adopt green technologies based on an economic factor and how many of its neighbors have adopted, and overall climate attitude. All of these hold equal weight in the decision to adopt.

To illustrate, we simulate the adoption rates of four vehicle types (VT) including a hybrid-electric vehicle (HEV), a plug-in hybrid-electric vehicle (PHEV), a battery electric vehicle (BEV), and a conventional vehicle (CV). In our model, agents are created based on the number of households in the census tract. Households are given an education level, income, political affiliation, and number of vehicles owned based on census tract data. What follows shows the model for vehicle adoption.

Equation 1 is used to compute the utility of or the intention for each vehicle type. For three type of vehicles (HEV, PHEV and BEV), attitude, social and economic factors are evaluated using Eqs. 2–4. However for CVs, attitude and economic factors are extracted using Eqs. 6 and 7.

In Eq. 4, ticks represents the duration time of simulation (in years) and this term causes obstacle impacts to be reduced over time. Also in this equation, α_{VT} is different for each vehicle type to reflect alternative rates of improvement of the technologies. This value is 1.3, 1.32, and 1.35 for HVs, PHEVs and BEVs, respectively. For these three vehicle types, Eq. 5 evaluates the payback where β_{VT} is considered as 12 for HEVs and 16 for PHEVs and BEVs to reflect differences in upfront cost for charger and battery.

$$Intention_{VT/MC} = \frac{1}{3}(Economic\_Factor_{VT/MC} + Social\_Factor_{VT/MC} + Attitude\_Factor_{VT/MC}) \tag{1}$$

$$Attitude_{VT/MC} = \frac{1}{2}(Political\ Party\ +\ Awareness\ Level) \tag{2}$$

$$Social_{VT} = \frac{Number\ of\ VT/MC\ in\ network}{Number\ of\ Vehicles/MC\ in\ Network} \tag{3}$$

$$\begin{aligned} Econimc_{VT} = (0.17 * Financial\ Factor + 0.5 * Payback\ Period + 0.33 \\ * (100 - Obstacles\ To\ Adoption + ticks^{\alpha_{VT}})) \end{aligned} \tag{4}$$

$$Payback_{VT} = \left(-2 * gas.price + \beta_{VT} + Unif(0, 4)\right) \tag{5}$$

The literature indicates that liberals favor green technologies [10]. Census data shows political affiliations which are captured in our census tract agents. Political party score is set to Republican (25), Independent (50), and Democrat (75). This causes liberals to care about VTs (HVs, PHEVs and BEVs) in Eq. 2. We also utilize Eq. 6 to reflect lower attractiveness of CVs to liberals relative to other VTs.

$$Attitude_{CV} = \frac{1}{2}((100 - Political\ Party) + Awareness\ Level) \tag{6}$$

Political Rating gives a positive affinity for CVs and obstacles to adoption and payback period both are zero for CVs. Also the tick is omitted from Eq. 4 for CVs. The obstacles to adoption value simulate common reasons a household would avoid buying a different type of vehicle. For example, "range anxiety" and lack of charging infrastructure available are two common reasons listed for not buying BEVs or PHEVs. The tick term in Eq. 4 tends to reduce these obstacles as the technology matures.

Price strategy also affects mode choice (MC). In practice, short-run fluctuations in gas price may lead to temporary changes in driving behavior (e.g. traveling at more fuel-economical speeds, avoiding rapid accelerations and breaking, making fewer trips or switching to other modes including public transportation, walking, and bicycling). However, consumers might return to old driving patterns when gas prices return to their previous level. On the other hand, long-run changes in gas price have more permanent effects on vehicle miles travelled and gas consumption (e.g. via buying a more efficient car, switching to an alternative fuel or hybrid vehicle, or increase of the tendency of living close to work places). Studies show that higher gas prices decreases GHG emissions from vehicles, improves air quality with benefits for health, and reduces congestion, with benefits for the economy [11].

Equation 1 is used to compute the utility of or the intention for each mode choice including vehicle (V), bike, walk, or public transportation. For each VT mode choice (MC), attitude and social factors are evaluated using Eqs. 2 and 3. For three other types of mode choice (walk, bike, and public transit), economic factors are extracted using Eq. 7. For these three MCs, β is considered as 30 for walk and 27 for bike and 31 for public transit using Eq. 7. For all vehicle types, economic factor is extracted using Eq. 8.

$$Econimc_{MC} = 1.5 * (\beta - distance\ to\ work) + Unif(0, 2.5) \tag{7}$$

$$Econimc_{VT(MC)} = distance\,to\,work + 10 * ((1.5 + (2.5 - (0.5 * gas.price)))) \quad (8)$$

3 ABM Model Experiments

In this section we experiment with ways to extend the baseline model so as to add the agents with the aforementioned preferences for alternative transportation modes and environment quality. A guiding principle in the design of agents is to Keep It Simple, Stupid (KISS). Following KISS, we will look at a few discrete agent differences, mostly linear approximations, and very few choices. At this stage, we are just testing ideas and approaches and can always make things more sophisticated later if we find ideas that are useful.

3.1 Transportation Behavior Model

A map of the DVRPC region is created and divided into census tracts. As shown in Fig. 1, for the current model, initially a map of the DVRPC is split into 5 zones based on the population density data. Zones consist of census tracts with population (1,000s) per acre density; Zone 1 (0–20), Zone 2 (20–40), Zone 3 (40–60), Zone 4 (60–80), and Zone 5 (80+). From these intervals, 1072 tracts are in Zone 1(Purple), 194 are included in Zone 2 (Light Grey), 77 reside in Zone 3 (Yellow), 20 include in Zone 4 (Orange), and finally 6 reside in Zone 5 (Red). Data is not available for the regions shown in black in this figure.

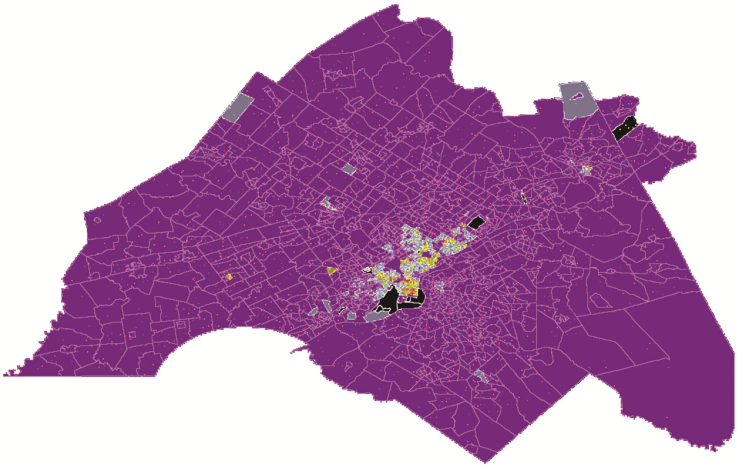


Fig. 1. Zones in DVRPC (Color figure online)

As part of this research on agents, we are investigating different archetypes. We have three different political affiliations. Education levels are less than high school, high school graduate, some college or associate's degree, and bachelor's degree or higher. Income are put in five groups including under $25,000, $25,000 – $49,999,

$50,000 – $74,999, $75,000 – $99,999, and over $100,000. People hold varying views and values about how they personally, and also their community, should manage sustainability tradeoffs. This has been widely researched in the climate change and sustainable development literature. As an example, a useful framework for considering people's values is described in de Vries et al. [12] using the two dimensions (or axes) of globalizing versus regionalizing and private/material/market versus public/immaterial/government. The four quadrants have come to be known as: free market-oriented globalizing world (Global Market); the market-oriented protectionist world (Fortress); the government oriented globalizing world (Global Solidarity); and community-oriented regions (Caring Region). We also classify the agents along other factor sets such as, among others, level of being informed (important for informational campaigns) and willingness to change and adopt new technology, products, and behaviors: e.g. see [13].

Agents are created based on the number of total commuters within a census tract. For every 333 commuters, one agent is created due to limitation of personal computer such as memory and CPU. This agent is given an income level, social awareness level, transit mode, and commute time all based on relevant census tract information. Income level and commute time are set according to the median values for that tract. While transit mode is based off the percentage of each mode in that census tract. For example, if the census tract contains 60% public transit commuters, this agent has a 60% chance of choosing public transit as his initial mode of choice. Social awareness level is based on [14] so that the overall number of each type is in line with these numbers. The agent is also given a social network based on a preferential network either within the same social awareness level or solely with others outside of his awareness group. This social network is used to influence decisions on adoption rates and transit mode choices.

Change of Vehicle. Strategies to promote adoption of BEVs, HEVs and PHEVs reduce air emission and oil dependency impacts from passenger vehicles. Initial vehicle types are set by the current market share of each relevant vehicle type. Each vehicle type is given an up-front cost, payback period, and obstacles to adoption value. Up-front cost is based on the average cost of each vehicle type currently on the market. The length of payback period is created by taking into account any additional costs to adopt, such as installing a home charging unit, and the current gas price, which is an input to the model. As the simulation progresses in years (ticks), the obstacles to adoption are decreased, as to simulate innovation and improvements in each technology that will occur over time (see Eq. 4).

Mode Choice. To help DVRPC to exceed the goal of 80% reduction in the emission of GHG by 2050, it's vital for people to shift their mode from personal vehicle to walk, bike, or public transit. The decision criterion for transit mode is a function of distance to work, social awareness level, the transit mode of others within that agent's social network, and gas price. The simulation runs for 36 years, during which agents make decisions for which transit mode to use, and whether to adopt greener technologies. Adoption rate is modeled using Eq. 1.

3.2 Results Analysis and Discussion

As gas price is increased, the number of drivers decreases drastically (see Fig. 2a). The number of walkers, transit riders, and bikers all increase at different rates when gas price is increased (Figs. 2b-d). An increase of gas price from $3 to $5 results in about 150,000 less people using a car as their primary commuting mode. These 150,000 people convert mostly to transit ridership or walkers. This is 2.39% of DVRPC population (6,261,673 people).

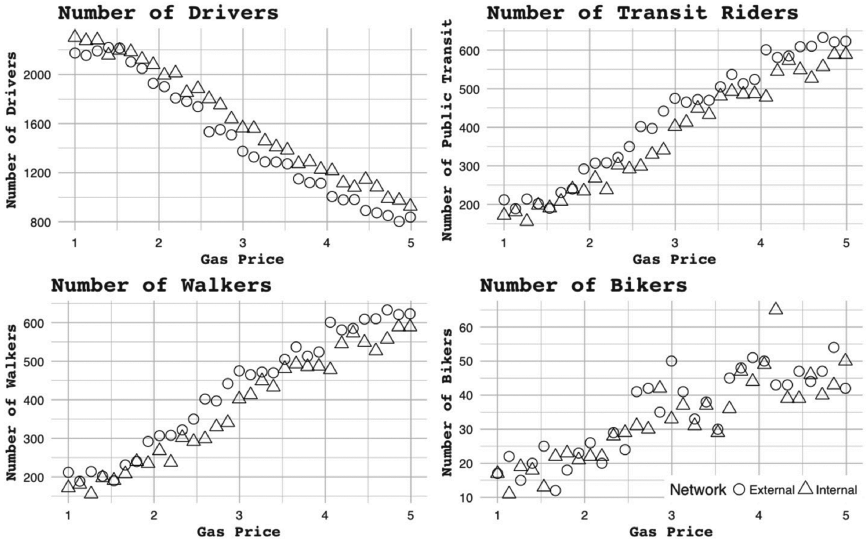


Fig. 2. The roles of social network and gas price in transportation mode

The effects of the different social networks can be seen as the two sets of projections in each plot. There are 2 projections in each plot of Fig. 2 – one for each type of network that is influencing the agents (internal vs external). When a person with a low level of social awareness is surrounded by a network of people who are more socially aware than themselves, then the person is more likely to make environmentally conscious decision. For instance, there are more people walking and less people driving when there is an external network type. This outcome can only happen when using the external network type. Networks composed of agents with similar social awareness levels only reinforce the behaviors to which they are already predisposed.

Using the theory of planned behavior, each household will make decisions on which vehicle type to buy. Equation 1 leads the agents to compute an intention value for each vehicle type, purchasing the one with the largest intention. The decision to

replace a vehicle varies according to a triangular distribution over 3–11 years. Therefore, each household will not decide on a new vehicle until their current vehicle is adequately used.

Figure 3 shows, the crossover point where CV is no longer the dominant vehicle type occurs around year 23, which corresponds to the year 2037. Initially, HEV is the main competition as the obstacles to adopt HEV right now are relatively low, and since the market share is currently highest. Thus, the social component is the strongest among all the other choices. However, as the simulation progresses both PHEV and BEV overtake HEV. This is mainly due to the simulated decrease in the obstacles to adoption. As the range of electric car batteries continues to improve and charging infrastructure is constructed, more households will gravitate towards PHEV or BEV. Assuming this technological innovation for PHEV and BEV continue, CV sees a significant drop in market share by 2050.

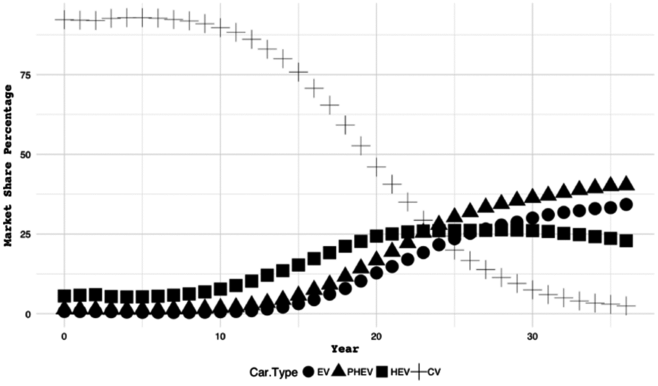


Fig. 3. Market share of vehicle types

Based on these market share results, we can now estimate the GHG impact of agent mode choice and adoption decisions. The GHGs are estimated from three VTs (CV, HEV, and PHEV). The GHG or CO_2 emissions are calculated using Eq. 9 (Where VT_1, VT_2, and VT_3, represents CV, HEV, and PHEV respectively). Recalling that, each agent represents 333 households. Figure 4 shows the outcome. The horizontal line across the top is DVRPC's 2010 estimate of GHGs from vehicles. The predicted line shows a reduction of just over 50% by 2050.

$$CO_2 \text{ Emissions } = \sum_{i=1}^{n=3} \text{Number of } VT_i * \text{Average Annual } VT_i \text{ Emissions} * 333 \quad (9)$$

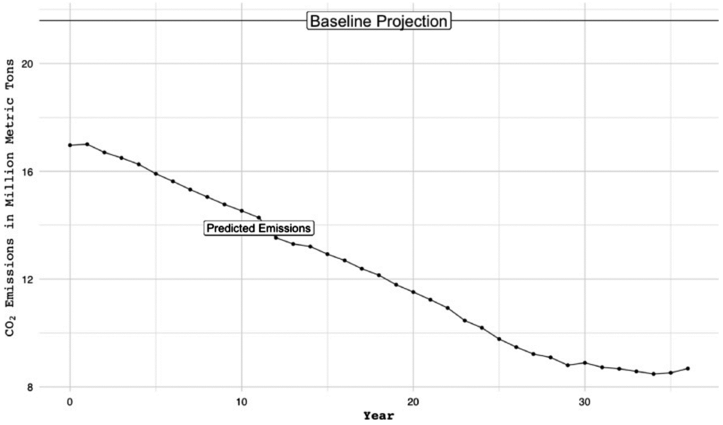


Fig. 4. Predicted CO_2 vehicle Emission vs Baseline

4 Conclusion

As systems engineers, it is tempting to think that as better technologies become available for conserving energy and reducing GHGs, they will naturally lead to a better outcome. In Sect. 1 we reviewed a number of innovations for the transportation sector producing GHGs. But cities and regions are complex sociotechnical systems filled with people with divergent objectives.

Using a flat line forecast from DVPRC's 2010 Vehicle Emissions Report, we expect to be pumping 21.6 million metric tons of CO_2/year into the atmosphere from vehicles alone in the business as usual scenario. Using TBP and DDC discussed in Sect. 2, we have begun the effort to look how people will react to large social challenges. Section 3 then delved into the equations used to model the adopters of these innovations. The goal is to incorporate such theories within an ABM that can in turn help policy-makers seeking to bring new technologies and approaches into the marketplace.

In Sect. 4, we applied our agent model for transportation system to see if CO_2 emissions decrease relative to the baseline. We simulated our model and presented the simulation results. In transportation sector, the 51.17% CO_2 reduction that actually occurred relative to the baseline.

Pushing gas prices higher than $5/gallon and providing education and economic incentives for alternative transportation options appears to be needed to get closer to the 80 by 50 goal. Our future research will explore these issues further. Future work will also focus on the residential and commercial building sector to model households' energy behavior aiming at reducing CO_2 emissions through applying strategies such as moving towards green buildings, smart grid, and renewable energies. Finally, there were innumerable assumptions, significations, and guesstimates that needed to be utilized to get this prototype built. We need to go back and ground our equations more thoroughly by conducting surveys and more fully utilizing available data sets.

Acknowledgements. We thank Kleinman Center for Energy Policy, the Mellon Foundation: Humanities, Urbanism, and Design Initiative, and the Delaware Valley Regional Planning Commission for supporting us in this research. Any opinions or errors are those of the authors alone.

References

1. Khansari, N., et al.: An agent-based decision tool to explore urban climate & smart city possibilities. In: 11th Annual IEEE International Systems Conference (SysCon) 2017 (2017)
2. Moss, S., Pahl-Wostl, C., Downing, T.: Agent-based integrated assessment modelling: the example of climate change. Integr. Assess. **2**(1), 17–30 (2001)
3. Robinson, S.A., Rai, V.: Determinants of spatio-temporal patterns of energy technology adoption: an agent-based modeling approach. Appl. Energy **151**, 273–284 (2015)
4. Sopha, B.M., Klöckner, C.A., Hertwich, E.G.: Adoption and diffusion of heating systems in Norway: coupling agent-based modeling with empirical research. Environ. Innovation Societal Transitions **8**, 42–61 (2013)
5. Rai, V., Henry, A.D.: Agent-based modelling of consumer energy choices. Nat. Clim. Change **6**(6), 556–562 (2016)
6. Ma, T., Nakamori, Y.: Modeling technological change in energy systems–from optimization to agent-based modeling. Energy **34**(7), 873–879 (2009)
7. Roozmand, O., Ghasem-Aghaee, N., Hofstede, G.J., Nematbakhsh, M.A., Baraani, A., Verwaart, T.: Agent-based modeling of consumer decision making process based on power distance and personality. Knowl.-Based Syst. **24**(7), 1075–1095 (2011)
8. Silverman, B.G., Hanrahan, N., Bharathy, G., Gordon, K., Johnson, D.: A systems approach to healthcare: agent-based modeling, community mental health, and population well-being. Artif. Intell. Med. **63**(2), 61–71 (2015)
9. Rai, V., Robinson, S.A.: Agent-based modeling of energy technology adoption: empirical integration of social, behavioral, economic, and environmental factors. Environ. Model Softw. **70**, 163–177 (2015)
10. McCright, A.M., Dunlap, R.E., Xiao, C.: Perceived scientific agreement and support for government action on climate change in the USA. Clim. Change **119**(2), 511–518 (2013)
11. Circella, G., Handy, S., Boarnet, M.: Impacts of Gas Price on Passenger Vehicle Use and Greenhouse Gas Emissions, 30 September 2014. http://arb.ca.gov/cc/sb375/policies/policies.htm
12. De Vries, B.J., Petersen, A.C.: Conceptualizing sustainable development: an assessment methodology connecting values, knowledge, worldviews and scenarios. Ecol. Econ. **68**(4), 1006–1019 (2009)
13. Khansari, N., Vesaghi, A., Mansouri, M., Mostashari, A.: The multiagent analysis of social progress in energy behavior: the system dynamics methodology. IEEE Syst. J. **PP**(99), 1–10 (2015)
14. Dunlap, R.E.: At 40, environmental movement endures, with less consensus. Gallup Poll. (2010)

A Parametric Study of Opinion Progression in a Divided Society

Farshad Salimi Naneh Karan and Subhadeep Chakraborty(✉)

Department of Mechanical, Aerospace, and Biomedical Engineering,
University of Tennessee, Knoxville, TN, USA
schakrab@utk.edu

Abstract. In this paper, a probabilistic finite state automaton framework is used to model the temporal evolution of opinions of individuals in an ideologically divided society in the presence of social interactions and influencers. In such a society, even quantifiable and verifiable facts are not unqualified absolute but are only viewed through the prism of the individual's biases which are almost always strongly aligned with one of the few prominent actors' viewpoint. The gradual progression of divisiveness and clustering of opinion or formation of consensus in a scale free network is studied within the framework of bounded-confidence interaction between nodes. Monte Carlo simulations were conducted to study the effect of different model parameters, such as the initial distribution of opinion, confidence bound, etc. in the behavior of the society. We have shown that in absence of influencers, government policies are the important factors in the final distribution of the society unless a specific group has higher number of members initially. Also, even very small groups of influencers proved to be highly effective in changing the dynamics.

Keywords: Decision making · Opinion dynamics · Influence

1 Introduction

The field of social psychology has a rich history of studies in social influence and group behavior. Such work generally investigates a number of disparate motives and contextual factors to explain individual-level and group behavior, traditionally in small-scale laboratory settings, but increasingly motivated by data mined from online resources, such as social media. In fact, social media, with its always-on persistence and open infrastructural base has quickly developed into a platform for news-storytelling, collaborative filtering and curating of news [1]. Twitter has enabled non-elites to emerge as gatekeepers of information within networked, crowdsourced environments, but at the same time, has provided a forum for conflict, confrontation and polarization. This has never been so relevant as now and particularly in the political scene in the US in the pre-and post-time period of the 2016 presidential election.

While the extremity of the current political rhetoric may feel unprecedented, confrontation has always focused on the singular goal of winning over the other

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 182–192, 2017.
DOI: 10.1007/978-3-319-60240-0_22

side. As an entire population becomes consumed by this mindset, we reenact partisan patterns of conflict that may comfort our fears, promote strong clustering between like-minded people, but undermine cooperation across society and the chance of ever achieving unity and consensus.

Studies suggest that the operation of homophily - the tendency to follow like-minded individuals and to shun those with opposing opinions -is strongly prevalent in social media applications [2]. Furthermore, shared geo-locality and communal bonds are strengthened via Twitter posts, permitting forms of "peripheral awareness and ambient community" [3]. To model these findings, several nonlinear interaction models among individuals, have been studied which illustrate polarized decision, the self-organization of behavioral conventions, and the transition from individual to mass behavior. In one such study by Shutters [4], the cultural polarization phenomenon has been studied on different network structures in the presence of extremists in the framework of bounded confidence. In this study, the change of opinion after each interaction is guaranteed for non-extremists, rendering the imitation and simulation of human decision making rather unrealistic.

This paper is an attempt to construct a parametric study that can address at least a portion of these ideas within a mathematically tractable agent-based modeling (ABM) framework. The nodes in an extended Barabási-Albert (BA-extended) network are modeled as rational actors, their choices and logical mechanism are related to the historically dominant expected utility family of theories made popular by von Neumann and Morgenstern [5]. The heart of the theory, sometimes called the rational expectations principle [6], proposes that each alternative course of action or choice should be evaluated by weighting its global expected satisfaction-dissatisfaction with the probabilities that the component consequences will occur and be experienced. The Probabilistic Finite State Automata (PFSA) based discrete choice model, proposed and studied in [7] has been modified with one key difference.

The scenario under discussion and this key difference are explained in the next sections. First, the PFSA framework and its assumptions are briefly explained. Next, we discuss how the Bounded Confidence interaction model is applied in the simulations. Then the results of simulations for different scenarios are presented and discussed. In the end, all findings of this paper are summarized and concluded.

2 Simulation Setup

2.1 Probabilistic Finite State Automata

In this paper, a Probabilistic Finite State Automata (PFSA) framework is used to represent the decision making routine. In this framework, it is assumed that every agent has the same *finite set of discrete choices (or states)* at each time instant. Also, we assume that agents subscribe to the normative perspective of the group they belong to; social norms of groups can differ from each other, but inside a group the same social norm is shared with everybody. Moreover, it is assumed that, even

Table 1. List of PFSA states and events

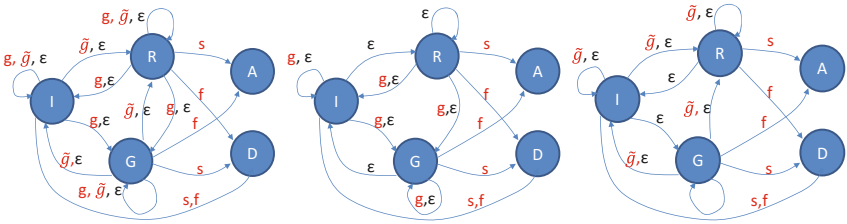
States	Description	Events	Description
I	State of being undecided/neutral	g	A popular act by the government
R	State of supporting the Revolution	$\tilde{g}$	An unpopular government act
G	State of supporting the Government	ε	An internal decision
A	State of political advantage	s	Success of the revolution
D	State of political disadvantage	f	Failure of the revolution

when presented with the exact same choices with the exact same pay-offs, different individuals, and possibly even the same individual, may probabilistically make alternate decisions. Furthermore, it is assumed that there are two types of events, *external/global* and *internal*; when a global event happens all the population is affected, however, an internal event is similar to an individual's personal choice. A complete description of the mentioned assumptions can be found in [7].

The assumption of *normative perspective* allows rational behavior, to be encoded as a PFSA. In our simplified depiction of the situation, each individual faces the *internal* decision of supporting the existing government, supporting the rebelling group, or remaining in a state of indecision. Additionally, the individual can reach a state of political advantage, or disadvantage, but the uncontrollable transition to these two states can only occur through an *external* event, namely, the success or failure of the revolution. The five PFSA states and events are described in Table 1.

Assumption. Different normative perspective for different groups In this study, it is assumed that the society is divided in three groups: the Independent group, the group with preexisting bias towards Revolution (R-Leaning), and the group biased towards the government (G-Leaning). The way each group perceives the actions of the government is different. Members of the Independent group perceive the popular actions as good and the unpopular actions as bad. However, members of the R-Leaning group perceive all actions of the government as bad. On the other hand, members of the G-Leaning group perceive all actions of the government as good. This assumption results in having three different normative perspectives in the population, Fig. 1.

Figure 1a gives a schematic of the assumed normative perspective encoded as a PFSA for the Independent group. It may be noticed that transitions such as $g : G \to R$ or $g : I \to R$ are unauthorized, since it is assumed that a favorable act by the government should not make anyone decide to join the opposing group. Also, the same event can cause alternate transitions from the same state; the actual transition will depend probabilistically on the measure of attractiveness of the possible target states. In Fig. 1b, all actions of the government are clubbed together and recognized as good for the G-Leaning group. The elimination of some transitions between states in Figs. 1b and c is rooted in the biased views that the G-Leaning and R-Leaning groups have. For example, the only way



(a) Independent (b) G-Leaning (c) R-Leaning

Fig. 1. Normative perspective of different groups

the transition $G \rightarrow R$ can be authorized is when an unpopular action by the government, $\tilde{g}$, happens. Since unpopular actions are perceived as good by the G-Leaning group, $G \rightarrow R$ can never happen for this group and must be eliminated.

In the PFSA framework, the probability of transitioning to a different state is dependent on the reachability of that state from the current state, the current event (external or internal), and also the relative degree of attractiveness of the target state. The state attractiveness measure is calculated using the concept of positive real measure attributed to a string of events [8]. It depends on the reward from each state (χ), the state transition matrix (Π), and the distribution of states ($\bar{v}_i$). A real measure ν_θ^i for state i is defined as

$$\nu_\theta^i = \sum_{\tau=0}^{\infty} \theta \, (1-\theta)^\tau \, \bar{v}^i \Pi^\tau \bar{\chi} \tag{1}$$

where $\theta \in (0, 1]$ is a user-specified parameter. Mathematical structure of the mentioned parameters are available in [7].

It should be noted at this point that the premise for these assumptions are the authors' hypotheses and conjectures based on observations from the political sphere, but these are as of now unsubstantiated by studies or data analytics. The basic observation is that in response to the SAME political event two groups of people respond in diametrically opposite manner (strongly in favor, or strongly against) - each group with the same passion and conviction that they are correct. Thus, although within their own logical construct, the two groups are not dissimilar, but the perturbation that drives the two groups manifest itself in two completely different ways - to the G-leaning group each action by the government seems perfect, while to the R-leaning group, the same actions are detestable.

The other interesting dynamic in these composite groups is the possibility of gradual drift of opinion due to interaction and inter and intra group communications. It seems that the KH bounded confidence model of interactions is appropriate since this model predicates that two nodes only communicate if they are not too dissimilar. This is described next.

2.2 Bounded Confidence Dynamic

In the Krause-Hegselmann (KH) bounded confidence model, an agent is chosen at random; then, the agent interacts with its compatible neighbors. Compatibility between two nodes is determined by the distance between the current opinions held by the two nodes. The procedure is repeated by selecting another agent randomly and so on. The type of final configuration reached by the system depends on the value of the confidence bound d. In this paper, it is assumed that the interaction is entirely through the characteristic function χ of the states. The update rule for the reward vector of agent i, due to interactions with its neighbors is as follows:

$$\bar{\chi}^i_{t+1} = \bar{\chi}^i_t + \mu.(\bar{\chi}^i_{t_{neighbors}} - \bar{\chi}^i_t) \tag{2}$$

where $\bar{\chi}^i_{t_{neighbors}}$ is the *mean reward vector* of the first-order neighbors in the network of agent i at time step t. The averaging process is used to combat the minor fluctuations in the local reward vector [9]. Here μ (or the convergence parameter) is the weight which determines how much an agent is influenced by the other one.

Since many networks in the real-world are conjectured to be scale-free, including the World Wide Web, biological networks, and social networks, in this study, a BA extended model network created by the Pajek software program is used [10]. Table 2 presents the parameters of the network.

Table 2. List of parameters used for BA scale-free network

Number of vertices	100
Number of initial disconnected nodes	3
Number of added/rewired edges at a time	2
Probability to add new lines	0.3333
Probability to rewire edges	0.33335

Influence Model: The influencers are treated as indistinguishable except for the fact that they never update or change their $\bar{\chi}$ values; moreover, they do not make decisions, and stay in the same state of mind during the entire simulation. Also, it is typical that the influencers are serving a certain agenda, in this case, trying to mobilize forces to join the Revolution. But, this influence is exerted very passively, by advertising a higher value for $\chi(R)$ and lower value for all other states.

$$\chi^I(q_j) = \begin{cases} \chi_m(q_j) - \Delta & \text{if } j = 1, 3, 4, 5, \\ \chi_m(q_j) + \Delta & \text{if } j = 2. \end{cases} \tag{3}$$

in which $\chi^I(q_j)$ represents the reward associated with state q_j for influence nodes, and $\bar{\chi}_m$ is an estimate of the reward values expressed by the whole society on an average. Δ is a parameter adjusting the strength of influencers (control input).

Simulation Process: A population of 100 people are divided in three groups with specific ratios (G_i, R_i, I_i where $G_i + R_i + I_i = 1$). Each group is initialized and given the respective normative perspective. All agents are assigned a random number drawn from a uniform distribution $U(0, 1)$, representing the time remaining before that person makes a decision. This imposes an ordering on the list of people in the network. As soon as someone makes a decision, the time to his next decision, drawn from $U(0, 1)$, is assigned and the list is updated. Additionally, external events g and $\tilde{g}$ are also associated with a random time drawn from $U(a, b)$.

At the time epoch t_k, when it is the i^{th} person's turn to make a decision, he updates his personal estimate of the reward vector according to the update equation (Eq. 2). He then calculates the degree of attractiveness of the states based on the normalized measure, using Eq. 1. The only difference in the case of an external event such as $g, \tilde{g}, s$ or f is that everyone simultaneously updates their states rather than asynchronously, as in the case of internal events. Each simulation is run 50 times and the average of all the runs is analyzed.

3 Simulation Scenarios and Results

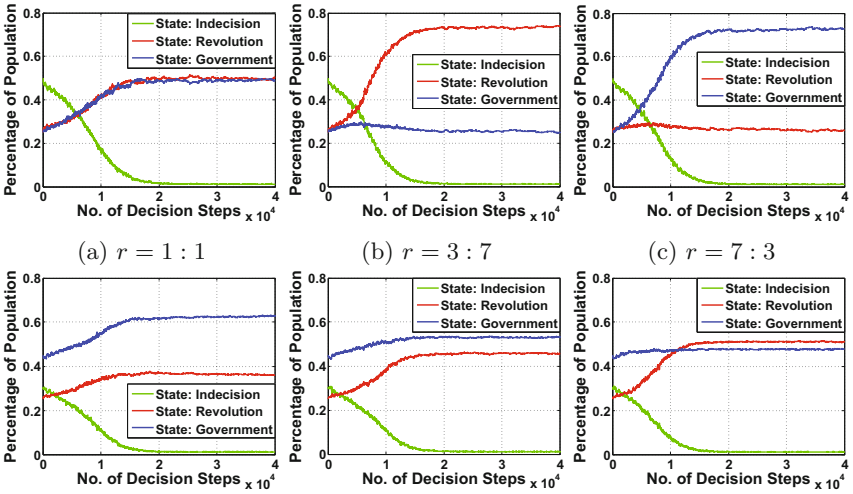
3.1 Effect of Global Events

This experiment studies the effect of the ratio of good to bad external events, or equivalently $r = \frac{P(g)}{P(\tilde{g})}$, on the opinion dynamic of the population. First, we consider equal initial distributions for G-Leaning and R-Leaning groups ($G_i = R_i = 0.25$). The first row of Fig. 2 provides the results of this type of initial distribution. In all three cases, as soon as events happen, members of the Independent group change their opinions.

In Fig. 2a, the same probability of good and bad external events causes the Independent group to equally divide between the G-Leaning and R-Leaning groups. However, in Fig. 2b, because of the higher probability of bad events, people from the Independent group lean towards the Revolution group. The exact same reasoning can be used for Fig. 2c where good actions by the government outnumber the bad ones causing people from the Independent group to support the Government. It can be concluded that in the absence of influencers, for equal initial distribution, r is the determining factor of the state distribution of the population at equilibrium.

In a second set of experiments, we consider unequal initial distributions for G-Leaning and R-Leaning groups ($G_i = 0.45, R_i = 0.25$). Similar to the previous experiment, when equal number of good and bad events happen, the Independent group leans towards either groups somewhat equally, Fig. 2d.

In Fig. 2e, when bad events slightly outnumber the good events, the Revolution state starts to increase but it is not the dominant opinion of the population as it was in Fig. 2b. The reason is that a high percentage of the population is initially in the G-Leaning group. As a result, Independent members have a higher chance of interacting with G-Leaning members in their network, and



(a) $r = 1:1$ (b) $r = 3:7$ (c) $r = 7:3$

(d) $r = 1:1$ (e) $r = 4:6$ (f) $r = 3:7$

Fig. 2. Effect of external events on population opinion without influence group

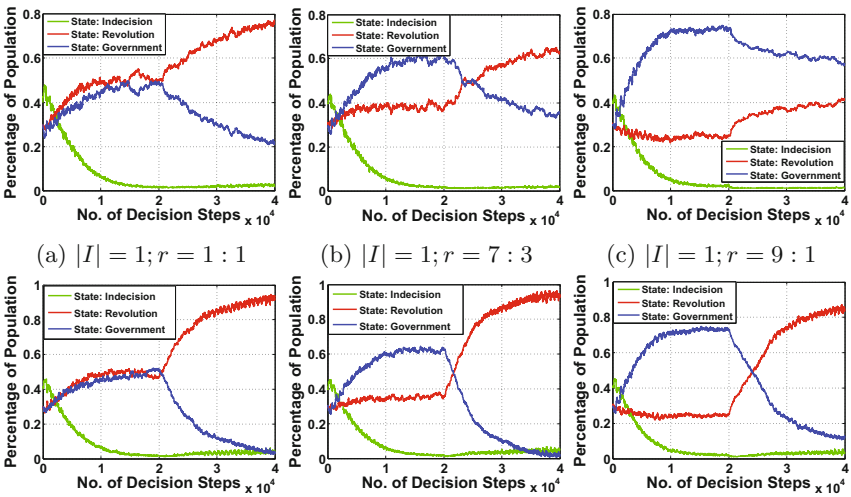
consequently, changing their opinions to G. Nonetheless, when bad policies outnumber the good policies significantly, the Revolution state rises to the top and becomes the dominant opinion of the population, Fig. 2f. The slight increase in the G state in the beginning of the simulation is because of this phenomenon. In conclusion, in absence of influencers, for unequal initial distribution both initial distribution and r are important in the steady state opinion distribution of the population.

3.2 Effect of Influencers

This experiment investigates the effects of presence of Influencers and their quantity on the steady state behavior of the population. In order to specifically observe the effect of Influencers, they are activated at decision step 10000, and they are biased towards the Revolution state. The influencers randomly choose people to form a links with, where the probability of forming a link is 0.25. Figure 3 presents the results of this experiment for $G_i = R_i = 0.3$. In the first experiment, only one Influencer is available in the society, Figs. 3a, b and c.

When $r = 1:1$, as discussed earlier, Independent members start joining the R or G state equally, Fig. 3a. However, as soon as the Influencer is activated, the percentage of people in the R state starts rising drastically because of the interactions which happen among the population. The interesting point is that there is a slight increase in the state of Indecision too. The reason can be found in the normative perspective of the G-Leaning group, Fig. 1b. As a result of continuous interactions with influencers, even a G-Leaning member might change his opinion to R, and the only path he can change his opinion from the G state

is through state I. This causes a slight increase in the number of agents in the state of Indecision.



(a) $|I| = 1; r = 1:1$ (b) $|I| = 1; r = 7:3$ (c) $|I| = 1; r = 9:1$

(d) $|I| = 5; r = 1:1$ (e) $|I| = 5; r = 7:3$ (f) $|I| = 5; r = 9:1$

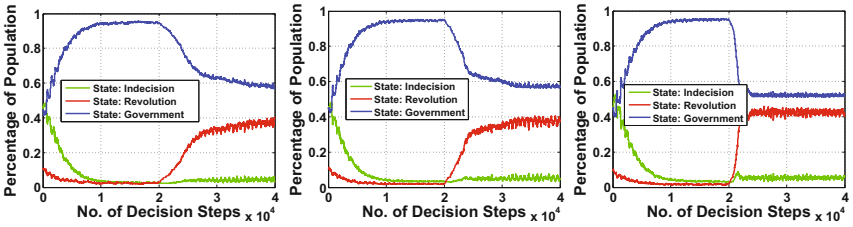
Fig. 3. Effect of external events on population opinion in presence of influencers

As the number of popular acts by the government increases, in the first phase of the simulation, the Independent members join the G state making it the dominant opinion of the whole population, Fig. 3b. However, the presence of the influencer affects peoples opinions through interactions causing the R state to be the dominant opinion of the society although there are more popular acts by the government. In this scenario, higher probability of the good actions just makes the transition to the R state slower. The same reasoning is applicable for Fig. 3c, with the only exception that presence of one influencer is not able to overcome the effect of significantly higher probability of the popular acts, to destabilize the society.

In another set of experiments, higher number of influencers were added to the population, Figs. 3d, e and f. It is observed that with more influencers present, the transition is faster, and almost all the populations joins the R state. Also, higher number of popular act by the government cannot prevent the destabilization of the society. This raises the question that "Can high number of Influencers guide the population towards destabilization under any condition?"

To answer this question, we consider an extreme case where the initial distribution is highly in favor of the G-Leaning group ($G_i = 0.45, R_i = 0.1$), and the probability of popular actions by the government is significantly higher. Figure 4 represents the results of simulation for such a case with varying number of influencers. As number of Influencers increases, the percentage of population

in the R state increases. Moreover, this transition is faster. However, the Influencers are not able to dominate the majority opinion. Both the dominant initial G-Leaning distribution and the high number of popular acts by the government cause this behavior. So, a large group of influencers is not a guarantee for guiding a population towards a predetermined state under all conditions.

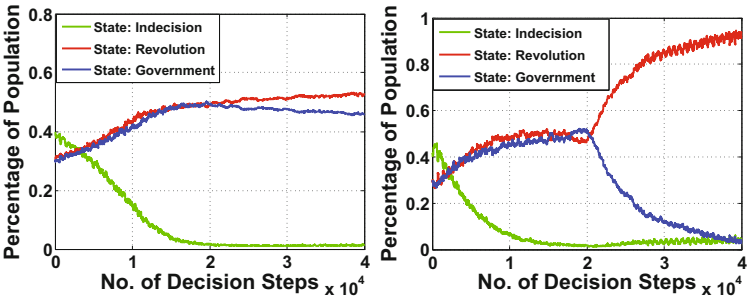


(a) $|I| = 5$ (b) $|I| = 10$ (c) $|I| = 50$

Fig. 4. Effect of number of influencers on a G-dominant society with $r = 9 : 1$

3.3 Effect of Distance Parameter (d)

Influencers deliberately advertise biased reward values in an attempt to pull the population slowly towards the state of their choice (R, in this study). Nonetheless, they would be successful in doing so if they are reachable for agents in the society. The parameter which controls reachability is the distance parameter d. In this section, we study the effect of the distance parameter on the steady state behavior of the population. Figure 5 presents the behavior of a society with two values of d. In Fig. 5a, because of the low distance parameter, most of the agents are not able to interact with Influencers. As a result the change in the population behavior is not significant. However, in Fig. 5b, the distance parameter is higher,



(a) $d = 0.5$ (b) $d = 1$

Fig. 5. Effect of distance parameter, $|I| = 5$, $r = 1 : 1$

more agents are affected by the Influencers as a result of interacting with them, and the change in the behavior of the society is significant.

4 Conclusion and Future Work

This paper studies the temporal evolution of opinions of individuals in an ideologically divided society by incorporation a PFSA framework along with the KH bounded confidence model. Three ideological groups called Independent, G-Leaning and R-Leaning, form a population in which people are connected. Indistinguishable influencers are also present in some experiments. There are two motives for individuals to change their decisions: popular/unpopular acts by the government, and interactions between people.

Results show that, in absence of influencers, ratio is the determining parameter in the equilibrium state of the population unless one of the groups includes a significantly higher number of members. The results also reveal that although very small in number, influencers are capable of creating drastic changes in the opinion dynamic of the society. Higher number of influencers result in faster transitions and attracting more people to the target state. It is shown that in the presence of a major group, there are situations where influencers, although very high in number, are not able to push the population's opinion towards the opinion of the minor group. Finally, it is presented that with a low distance parameter, the societies behavior is not greatly affected by the influencers.

References

1. Schonfeld, E.: Costolo: twitter now has 190 million users tweeting 65 million times a day, Techcrunch, vol. 8, June 2010
2. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery, pp. 36–43. ACM (2005)
3. Erickson, I.: Geography and community: new forms of interaction among people and places. Am. Behav. Sci. **53**(8), 1194–1207 (2010)
4. Shutters, S.T.: Cultural polarization and the role of extremist agents: a simple simulation model. In: Greenberg, A.M., Kennedy, W.G., Bos, N.D. (eds.) SBP 2013. LNCS, vol. 7812, pp. 93–101. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37210-0_11
5. von Neumann, J., Morgenstern, O., et al.: Theory of Games and Economic Behavior. Princeton University Press, Princeton (1944)
6. Fishbein, M., Ajzen, I.: Belief, attitude, intention, and behavior: an introduction to theory and research (1977)
7. Chakraborty, S., Mench, M.M.: Socio-cultural evolution of opinion dynamics in networked societies. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 78–86. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29047-3_10
8. Ray, A.: Signed real measure of regular languages for discrete event supervisory control. Int. J. Control **78**(12), 949–967 (2005)

9. Boyd, S., Ghosh, A., Prabhakar, B., Shah, D.: Randomized gossip algorithms. IEEE/ACM Trans. Netw. (TON) **14**(SI), 2508–2530 (2006)
10. Batagelj, V., Mrvar, A.: Pajek: program for analysis and visualization of large networks. In: Timeshift-The World in Twenty-Five Years: Ars Electronica, pp. 242–251 (2004)

Integrating Simulation and Signal Processing with Stochastic Social Kinetic Model

Fan Yang and Wen Dong(✉)

Department of Computer Science and Engineering,
State University of New York at Buffalo, Buffalo, USA
wendong@buffalo.edu

Abstract. Data that continuously track the dynamics of large populations have spurred a surge in research on computational sustainability. However, coping with massive, noisy, unstructured, and disparate data streams is not easy. In this paper, we describe a particle filter algorithm that integrates signal processing and simulation modeling to study complex social systems using massive, noisy, unstructured data streams. This integration enables researchers to specify and track the dynamics of complex social systems by building a simulation model. To show the effectiveness of this algorithm, we infer city-scale traffic dynamics from the observed trajectories of a small number of probe vehicles uniformly sampled from the system. The experimental results show that our model can not only track and predict human mobility, but also explain how traffic is generated through the movements of individual vehicles. The algorithm and its application point to a new way of bringing together modelers and data miners to turn the real world into a living lab.

1 Introduction

In this paper, we describe a particle filter approach to integrate signal processing and simulation modeling in the study of complex social systems using massive, noisy, unstructured data streams. As a result of this integration, researchers will be able to specify the dynamics of complex social systems by building a simulation model—a model that projects system-state changes over time—and applying the model to track complex systems and conduct thought experiments. We demonstrate this predictive ability in an application that tracks road transportation dynamics at city scale from the trajectories of probe vehicles and a state-of-the-art transportation simulator.

Data that continuously track the dynamics of large populations [1] have spurred a surge in research on social diffusion [4,5], social network dynamics [8], and human mobility [12]. These data are often massive, noisy, and unstructured. Processing such data and turning them into useful knowledge is not easy. Traditional signal processing and pattern recognition models for capturing the dynamics in the data often lack intuitions about how component behaviors lead to predicted system behavior in a complex system, and have difficulty incorporating the effect of non-recurrent conditions.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 193–203, 2017.
DOI: 10.1007/978-3-319-60240-0_23

Simulation modeling [2] is a method that captures system dynamics by simulating and collecting runtime system states. It is a widely adopted method for solving problems about complex systems, which is characterized by complex interdependence among components and nonlinear relationships between system behavior and component behavior. These simulated models have found widespread application in biology, industrial and systems engineering, economics, the social sciences, and education. A simulation model is easier to implement than an analytical model, and is intuitive to understand. However, it is a significant challenge to verify and validate a simulation model, and it is complicated to translate such a model from theoretical speculation into an enabling tool.

Our approach is to identify the simulation model of a complex system driven by a collection of events as a Markov process; to define the massive, noisy, and unstructured data streams as observations about this system; and to develop machine learning algorithms to track and learn the evolution of the latent state Markov process from these noisy observations. The key observation behind this approach is that a simulation model generates different sample paths with different probabilities, which are unambiguously identified by a sequence of events and the corresponding times at which these events occur. As such, the simulation defines a stochastic process with a probability measure assigned to the space of the sample paths that describe the interactions among elements in the system.

In this paper, we introduce the stochastic kinetic model to specify the dynamics of a complex social system with a set of events described by production rules, and design a particle filter algorithm to track the individuals in the system from noisy and partial observations and learn the system dynamics. A dynamic Bayesian network (DBN) [3,16] can alternatively represent the dynamics of a complex system because the values of the random variables describing the system at time t are probabilistically dependent on those at time $t-1$. In comparison with a DBN, an event-based model describes the system dynamics more succinctly by factoring the conditional probability of the system variables at two time steps into the probabilities for a sequence of events between these two time steps that incrementally change the system variables in simple ways. A deep neural network can potentially represent the arbitrarily complex dynamics of a system through a huge number of synaptic weights [9]; However, it requires a very large set of training data and huge computational resources to train these weights, and does not tell us how complex systems work microscopically, or what the consequences are of non-recurrent events. Variational inference is an alternative algorithm to track and learn complex system dynamics [20]. But it suffers from numerical stability issues when applied to complex systems.

The most creative and innovative aspect of this work is the integration of simulation modeling and signal processing in the study of complex social systems. This approach has not been explored because the intersection of the signal processing community and the simulation community is presently very small. However, this intersection is nonetheless very powerful because it affords an intuitive interpretation of the information extracted from massive, noisy, unstructured data streams. It has the potential to revolutionize how researchers use

simulation models, from running computer programs and analyzing program outputs to making inferences about a real-world system. For example, by integrating an agent-based transportation model with signal processing, we can not only simulate traffic jams during rush hour but also predict from the trajectories of probe vehicles whether today's traffic jams will be formed earlier or last longer than usual, and help drivers use the road network more efficiently. This kind of practical, transformative result is what happens when we bring together modelers and data miners to turn the real world into a living lab.

2 Stochastic Social Kinetic Model

We introduce the stochastic kinetic model to capture the dynamics of a complex social system driven by a set of events. A *stochastic kinetic model* is a biochemist's way of describing the temporal evolution of a biological network with M species driven by V mutually independent events [7,19], where the stochastic effects are particularly prevalent (e.g., a transcription network or signal transduction network). Let $\mathbf{X} = (\mathbf{X}^{(1)}, \cdots, \mathbf{X}^{(M)})$ denote the M species in the network. An event (chemical reaction) v is specified by a production

$$\alpha_v^{(1)}\mathbf{X}^{(1)} + \cdots + \alpha_v^{(M)}\mathbf{X}^{(M)} \xrightarrow{c_v} \beta_v^{(1)}\mathbf{X}^{(1)} + \cdots + \beta_v^{(M)}\mathbf{X}^{(M)}. \tag{1}$$

The production is interpreted as having *rate constant* c_v (probability per unit time, as time goes to 0). $\alpha_v^{(1)}$ individuals of species 1, $\alpha_v^{(2)}$ individuals of species 2 ... interact according to event v, resulting in their removal from the system. $\beta_v^{(1)}$ individuals of species 1, $\beta_v^{(2)}$ individuals of species 2 ... are introduced into the system. Hence event v changes the populations by $\Delta_v = (\beta_v^{(1)} - \alpha_v^{(1)}, \cdots, \beta_v^{(M)} - \alpha_v^{(M)})$, and $S = (\Delta_1^\top, \cdots, \Delta_V^\top)$ is therefore the *stoichiometry matrix*. The species on the left side of the production are *reactants*, the species on the right side of the production are *products*, and the species m with $\alpha_v^{(m)} = \beta_v^{(m)}$ are *catalysts*.

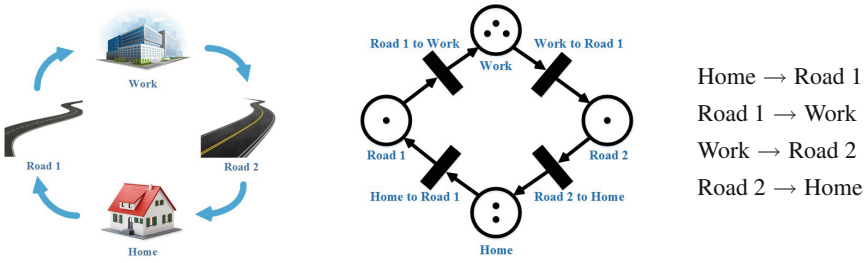
At the system level, let $x_t = (x_t^{(1)}, \ldots, x_t^{(M)})$ be the populations of the species in the system at time t. A stochastic kinetic process initially in state x_0 at time $t = 0$ can be simulated through the Gillespie algorithm [7] by iteratively (1) sampling the time τ to the next event according to exponential distribution $\tau \sim \text{Exponential}(h_0(x_t, c))$, where $h_0(x, c) = \sum_{v=1}^{V} h_v(x_t, c_v)$ is the rate of all events and $h_v(x_t, c_v)$ is the rate of event v, (2) simulating the event v according to categorical distribution $v \sim (\frac{h_1}{h_0}, \ldots, \frac{h_V}{h_0})$ conditional on event time τ, and accordingly (3) updating the system time $t \leftarrow t + \tau$ and populations $x \leftarrow x + \Delta_v$, until the termination condition is satisfied. In this algorithm, event rate $h_v(x_t, c_v)$ is the rate constant c_v multiplying a total of $\prod_{m=1}^{M}(x_t^{(m)})^{\alpha_v^{(m)}}$ different ways for individuals to interact in the system, assuming homogeneous populations. Exponential distribution is the maximum entropy distribution given the rate constant, and consequently is favored by the nature. The stochastic kinetic model thus assigns a probabilistic measure to a sample path induced by a sequence of events $v_1, \ldots, v_n$ happening between times 0 and T, $0 < t_1 < \cdots < t_n < T$, which is

$$P(v_{1:n}, t_{1:n}, x_{1:n}) = \prod_{i=1}^{n} h_{v_i}(x_{t_{i-1}}, c_{v_i}) \exp(-\sum_{i=1}^{n} h_0(x_{t_{i-1}}, c)(t_i - t_{i-1})), \quad (2)$$

where

$$h_v(x, c_k) = c_v g_v(x) \text{ for } v = 1, \cdots, V, \text{and } h_0(x, c) = \sum_{v=1}^{V} h_v(x, c_v). \quad (3)$$

The stochastic kinetic model is one way to define a discrete event process, and its equivalents in other fields include the stochastic Petri net [10,13], the system dynamics model [6], the multi-agent model specified through a flow chart or state chart [2], and the production rule system. As such, we argue that the stochastic kinetic model can capture the dynamics in not only biological networks but also social networks. For example, epidemiologists use productions such as "$S + I \rightarrow 2I$" and "$I \rightarrow R$" to represent infection and recovery events, and economists have developed the predator-prey model ("prey $\rightarrow$ 2 prey," "prey + predator $\rightarrow$ 2 predator," and "predator $\rightarrow \emptyset$") to represent the interactions among different industries. In our research, we use events of the form "$p_i \circ l_j \rightarrow p_i \circ l_k$" to specify the dynamics in a road transportation network in terms of vehicle movements: a vehicle previously at location j moves to location k (Fig. 1).



(a) Real world dynamics (b) Stochastic Petri net (c) Stochastic kinetic model

Fig. 1. Several representations of a complex system. (a) Real-world complex system dynamics, (b) a stochastic Petri net representation, (c) a stochastic kinetic model representation.

Computational social scientists often specify the complex dynamics of a social system with discrete event simulator software. To identify such a discrete event simulator as a Markov process and use the simulator to track real-world social systems with continued observations, we exploit the fact that all discrete event simulators (at least, to the best of our knowledge) have a way to dump the events happening in a simulation run. As such, we can reconstruct simulation runs according to the event sequences and so reconstruct the stochastic discrete event model from simulation runs outside the simulator—instead of unearthing the source code of a specific simulator over many man-months. For example, rather than hacking through the 140 thousand lines of code for MATSIM (which is a

state-of-the-art multi-agent transportation simulator) to make real-time inferences with real-world data, we can dump four events: vehicle leaving a building, vehicle entering a link, vehicle leaving a link, and vehicle entering a building. From these four events, we can construct a state transition matrix to represent vehicle dynamics.

Although the stochastic kinetic model is a continuous time model, we work with a discrete time stochastic model in the rest of this paper because our goal is to track stochastic kinetic dynamics from observations of populations or individuals with countably many computational steps. To this end, we approximate the continuous time process with a discrete time process on a countable set of equally spaced time points $0, \tau, 2\tau, \ldots$, with a time interval so small that the probability of more than one event happening in the interval τ is negligible. This approximation works because the state transition kernel from time 0 to time τ is $p(x_0 \to x_\tau) = \sum_{n=0}^{\infty} \left(I + \frac{Q}{\gamma}\right)^n \exp(-\gamma\tau)\frac{(\gamma\tau)^n}{n!}$ according to the uniformization method [11], where γ is a uniformization rate, I the identity matrix and Q the infinitesimal generator defined by $h_k, k = 1, \ldots, V$. With $\gamma \to \infty$ and $\gamma\tau = 1$, we get a first-order approximation of the state transition kernel $I + Q \cdot \tau$.

Specifically, let $v_1, \ldots, v_T$ be a sequence of events in the discrete time stochastic kinetic system, $x_1, \ldots, x_T$ a sequence of states (populations of species), and $y_1, \ldots, y_T$ a set of observations about the populations. Our goal is to make inferences about $\{v_t, x_t : t = 1, \ldots T\}$ from $\{y_t : t = 1, \ldots, T\}$ according to the following probability measure, where indicator function $\delta(x_t - x_{t-1} = \Delta_{v_t})$ is 1 if the previous state is x_{t-1} and the current state is $x_t = x_{t-1} + \Delta_{v_t}$, and 0 otherwise.

$$P\left(v_{1,\ldots,T}, x_{1,\ldots,T}, y_{1,\ldots,T}\right) = \prod_{t=1}^{T} P(x_t, y_t, v_t | x_{t-1}), \tag{4}$$

$$\text{where } P(x_t, y_t, v_t | x_{t-1}) = P(v_t | x_{t-1}) \delta(x_t - x_{t-1} = \Delta_{v_t}) P(y_t | x_t), \tag{5}$$

$$\text{and} \quad P(v_t | x_{t-1}) = \begin{cases} c_k \tau \cdot g_k\left(x_{t-1}\right) & \text{if } v_t = k \\ 1 - \sum_j c_j \tau g_j\left(x_{t-1}\right) & \text{if } v_t = \emptyset \end{cases}. \tag{6}$$

3 Particle Filter with Stochastic Kinetic Model

We apply a particle filter to track the dynamics of a stochastic kinetic process. The particle filter maintains a collection of particles x_t^k for $k = 1, \cdots, N$ and $t = 1, \cdots, T$ to represent the likelihood of the latent state of a stochastic process x_t at different regions of the state space with each particle representing a system state, given noisy and partial observations $y_{1,\cdots,T}$. It tracks the evolution of a stochastic process by alternately mutating the collection of particles according to stochastic process dynamics $P(x_t|x_{t-1})$ and selecting the particles according to observations $P(y_t|x_t)$. In comparison with a particle filter, a simulation run uses only one particle and does not use observations to perform particle selection.

Specifically, let x_t^k for $k = 1, \cdots, N$ be the collection of particle positions and v_t^k the corresponding events from particle mutation, and $i_t^k \in \{1, \cdots, N\}$ be the

collection of particle indexes from particle selection. To make inferences about the latent state x_t of a stochastic process starting at state x_0 from observations $y_{1:t}$, we initialize particle positions and indexes as $x_0^1, \cdots, x_0^N = x_0$ and $i_0^1 = 1, \cdots, i_0^N = N$, and iteratively sample the next event v_t^k according to how likely it is that different events will occur conditioned on system state $x_{t-1}^{i_{t-1}^k}$ for $k = 1, \cdots, N$ (Eq. 7), then update $x_t^k = x_{t-1}^k + \Delta_{v_t^k}$ accordingly (Eq. 8) and resample these events per their likelihoods with regard to the observation y_t for $t = 1, \cdots, T$ (Eq. 9).

$$v_t^k | x_{t-1}^{i_{t-1}^k} \sim \text{Categorical}(1 - \frac{h_0(x_{t-1}^{i_{t-1}^k})}{\gamma}, \frac{h_1(x_{t-1}^{i_{t-1}^k})}{\gamma}, \cdots, \frac{h_V(x_{t-1}^{i_{t-1}^k})}{\gamma}), \tag{7}$$

$$x_t^k = x_{t-1}^{i_{t-1}^k} + \Delta_{v_t^k}, \tag{8}$$

$$i_t^k | (x_t^{1:N}, y_t) \sim \text{Categorical}(p(y_t|x_t^1), \cdots, p(y_t|x_t^N)). \tag{9}$$

To determine a particle trajectory from the posterior distribution of a stochastic kinetic process with respect to observations, we trace back the events that lead to the particles $x_T^{i_T^k}$ for $k = 1, \cdots, N$:

$$x_0, v_1^{j_1^k}, x_1^{j_1^k}, \cdots, v_T^{j_T^k}, x_T^{j_T^k}, \text{where } j_T^k = i_T^k, j_{T-1}^k = i_{T-1}^{j_T^k}, \cdots, j_1^k = i_1^{j_2^k}. \tag{10}$$

To learn the rate constants (the parameters) of a stochastic kinetic model, we sample from the posterior distribution of these parameters conditioned on the particle trajectory, using a beta distribution as conjugate prior. Let $v_1, \cdots, v_T$ be the sequence of events in a particle trajectory (Eq. 10). The posterior probability distribution of a rate constant is a beta distribution that matches the expected number of events in a sample path with the number of events that actually occur, where a_v and b_v are hyper-parameters:

$$c_v | v_{1:T}, x_{1:T} \sim \text{Beta}(a_v + \sum_{i=1}^{T} \delta_v(v_t), b_v + \sum_{t=0}^{T} g_v(x_t) - \sum_{t=1}^{T} \delta_v(v_t)). \tag{11}$$

Overall, then, we have developed a particle-based algorithm to make inferences about a complex system using a stochastic kinetic model and noisy observations (Algorithm 1).

The discrete event model developed in this paper can be extended into a Markov discrete event decision process (MDEDP) to study how individuals make decisions probabilistically in order to jointly maximize a time-discounted future reward according to their perceptions of the world. To this end, we introduce action variables that are probabilistically dependent on the current system state for individuals to control the event rates, and specify the rewards for an individual to be in different states for a unit time. To tractably solve the MDEDP, we can first reduce its state value function into the probability of receiving a reward in a mixture of dynamics Bayesian networks [18], then search into the future for the best individual actions with the particle filter algorithm.

Algorithm 1. Particle-Based Inference with Stochastic Kinetic Model

Input: Observations $y_1, \cdots, y_T$ of a stochastic kinetic process (Eq. 4) specified by a set of events (Eq. 1 and 6 for $v = 1, \cdots, V$).

Output: Resampled particles $(v_t^{i_t^k}, x_t^{i_t^k})_{t=1:T}^{k=1:N}$ from particle filter, particle trajectories $(v_t^{j_t^k}, x_t^{j_t^k})_{t=1:T}^{k=1:N}$ from particle smoother.

Procedure:

- Initialize $x_0^1 = \cdots = x_0^N = x_0$, $i_0^1 = 1, \cdots, i_0^N = N$.
- (Filtering) For t in $1, \cdots, n$ and k in $1, \cdots, N$, sample v_t^k and i_t^k according to Eq. 7, 8 and 9, where $p(y_t|x_t)$ is defined in Eq. 4.
- (Smoothing) Back-track particle trajectory from $x_T^{i_T^k}$ according to Eq. 10, for $k = 1, \cdots, N$.
- (Parameter Learning) Sample rate constants according to Eq. 11.

4 Tracking City-Scale Transportation Dynamics

In this section, we evaluate the performance of the particle filter in continuously tracking current and future traffic conditions at city scale from an event-based transportation model and the sporadically observed locations of probe vehicles uniformly sampled from the system.

4.1 Modeling Traffic Dynamics

We model road traffic dynamics through a single type of event—$p_i \circ l_j \rightarrow p_i \circ l_k$—a vehicle i moving from link/building j to link/building k with rate constant c_{l_j,l_k}, changing its location from $X_t^{(p_i)} = l_j$ to $X_{t+1}^{(p_i)} = l_k$, changing the number of vehicles on link l_j from $X_t^{(l_j)} = x_t^{(l_j)}$ to $X_{t+1}^{(l_j)} = x_t^{(l_j)} - 1$, and changing the number of vehicles on link l_k from $X_t^{(l_k)} = x_t^{(l_k)}$ to $X_{t+1}^{(l_k)} = x_t^{(l_k)} + 1$. According to this model, a vehicle stays at link/building j for an average duration $1/\sum_k c_{l_j,l_k}$ and on exit chooses a downstream link/building with a probability proportional to the rate constant $c_{l_j,l_k} / \sum_{k'} c_{l_j,l_{k'}}$. Here we use "$\circ$" to represent a bond: person i binds to location j before the event and binds to location k after the event.

We assume that probe vehicles are uniformly sampled from the system. Let x_{ttl} be the total number of vehicles in the system and y_{ttl} be the total number of observed vehicles. The probability of observing $y_t^{(l_j)}$ probe vehicles at location j conditioned on that there are $x_t^{(l_j)}$ vehicles in total is $p(y_t^{(l_j)}|x_t^{(l_j)}) = \binom{x_t^{(l_j)}}{y_t^{(l_j)}} \binom{x_{\text{ttl}} - x_t^{(l_j)}}{y_{\text{ttl}} - y_t^{(l_j)}} \Big/ \binom{x_{\text{ttl}}}{y_{\text{ttl}}}$. When the total number of vehicles in the system is large, the percentage of probe vehicles at a given link/building is roughly the percentage in the system.

4.2 Experimental Setup

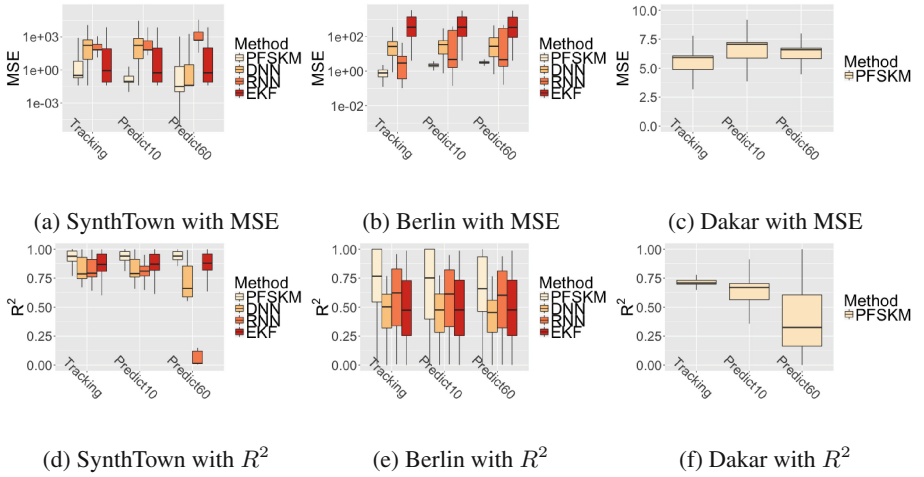
Here we compare the performance of the particle filter against other algorithms on three datasets of human mobility: SynthTown, Berlin and Dakar. The SynthTown dataset is comprised of a synthesized network of one home location, one work location, and 23 single-direction road links, with the trips of 2000 synthesized inhabitants going to work in the morning and going home in the evening [14]. This dataset is small enough for studying how different algorithms work. The Berlin dataset is comprised of a network of 24,000 single-direction road links derived from Open Street Map and the trips of 9000 synthesized vehicles representing the travel behavior of one million vehicles. The trips in the Berlin dataset were carefully validated with survey and sensor network data, and thus provide the ground truth for evaluating algorithms in a semi-realistic configuration [21]. The Dakar dataset is comprised of a network of 8000 single-direction road links derived from Open Street Map and 12,000 real-world vehicle trips derived from Data for Development call detail records [15].

To evaluate the effectiveness of our model, we compare the stochastic kinetic model (PFSKM) with a deep neural network (DNN) [9], with a recurrent neural network (RNN) [9], and with a dynamic Bayesian network (EKF) [17]. The vanilla neural networks represent the power of a general-purpose non-parametric model that does not involve a problem-specific structure. The dynamic Bayesian network characterizes the problem-specific structure through probability dependence among its random variables, and includes a suite of inference and structure learning algorithms. All models are trained with 30 days of traffic data and evaluated with a separate single day of testing data.

We use two metrics to evaluate the performance of our model: coefficient of determination (R^2) and mean squared error (MSE). We use R^2 to evaluate the goodness of fit between a time series of the estimated vehicle counts at a location and the ground truth. Let f_t be the estimated vehicle count at time t, y_t the ground truth and $\bar{y}$ the average of y_t. We define $R^2 = 1 - \sum_t (f_t - y_t)^2 / \sum_t (y_t - \bar{y}_t)^2$. A higher R^2 indicates a better fit between the estimated time series and the ground truth, with $R^2 = 1$ indicating a perfect fit and $R^2 < 0$ a fit worse than using the average. We use MSE to measure the average squared error difference between the estimated vehicle counts at all locations at a given time t and the ground truth. Let $f^{(i)}$ be the estimated vehicle count at location i and $y^{(i)}$ the ground truth. We define MSE $= \frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - f^{(i)})^2$. A lower MSE indicates an estimation closer to the ground truth.

4.3 Evaluation Results

Figure 2 compares the summary MSE and R^2 performance statistics for the four models in vehicle tracking—i.e., estimating the numbers of vehicles up to now, short-term prediction (10 min) and long-term prediction (1 h) on all data sets. The Dakar dataset is too large for DNN, RNN, and EKF, which indicates the better scalability of PFSKM. PFSKM has the lowest MSE across different times of day, which is followed by DNN, EKF, and RNN in order (top row, lower



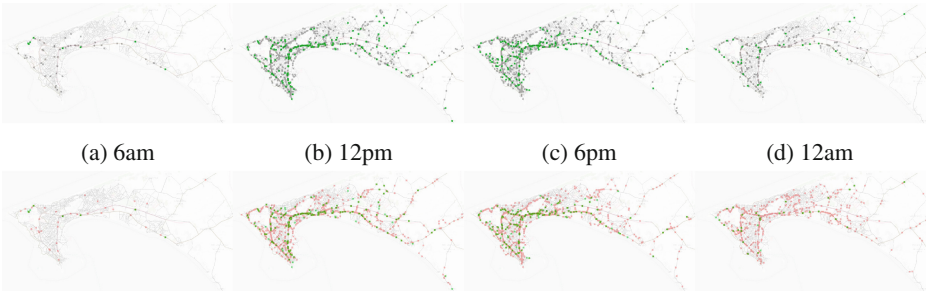
(a) SynthTown with MSE (b) Berlin with MSE (c) Dakar with MSE

(d) SynthTown with R^2 (e) Berlin with R^2 (f) Dakar with R^2

Fig. 2. Performance of PFSKM, DNN, RNN and EKF on the SynthTown, Berlin and Dakar datasets using MSE (top, lower MSE indicating better performance) and R^2 (bottom, higher R^2 indicating better performance).

is better). PFSKM also has the highest R^2 across different locations, which is followed by DNN, EKF, and RNN (bottom row, higher is better). PFSKM outperforms RNN and DNN because it can explicitly leverage problem-specific structures such as road topology. It outperforms EKF because it can work with arbitrary probability distributions, and sometimes Gaussian assumption is not a good approximation for real-world applications.

Figure 3 illustrates the output of PFSKM with four snapshots of the Dakar road network (bottom row) at 6am, noon, 6pm, and midnight, corresponding to four points in the particle trajectory. Each particle trajectory is an uninterrupted



(a) 6am (b) 12pm (c) 6pm (d) 12am

(e) 6am (f) 12pm (g) 6pm (h) 12am

Fig. 3. Dakar region traffic estimation and ground truth at four time points. The bottom figures are the estimation result, and the top figures are the ground truth. (Color figure online)

simulation in which the vehicles move from one location to another, explaining how traffic is generated with the movements of individual vehicles and what the consequences of non-recurrent events are. We have also provided the ground truth snapshots at the same time for reference (top row).

In these snapshots, probe vehicles are shown as green dots, simulated vehicles as red dots, and non-probe vehicles from the ground truth as black dots. There is no correspondence between the non-probe vehicles in the particle trajectory and those in the ground truth, because the non-probe vehicles are simply not observable. However, the vehicle densities at different locations of the road network and in the ground truth agree with each other, and both are proportional to the densities of probe vehicles. This is because for the particle filter, we have continually selected the most likely system evolution directions conditioned on the probe vehicle movements, and we backtrack these system evolution in particle smoother.

5 Conclusions

In this paper, we have described a particle filter algorithm for our stochastic social kinetic model that integrates signal processing and simulation modeling to study complex social systems using massive, noisy, unstructured data streams. Our method outperforms neural networks and the extended Kalman filter in inferring city-scale road traffic from continued observations of a small number of probe vehicles uniformly sampled from the system. This method points to a new way of bringing together modelers and data miners by turning the real world into a living lab.

References

1. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis (2015). arXiv preprint: arXiv:1502.03406
2. Borshchev, A.: The Big Book of Simulation Modeling: Multimethod Modeling with AnyLogic 6. AnyLogic North America, Chicago (2013)
3. Boyen, X.: Inference and learning in complex stochastic processes. Ph.D. thesis, Stanford University (2002)
4. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81**(2), 591–646 (2009)
5. Dong, W., Heller, K., Pentland, A.S.: Modeling infection with multi-agent dynamics. In: Yang, S.J., Greenberg, A.M., Endsley, M. (eds.) SBP 2012. LNCS, vol. 7227, pp. 172–179. Springer, Heidelberg (2012). doi:10.1007/978-3-642-29047-3_21
6. Forrester, J.W.: Industrial Dynamics. MIT Press, Cambridge (1961)
7. Gillespie, D.T.: Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. **58**, 35–55 (2007)
8. Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M.: A survey of statistical network models. Found. Trends® Mach. Learn. **2**(2), 129–233 (2010)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press, Cambridge (2016)

10. Goss, P.J., Peccoud, J.: Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. Proc. Natl. Acad. Sci. **95**(12), 6750–6755 (1998)
11. Grassmann, W.K.: Transient solutions in Markovian queueing systems. Comput. Oper. Res. **4**, 47–53 (1977)
12. Guan, T., Dong, W., Koutsonikolas, D., Qiao, C.: Fine-grained location extraction and prediction with little known data. In: Proceedings of the 2017 IEEE Wireless Communications and Networking Conference. IEEE Communications Society (2017)
13. Marsan, M.A., Balbo, G., Conte, G., Donatelli, S., Franceschinis, G.: Modelling with Generalized Stochastic Petri Nets. Wiley, New York (1994)
14. MATSim Development Team (eds.): MATSIM-T: aims, approach and implementation. Technical report, IVT, ETH Zürich, Zürich (2007)
15. de Montjoye, Y.A., Smoreda, Z., Trinquart, R., Ziemlicki, C., Blondel, V.D.: D4D-Senegal: the second mobile phone data for development challenge (2014). arXiv preprint: arXiv:1407.4885
16. Murphy, K.P.: Dynamic Bayesian networks: representation, inference and learning. Ph.D. thesis, University of California, Berkeley (2002)
17. Smith, G.L., Schmidt, S.F., McGee, L.A.: Application of statistical filter theory to the optimal estimation of position and velocity on board a circumlunar vehicle. National Aeronautics and Space Administration (1962)
18. Toussaint, M., Storkey, A.: Probabilistic inference for solving discrete and continuous state Markov decision processes. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 945–952. ACM (2006)
19. Wilkinson, D.J.: Stochastic Modelling for Systems Biology. CRC Press, Boca Raton (2011)
20. Xu, Z., Dong, W., Srihari, S.N.: Using social dynamics to make individual predictions: variational inference with stochastic kinetic model. In: Advances in Neural Information Processing Systems, pp. 2775–2783 (2016)
21. Ziemke, D., Nagel, K., Bhat, C.: Integrating CEMDAP and MATSim to increase the transferability of transport demand models. Transp. Res. Rec. J. Transp. Res. Board **2493**, 117–125 (2015)

Learning Network Dynamics from Tumblr®: A Search for Influential Users

Steven Munn[1(✉)], Kang-Yu Ni[2], and Jiejun Xu[2]

[1] Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 92092, USA
sjm@ece.ucsb.edu
[2] HRL Laboratories, Malibu, CA 90265, USA
{kni,jxu}@hrl.com

Abstract. This work offers an original analysis of a unique data set gathered from the blogging website Tumblr by developing and applying a new data driven method for investigating network dynamics. To our knowledge, this is the first effort to analyze the spread of information on Tumblr on a such a large scale, and our method generally applies to networks where nodes have time-evolving states. We start by testing our method on simulated data, then we follow over 50,000 blogs on Tumblr over a year of activity to determine not only which blogs are influential, but more importantly, how these blogs spread their content.

Keywords: Tumblr · Dictionary learning · Social networks · Influence maximization · Non-negative matrix factorization

1 Introduction

Given a large-scale social media data set from Tumblr® (a popular online microblogging platform), our goals are to, first, find the influential users and, second, identify the processes by which they spread their ideas. The first problem of finding influential users, also known as the influence maximization problem, is heavily dependent on having a good solution to our second objective. Up until now, however, most published work focuses on solving one problem or the other. To achieve our goals we will thus propose a new technique. By solving both problems together, we aim to do better at each one individually.

Our method builds on existing work in the study of network dynamics, such as state of the art research in traffic analysis, electrical grid balancing, or anomaly detection in sensor systems. This is because modeling the way blogs spread their content on Tumblr is a type of network dynamics problem itself. The blogs are interconnected in an irregular and time-varying fashion depending on how many users follow a blog or forward its content. And, the content of each blog is constantly changing depending on its activity. We can thus use ideas from general network dynamics research to reach our goals. So far, most research work has focused on how network connectivity changes over time [10]; however, the

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 204–213, 2017.
DOI: 10.1007/978-3-319-60240-0_24

processes by which nodes change their states (this corresponds to blogs changing their content in the context of analyzing Tumblr) are often ill-understood.

To shed light on these processes, we take inspiration from signal processing on graphs, a new area of research dedicated to analyzing the states of nodes on a graph by using well-understood techniques from electrical engineering [13]. We examine the changing states of nodes by representing these states as a series of attributes on the network, and then aggregating the attributes into a matrix representation. This allows us to use dictionary learning (DL) methods to obtain either a lower-dimensional representation of the attributes or a sparse representation thereof.

The main advantage of our approach is that it is agnostic of graph structure because this graph structure is often inferred from incomplete data. More importantly though, we typically do not know which aspects of the structure are important. The lower-dimensional representation obtained from dictionary learning helps us find the relevant structural features of a network from which we can infer the key nodes that drive widespread changes of state. To illustrate this idea, consider a protein interaction network. Although some proteins may react with many others, due to physical limitations, often there are restrictions on how many of these reactions can occur at the same time. These proteins may be of high degree in the network; however, only a few edges are actually relevant at any given time during the unfolding of the network process.

In this paper we motivate the use of DL with some background on signal processing on graphs. Then we explain our methodology and apply it to the study of opinion dynamics (or the spread of information), first on a set of simulation data, then with data collected from Tumblr.

2 Transforming Network Attributes

2.1 Signal Processing on Graphs

Our interest in using signal processing techniques to answer network science questions stemmed from the growing literature regarding signal processing on graphs [13]. A mainstay of this literature is the development of transforms for representing graph attributes as the linear combination of basis vectors (see [3,4,6,12] for a sample of relevant papers).

The starting point for these transforms is to consider graph attributes as a vector. A graph comprises a set of nodes $\mathcal{N}$ and edges $\mathcal{E}$. Assuming we have real-valued attributes (they could belong to any set, but we focus on the set of reals), then we can think of the attributes as the result of a function a that maps nodes to reals as in, $a : \mathcal{N} \to \mathbb{R}$.

If we label $a(n_i)$ the attribute corresponding to node i, then we define the attribute vector $\mathbf{x}$ as,

$$\mathbf{x} = [a(n_0), a(n_1), \ldots, a(n_{N-1})]^{\mathrm{T}}. \tag{1}$$

By extension, if the attributes at a given node are a time series, we can pack the attribute vectors at each time step into a matrix of attributes $\mathbf{X}$ that is $N \times t$ dimensional, where t is the number of time steps in the signal.

Perhaps the most well-known transform for graphs, the graph Fourier transform (GFT) expands attribute vectors as a linear combination of the graph Laplacian eigenvectors [5]. The GFT is especially useful for network models that represent a discretization of an underlying continuous space. The GFT assumes, however, that the graph attributes are smooth with respect to the network structure [5], which is often not the case with big data.

Another approach is to design wavelet transforms based on a model for the process underlying the graph attributes [3,4,6]. This works well for many applications such as traffic analysis [4], but these methods typically rely on a simple dynamical model, which makes the wavelets difficult to apply to more complicated processes. Furthermore, both the GFT and graph wavelet approaches rely heavily on graph structure, which limits their flexibility.

2.2 The Dictionary Learning Approach

Our new approach is to learn the basis vectors for our transform from the graph attributes using dictionary learning [14]. Specifically, we demonstrate this idea with non-negative matrix factorization (NMF) [7,9] for social network activity because the matrix $\mathbf{X}$ defined in Sect. 2.1 has non-negative entries in that case.

NMF is an algorithm for approximating a matrix with non-negative entries as the product of two matrices, also with non-negative entries. In other words, given an input matrix $\mathbf{X}$ with entries $x_{ij} \geq 0$, NMF solves the following optimization problem,

$$\min_{\mathbf{W},\mathbf{H}} ||\mathbf{X} - \mathbf{WH}||_{\mathrm{F}}^2 \tag{2}$$

subject to $w_{ij} \geq 0$, $h_{ij} \geq 0$ for all i and j, where $\mathbf{W}$ is an $N \times k$ (a number of components we set for NMF) matrix and $\mathbf{H}$ is $k \times t$.

NMF transforms $\mathbf{X}$ into a dictionary $\mathbf{W}$, where each column vector is a dictionary atom, and a coefficients matrix $\mathbf{H}$. This algorithm is useful for dimensionality reduction (setting $k < N$) [9], or for sparse encoding ($k > N$) [7]–both of which are helpful for our purposes. Dimensionality reduction will serve to identify the different sources, or starting points, for dynamic process on networks; sparse encoding will help us understand the nature of the processes unfolding on the network.

2.3 Dictionary Learning Adaptations for Improved Parallelism

For large datasets, we propose a modified approach that makes dictionary learning more efficient to run in parallel. Given a graph $\mathcal{G}$ of nodes $\mathcal{N}$ and edges $\mathcal{E}$, we can divide the graph into a set of subgraphs $\mathcal{L} = \{\mathcal{G}_0, \mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_l\}$ using community detection [1]. For each subgraph, we perform the analysis described above. That is, for subgraph $\mathcal{G}_i$, we get a signal matrix $\mathbf{X}_i$ from which we obtain a coefficients matrix $\mathbf{H}_i$ and a dictionary atoms matrix $\mathbf{W}_i$ via NMF.

For clarity, and without loss of generality, we assume that the nodes of $\mathcal{G}$ are numbered such that the nodes in subgraph $\mathcal{G}_0$ start at zero and end at $N_0 - 1$, while the nodes in $\mathcal{G}_1$ start from N_0 and end at $N_0 + N_1 - 1$, and so on. Using the full matrix $\mathbf{X}$ as input, we initialize the dictionary learning algorithm with the following dictionary matrix:

$$\mathbf{W} = \begin{bmatrix} \underbrace{\begin{matrix} | & | \\ | & | \\ \mathbf{r^T} & \mathbf{r^T} \\ | & | \\ | & | \end{matrix}}_{m} & \dots & \begin{matrix} \mathbf{W_0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W_2} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{matrix} & \dots \end{bmatrix} \tag{3}$$

where m is the number of global-scale atoms we wish to learn, $\mathbf{0}$ are appropriately sized zero matrices, and $\mathbf{r}$ is a row vector with entries taken at random or following a dictionary learning initialization step (as in [2]).

Instead of optimizing the entire $\mathbf{W}$ matrix, however, we only update the first m columns. This means we are learning m new global-scale atoms, and keeping the local atoms intact.

3 Opinion Dynamics Simulations

To better explain how NMF is useful for understanding opinion dynamics, we illustrate our method with some examples using simulated data. In this section, we explain how we generated the data to best reflect the type of data available from social media.

3.1 The Decaying Cascade Model

The proposed decaying cascade model is an adaptation of the independent cascade model [8] meant to reflect the fact that ideas tend to have limited reach across a social network: Your friend's friend is usually less likely to adopt your opinions than your friend. It is also closely related to the continuous independent cascade model of Gomez et al., which better represents the temporal aspects of social media data [11]. Algorithm 1 summarizes the steps for the decaying cascade simulation.

The algorithm starts with a set of initially active persons (nodes) $\mathcal{V}_0$ and keeps track of nodes that change their states in the set of events E. After a random amount of time, a randomly selected node (the source) attempts to convince one of its neighbors (the target) to adopt its opinion. The probability of success is,

$$p\left(n_{\text{Target}}\right) = p_0 \exp\left(-d_{\mathcal{V}_0}\left(n_{\text{Target}}\right)\right) \tag{4}$$

where n_{Target} is the targeted node, p_0 a probability constant, and $d_{\mathcal{V}_0}\left(n_{\text{Target}}\right)$ is the shortest path distance on the network from the target to one of the nodes in $\mathcal{V}_0$.

Algorithm 1. Decaying Cascade

Set $t = 0$
Pick $\mathcal{V}_0$ from $\mathcal{N}$, and set $V = \mathcal{V}_0$
Initialize $E = \{(t = 0, n) : n \in \mathcal{V}_0\}$
while $t < T$ **do**
    $t \leftarrow t +$ Random increment
    Randomly pick $n_{\text{Source}} \in V$
    Randomly pick $n_{\text{Target}} \in$ Neighbors of n_{Source}
    Pick outcome with probability $p\,(n_{\text{Target}})$
    if Source convinces target **then**
        $V \leftarrow V \cup n_{\text{Target}}$
        $E \leftarrow E \cup (t, n_{\text{Target}})$
    else
        n_{Source} gives up on n_{Target} forever
    end if
end while

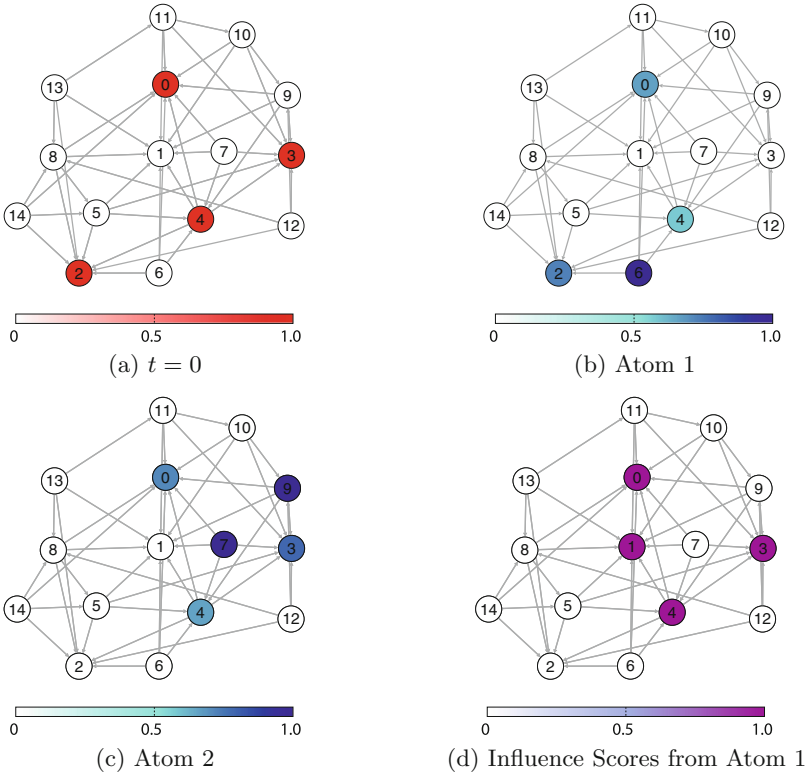
3.2 The Threshold Model

The proposed threshold model, modified from the linear threshold model [8], is exactly the same as our decaying cascade model up until the probability of convincing the target. Here, the target will only be convinced if the number of active neighboring nodes is above a predefined threshold. With a high enough threshold, the spreading of opinions will have similar properties to the decaying cascade model with respect to spatial localization across the nodes.

4 Simulation Analysis

To best reflect the idea that social media data typically involves the spreading of many ideas in parallel, we set up our simulation as follows: We select a number of "influential" nodes $\mathcal{V}_{\text{Influential}}$ (nodes that can initialize a process) and run a large number (one to two hundred) of threshold or decaying cascade model simulations where the initially actives nodes are a subset of $\mathcal{V}_{\text{Influential}}$. We then aggregate the set of events from each simulation into a global set of events for the network from which we derive the signal $\mathbf{X}$ that we use for dictionary learning.

Figure 1(a) shows the set of nodes $\mathcal{V}_{\text{Influential}}$ in red for a series of threshold simulations where three nodes are necessary to activate a new node. In figures (b) and (c) we can see two NMF atoms (of four emanating from the algorithm). The information spreads from the lighter blue nodes to the darker blue ones. Since the atoms are effective at showing the direction in which information is spreading, we define influence scores to find the most influential nodes as follows: We first find the top 20 atoms according to their coefficient magnitudes and then compute a weighted sum of these atoms. The influence score at each node is the average value of its neighboring nodes in the weighted sum atom. In figure (d), nodes whose neighbors have a high value in atom 1 (the top atom in our ranking for the influence scores) are colored in purple.



(a) $t = 0$ (b) Atom 1

(c) Atom 2 (d) Influence Scores from Atom 1

Fig. 1. Non-negative matrix factorization atoms for the threshold model of Sect. 3.2. In (a), the initially active nodes are in red. In (b) and (c), we plot the first two NMF atoms. Information is spreading from the lighter blue nodes to the darker blue ones. In (d) nodes in purple have neighbors whose values are high in atom 1. (Color figure online)

Table 1 summarizes the results of the following experiment: On a graph with four hundred nodes, we run a series of threshold and decaying cascade model simulations, with 20 influential nodes, and generate a signal that we use for our method. We rank the nodes in terms of importance using the influence scores. Of the top 20 nodes, if the nodes match one of the initial nodes set for the simulation we count that towards the "exact node" accuracy; if it is within one hop of one of an initial node, we count that towards "within one hop" accuracy. The accuracies are averaged over 100 experiments. For comparison, Table 1 includes a baseline method in which the nodes deemed most influential are the most active ones.

To better understand how information is spreading on the network, we can set up NMF to learn an overcomplete basis of atoms. These atoms will be different depending on the process. For example, in atoms learned from cascade model simulation data, only a single node tends to be active; however, four nodes are

Table 1. Accuracy measures for our NMF-based influence maximization algorithm. Using the two opinion dynamics models described earlier to generate data with ground truth key nodes, we compare the relative number of correctly identified sources, and the relative number of sources identified within a one-edge hop distance.

Algorithm	Accuracy	
	Exact node	Within one hop
Threshold model simulations		
NMF-based	$23.8 \pm 3.4\%$	$100.0 \pm 0.0\%$
PageRank	$15.0 \pm 0.0\%$	$15.0 \pm 0.0\%$
Baseline	$14.7 \pm 1.2\%$	$14.7 \pm 1.2\%$
Decaying cascade simulations		
NMF	$33.4 \pm 2.3\%$	$100.0 \pm 0.0\%$
PageRank	$15.0 \pm 0.0\%$	$15.0 \pm 0.0\%$
Baseline	$15.7 \pm 3.8\%$	$15.8 \pm 3.8\%$

usually active when the dynamics come from a threshold model with a three-node activation requirement.

5 Real Data Results

Tumblr is a microblogging website with tens of millions of active users where people write or respond to short blog posts. In this section we use our NMF approach to find influential users on Tumblr and better understand how people's ideas spread on this social network.

For this analysis, we use the reblog network as the underlying graph structure. Users are nodes, and directed edges correspond to users who rebloged each other's content. In other words, if user i reblogs user j's posted content, we have a directed edge from i to j in the reblog network.

We demonstrate our method on a group of 59,709 users, which are picked using a community detection method. First we build the reblog network for the entire dataset. Then we find the largest connected component (approximately six million users). Finally, we use Louvain components [1] to find a community of users with a large-enough number nodes. After selecting the users, we build an attribute matrix according to their activity on Tumblr for a year. The sampling rate is twelve hours, so $\mathbf{X}$ is a $59{,}709 \times 703$ matrix.

Using the parallelism method described in Sect. 2.3, we partition our graph of 59,709 users into two subcomponents, and set up NMF to learn 200 atoms for each subcomponent over 100 iterations. After combining the two dictionaries following Eq. 3, we then learn an additional 200 atoms over 100 iterations. All these steps take a total of 54 min to run on a single server node.

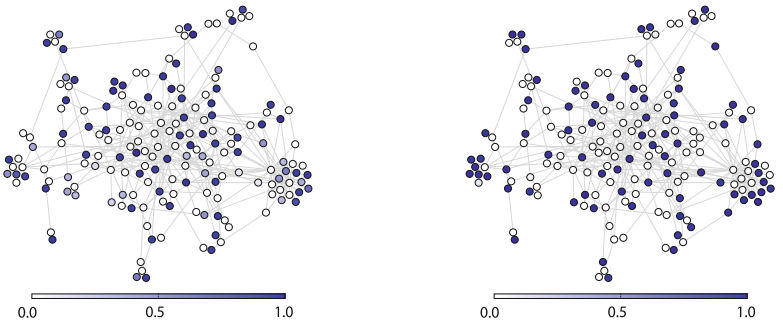
After ranking users according to the techniques of Sect. 4, we found that the most influential blogs according to our method mostly revolve around humor and fashion. Table 2 lists the top six blogs we found.

Table 2. The top six most influential blogs in a community of 59,709 users according to our method.

Rank	Description
1	Video collection features slapstick humor
2	Fashion and film
3	Humorous video collection with a voting system
4	Absurd-type humorous posts
5	Cartoon animation enthusiast's blog
6	Self-proclaimed novelist's posts

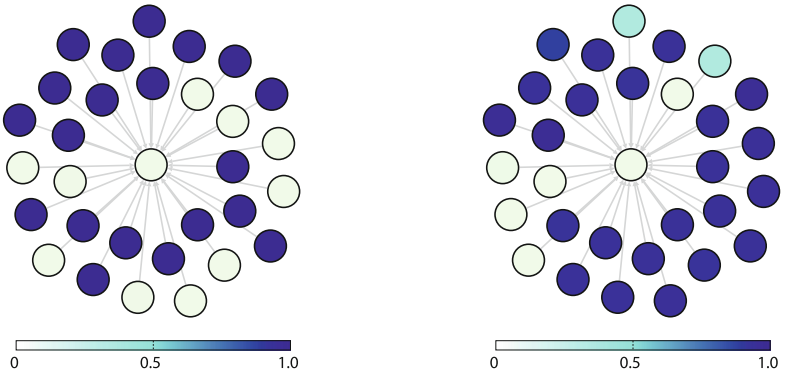
By comparison, other influence metrics including centrality measures (e.g., betweenness, degree, and PageRank®) and the popular online ranking system Alexa® also attribute high scores to nodes that our method deems influential. The ranking order, however, is usually different because the methods do not take into account how information spreads on Tumblr. This supports our assumption that even graph structures that should be highly reflective of network dynamics do not paint the full picture of how the processes are unfolding on graphs.

To examine the opinion dynamics processes in more detail, Fig. 2 shows the value of the atoms at the scale of a small component in the network. Although plotting software would allow us to look at larger components, for illustration purposes in this paper, a smaller component is preferable. In Fig. 2 there are two of the most important atoms. They show us how information is traveling within the component. Some paths are consistently active, while others are not.



(a) Atom 1 (b) Atom 2

Fig. 2. A Louvain component of two of the most important atoms according to average coefficient magnitude. The atoms exemplify the characteristic ways in which the users post in this component. Some paths in the network are active in one atom but not another, which suggests that information sometimes travels along these routes.



(a) Atom 1 (b) Atom 2

Fig. 3. Top two atoms by average coefficient magnitude values centered around an influential blog and its one-hop ego network. Both atoms typify the most prevalent ways in which users respond to the central blogs content. Some nodes with strong values in (a) have lower values in (b) this means that they respond more selectively to the center blogs posts.

Figure 3 offers a different perspective on the atoms by focusing on individual users. For this figure, first we rank the atoms in terms of their importance by computing the average magnitude of the coefficient corresponding to each atom over the whole signal and organize the atoms in decreasing order of this average. Then we plot the atom values in the ego network of a particular node using the top two atoms in the ranking. This shows us two common ways in which users linked to the influential node post content. Users with a strong value in one atom are likely to post within 12 h of others in the atom who also have strong values. Users with strong values across the different atoms are generally more responsive to the central user's content. We do find, however, users who only have strong values in only a few atoms, suggesting that they only respond to particular topics from the center node.

6 Conclusion

Network dynamics underlie many real-world phenomena, and understanding them is crucial for scientific and engineering research. Focusing our attention on opinion dynamics, we set out to shed light on the processes by which people share ideas, and we studied the task of finding key sources of information or opinions. Our analysis shows that our dictionary learning-based method accomplishes these goals in a novel way and opens the way to better understanding the spread of ideas and information. Possible extensions of our work include modeling competing opinions with positive and negative signal matrix values, or difference encoding the signal over time, which could potentially help us keep track of the sources and targets of user activity.

References

1. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech.: Theory Exp. **2008**(10), P10008 (2008). https://doi.org/10.1088%2F1742-5468%2F2008%2F10%2Fp10008
2. Boutsidis, C., Gallopoulos, E.: SVD based initialization: a head start for nonnegative matrix factorization. Pattern Recogn. **41**(4), 1350–1362 (2008). https://doi.org/10.1016%2Fj.patcog.2007.09.010
3. Coifman, R.R., Maggioni, M.: Diffusion wavelets. Appl. Comput. Harmon. Anal. **21**(1), 53–94 (2006). https://doi.org/10.1016%2Fj.acha.2006.04.004
4. Crovella, M., Kolaczyk, E.: Graph wavelets for spatial traffic analysis. In: IEEE INFOCOM 2003. 22nd Annual Joint Conference of IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428). Institute of Electrical and Electronics Engineers (IEEE) (2003). https://doi.org/10.1109%2Finfcom.2003.1209207
5. Grady, L.J., Polimeni, J.R.: Discrete Calculus. Springer, London (2010). Nature
6. Hammond, D.K., Vandergheynst, P., Gribonval, R.: Wavelets on graphs via spectral graph theory. Appl. Comput. Harmon. Anal. **30**(2), 129–150 (2011). https://doi.org/10.1016%2Fj.acha.2010.04.005
7. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**(November), 1457–1469 (2004)
8. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2003. Association for Computing Machinery (ACM) (2003). https://doi.org/10.1145%2F956755.956769
9. Lin, C.J.: Projected gradient methods for nonnegative matrix factorization. Neural Comput. **19**(10), 2756–2779 (2007)
10. Newman, M.E.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2006)
11. Rodriguez, M.G., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2010. Association for Computing Machinery (ACM) (2010). https://doi.org/10.1145%2F1835804.1835933
12. Sandryhaila, A., Moura, J.M.F.: Discrete signal processing on graphs. IEEE Trans. Signal Process. **61**(7), 1644–1656 (2013)
13. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process. Mag. **30**(3), 83–98 (2013). https://doi.org/10.1109%2Fmsp. 2012.2235192
14. Tosic, I., Frossard, P.: Dictionary learning. IEEE Signal Process. Mag. **28**(2), 27–38 (2011)

Modeling the Impact of Protraction on Refugee Identity

Erika Frydenlund(✉) and José J. Padilla

Virginia Modeling, Simulation and Analysis Center, Old Dominion University, Suffolk, VA, USA
{efrydenl,jpadilla}@odu.edu

Abstract. This study presents an agent-based model of identity shift based on refugees in protracted situations. The model is based on interactions between refugees, local citizens, and nongovernmental workers. Through repeated interactions at the individual level, we see the emergence of groups of identities that are generally neutral—not nationalistic nor specifically locally or globally focused. The model provides a starting point for understanding the processes of long-term protracted states of displacement on the general shift in identity among refugee populations in both camp and non-camp based settings.

Keywords: Agent-based modeling · ABM · Refugees · Protraction · Identity

1 Introduction

Massive forced migration of people is quickly becoming a defining feature of our time. From the waves of Syrian and African refugees collecting at the borders of Europe to the protracted refugee situations that have been going on for multiple decades, displacement is a humanitarian concern at the local, state, and global levels. Forced migration including refugees, asylum seekers, and internally displaced persons are at historic highs, with time spent in exile averaging around 25 years [1, 2]. This means that generations of people are born into refugeehood, without direct experience living in the state to which a repatriation effort would send them. What does this mean for traditional territorial notions of state sovereignty? How do refugees born, raised, and educated in exile conform to our understanding of citizen or nationality? This paper uses a modeling and simulation approach to theorize on the implications for identity shift as refugees are kept away from 'home' for extended periods of time. Using an agent-based model to look at identity shifts, we construct two scenarios: camp-based refugees with limited interaction with locals in the host population, and urban refugees with regular intermingling with local host nationals. While this only paints a theoretical picture of what may be happening among the millions of refugees in protracted situations globally, it demonstrates a methodology for theorizing about attitude shifts away from traditional notions of citizen and territorially-based sovereign states.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 214–222, 2017.
DOI: 10.1007/978-3-319-60240-0_25

2 Protraction and Refugee-Based Identity

The United Nations defines protracted refugee situations as those where a group of more than 25,000 people have been displaced for at least five years [3]. Conservatively, this definition may include around 70% of all refugees, with around 70% of protracted cases stretching at least two decades [4, 5].

Whether in camps or urban areas, decades spent in exile equates to generations born outside of the parents' country of origin and experiencing life outside of traditional notions of 'citizen' or 'nationality.' Physically, refugees exist at the local level of host country spaces. Given the extensive exposure to regional and international NGO workers and the values and norms these individuals transmit through daily interactions, refugees also occupy a supraterritorial space. This is also true of their involvement in the global economy [6] and potential for a role in international conflict [7] and peacebuilding [4]. Refugees also represent disruption of social contracts understood to exist between the citizen and state, as they flee from the failure of their home state to protect them [8–10].

Refugees embody a kind of 'imagined community' [11] among displaced persons that persists despite the realities of ongoing conflict, generations of separation, and inaccessible citizenship rights in their own country of origin. Whether urban or rural, refugees reconstruct identities of themselves that differentiate them both from the host population and their country of origin. Malkki's [12] research points to an idea that refugees construct different types of identities in urban spaces than in rural, presumably in response to increased interaction with the local population and opportunities to adapt to different cultural norms, languages, and economic opportunities.

In cases of long-term forced displacement, refugees may construct identities around the refugee label, where humanitarian intervention "schemes become a vehicle for transforming an identity where refugees are marginalized into a segregated and permanently transient and dependent status" [13]. In these contexts, refugees demonstrate agency to negotiate their identities as they adopt the label 'refugee' for "political currency" to reap certain benefits from humanitarian actors [13]. Zetter proposes the notion that reliance on the refugee label for aid dependency is an act of agency used to sustain an "image of transitory status" and lay claim to the politicized status of refugeehood in protraction [13].

This paper presents a preliminary model as part of a longer-term study on the identity of protracted refugees. Using an agent-based modeling and simulation approach, the current theoretical model examines how identities may transmit between individuals and how interactions with other identities—limited in the case of encamped refugees and more widespread among urban refugees—shapes the overall identity values. Unlike sociological and anthropological research, modeling and simulation necessitates abstraction in an effort to understand potential underlying dynamics of complex phenomena like identity shifts.

3 Modeling Approach

Agent-based models have been used with relatively limited scope in terms of understanding forced migration. In political science and sociology more broadly, however, agent-based models have been used to theorize on or understand the problem space of the spread of norms. In political science broadly, we see models of norm diffusion [14, 15], theories of social group formation [16], emergence of shared identities [15], and latent identities [17–19]. In more specific application contexts, there has also been work on ethnic mobilization and potential refugee militarization [20–23]. Edwards [24] made a case for the use of computational social science models, such as ABM, to explore refugees and forced migration factors in the *Journal of Refugee Studies*, which is contrasted with the field's larger emphasis on traditional statistical analysis or qualitative research methods. The model presented here is an extension of a model constructed to illustrate identity shifts in a contained refugee population over a long stretch of time [25]. In this paper, we compare the results from that model of encampment to one where the boundaries of the camp are removed to mimic interactions in an urban space. The foundational question this paper asks is: Do different identities arise in camps versus urban refugee settings? While this question has been asked from a qualitative perspective in other studies such as those by Malkki [12] and Turner [26], the ABM approach here allows us to explore the generalizability of such observations.

4 Model Development

The model extends Jager and Amblard's algorithm to explain how interactions shape individuals' attitudes by considering three types of agents (instead of two) [27]. In the original algorithm [27], two agents interact and assess each other's identity values. If agent A's identity value is beyond the threshold of rejection, then agent B will adjust her identity farther away from agent A. In that case, the identity was too extreme (say, the interaction between an extreme nationalist of the host country and a refugee who feels complete nationalism in her country of origin), so the identity shift away from agent A's represents a kind of backlash where agent B more staunchly embraces her identity. If agent A's identity is within the threshold of acceptance, then B finds A's values somewhat persuasive and moves her identity closer to A. If the difference between A's and B's identities is between the threshold of acceptance and threshold of rejection, then there is no change. At each interaction, refugee agents adjust their identity values in this way, as in the equations below:

If $\left\|x_i - x_j\right\| < t_i$	$\mathrm{dx_i} = \mu \cdot (\mathrm{x_j} - \mathrm{x_i})$
If $\left\|x_i - x_j\right\| > t_i$	$\mathrm{dx_i} = \mu \cdot (\mathrm{x_i} - \mathrm{x_j})$
If $u_i \leq \left\|x_i - x_j\right\| \leq t_i$	No change in identity

Where x_i is the identity of agent i, and x_j is the identity of agent j. The upper threshold, t_i, represents the threshold of rejection, while u_i represents the threshold of acceptance. The value μ in the equations above is a weighting mechanism to influence the change in identity value. Based on these calculations, in the following timestep, each agent updates its identity to:

$$x_i = x_i + dx_i$$

In the proposed model there are three types of actors (called 'agents') in the model: refugees, NGO workers, and local citizens. The model allows for testing both normal and uniform distributions of the starting identity values for each agent type of agent.

NGO workers (NGOs) can move freely into and out of the camp. They maintain individual, heterogeneous, static "global" identity values ranging from [0, 1) where zero is neutral and one is very globally oriented, perhaps not strongly identifying with any territorial nation. This simplification assumes that that even locals and refugees who work for NGOs will have some exposure to international norms and values that make their identities 'global.' NGOs do not update their static identity values.

Local citizens (Locals) cannot move into the camp. In the real world, markets and services inside of refugee camps result in porous borders and interactions between locals and refugees. We simplify since these numbers are often very low. Locals maintain a static "local" identity. Local identities heterogeneously vary from (−1, 0], where zero is neutral and negative value is very locally oriented, with only a national or local regional basis for identity and no alignment with the larger global population. This might represent someone who is identifies most strongly with a local town or ethnic identity but does not situate her identity in any large global context. Locals, upon interaction with refugees at the borders of the camp, can influence refugees' sense of local identity, but do not adjust their own static identity values.

Refugees cannot walk outside the camp boundary. This is assumed as an extreme containment case to contrast with the urban setting. Refugees begin with heterogeneous global and local identity values distributed normally or uniform (based on the experimental run) from [−1, 1]. A local identity of −1 indicates full alignment with the host population, zero is neither host nor country of origin, and 1 is full alignment with the country of origin. A global identity of −1 indicates completely territorial-based identity, zero is neither territorial nor global, and 1 is a fully global identity with no ties to a territorial state or nationality.

NGOs influence refugees' global identity values; Locals influence their local identity values; and other refugees influence both local and global identity values. Each refugee possesses heterogeneous, uniformly distributed threshold values for acceptance and rejection that dictate how they adjust identity values with each interaction.

In the model, Locals possess a static μ_L, and NGOs share the same μ_N. Refugees have a heterogeneous value μ_{Ri} that weights the influence they place on interaction with other refugees. The μ_L and μ_N determine the influential weight of refugees' interactions with Locals and NGOs respectively. These values are static throughout the simulation, and can thus be varied experimentally during simulation runs. The model was built in NetLogo [28] and run on a High Performance Computing Cluster. The experiment varied the following parameters: Number of Locals [5000, 10000] by increments of 2500; μ_L [0.1, 0.5] by increments

of 0.1; Number of NGOs [0, 2000] by increments of 500; μ_N [0.1, 0.5] by increments of 0.1; Number of Refugees [5000, 10000] by increments of 2500; Each parameter combination was run nine times and for 5000 time steps, for a total of 540,000 simulation runs. The ratio of NGOs and Locals to Refugees does not reflect any real-world scenario as this data could not be found, so we chose to vary the parameters widely to see the effects of differing ratios on refugees' identity formation. As the model is based on theory, we cannot currently know real-world estimates for t_i or u_i. Data collection efforts will start later this year.

We verified that the model was running as expected using a type of 'model docking' [29] where we compare the results of this model under similar values of factors illustrated in Jager and Amblard's original paper [27]. Through visual inspection, we determined the model performs as expected producing three main identity clusters for the same thresholds of acceptance (U = 0.2) and rejection (T = 1.6) for all agents. Additionally, we employed a set of genetic algorithms (Gas) to search the model's parameter space, looking for sensitivity in parameter combinations that might 'break' the simulation [30]. Using the Behavior Search application [31], we used GAs with a mutation rate of 0.03, crossover rate of 0.7, a tournament size of 3, and test population of 50 in each round to search the parameter space for combinations that minimized the variance of identity values for refugees. Each search was sampled 10 times. The GAs searched parameter space for abnormalities in updated identity values among refugee agents. No potential bugs or chaotic behavior were found using this method, indicating that the model is relatively robust to parameter value combinations.

5 Simulation Results

For both local and global identities, the environment and the starting identity distributions make statistically significant impacts on the distribution of identities at the end of the simulation (Tables 1 and 2). We conducted two-way ANOVAs to determine the effects of distribution (normal, uniform) and environment (camp, urban) on the identity values (local, global). The exceptionally large F-values indicate a statistically significant difference between these groups. In fact, starting distribution of agent identity values accounts for much of the difference between groups.

Table 1. Two-way ANOVA: average local identity by environment & distribution

	Type III sum of Sq.	d.f.	Mean Sq	F	Sig.
Environment	393.477	1	393.477	38,213.23	0.00
Distribution	84,549.59	1	84549.59	58,628.85	0.00
Envt * dist	294.788	1	294.788	28,628.85	0.00
Error	5,559.33	539,904	0.01		
Total	15,551.87	539,908			

$r^2 = 0.939$; adjusted $r^2 = 0.939$

Table 2. Two-way ANOVA: average global identity by environment & distribution

	Type III sum of Sq.	d.f.	Mean Sq	F	Sig.
Environment	13.975	1	13.975	937.784	0.00
Distribution	116,002.06	1	116,002.06	7,784,295.87	0.00
Envt * dist	25.873	1	25.873	1,736.21	0.00
Error	8,047.05	539,996	0.02		
Total	189,439.93	540,000			
$r^2 = 0.935$; adjusted $r^2 = 0.935$					

In Fig. 1 below, the grouping to the left centered near zero for all μ_L reflect those runs with uniformly distributed starting identity values. Those on the right that have shifted toward a strong local identity value of 1 derive from normally distributed starting identity values. Comparing them visually, the camp-based identities tend to form around a singular identity regardless of starting distribution type, where the urban refugees form smaller identity sub-groups as noted by the curves. These differentiating identities reflect the freedom of movement available to agents in the urban environment, where refugees are more likely to interact with urban host nationals. Most notably, when the starting distribution for identities is normal, refugees shift their identities dramatically away from the local host identity and towards that of their country of origin. In fact, the uniform starting distribution reflects preliminary fieldwork among protracted refugees, where identities converge closer to a neutral identity tied neither to the home nor host country. This disparity between simulation outcomes based on starting distribution will help to shape our fieldwork approach as well as future models that will incorporate additional factors such as age, education, religion, and language.

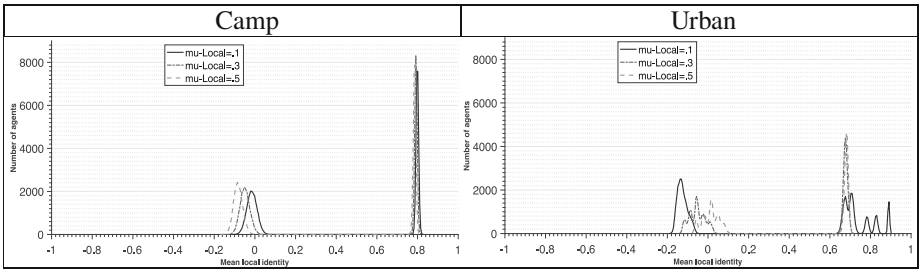


Fig. 1. Local identity of refugees by environment for varying μ_L

In Fig. 2, the cluster on the left represents data generated from normally distributed starting identity values; those on the right derive from uniformly distributed starting identities. Again, in the normally distributed cases, refugees tend to push in the opposite direction from strongly global identity values of 1. Preliminary fieldwork more closely supports the data depicted in the cluster on the right derived from uniformly distributed starting identities. In the uniform cluster, interaction with NGOs splits the refugees into two main identity groups, though very close on the identity scale. One group stays

centered around a neutral value of zero, while the other edges toward a slightly more global value. Note that the transition to a second grouping of identity values occurs after 5,000 timesteps and is very subtle in its shift toward a global identity. This is subtle even though we vary the number of NGO agents in the model widely to represent a high ratio at times of NGO worker to refugee. The visual difference between camp and urban refugees in Fig. 1 is more dramatic than that in Fig. 2 likely because of the constraints on refugees related to interactions. In both scenarios (camp, urban), refugees can interact with NGO workers who can move freely in and out of the camps. Refugees' interactions with locals, however, are limited to the camp in the first scenario, but freely available in the second which may account for the visual shift in identity distributions.

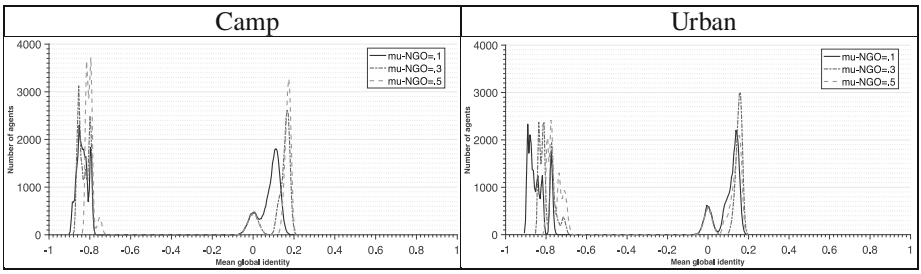


Fig. 2. Global identity of refugees by environment for varying μ_N

6 Conclusions

We proposed a model to capture theorized dynamics of the impact of protraction on refugee identity. The model is based on Jager and Amblard's algorithm that considers identity changes between two agents [27]. What the model demonstrates is that, with relatively minimal behavioral rules for interactions, there appears to be the potential for identities to evolve over long periods of time. Additionally, the data demonstrates that there is a high variability of outcome depending on the starting distribution of identity values. This indicates that future fieldwork must investigate underlying identity distributions as well as additional contributing factors such as age, education, religion, and age that may affect transition of identities over time. The uniform distribution in both cases more closely reflects fieldwork conducted among protracted refugees in Rwanda, but this may be a special case. The model suggests that, since starting distributions have such a dramatic effect on identity shift, fieldwork should incorporate multiple protracted contexts and cases as well as alternative explanatory factors.

In general, identities shift over time away from starting values and distributions. If reflective of real life, these shifts call into question both the idea of nationality and sovereignty in a modern, increasingly displaced world. Additionally, it begs even more critique of the UN's 'durable solutions' for refugee situations. As people live removed from their country of origin, and they shift their identity values toward that of a non-citizen, what will this mean for the prospects of repatriation or local integration? What can we do to capitalize on non-territorial identities and fold refugees more completely

into the modern economy? How would we adapt durable solutions to accommodate hyper-territorialized identities as those demonstrated with normally distributed starting values? These steps are necessary to ensure that refugees do not accumulate on the fringes of the international political space as 'wasted lives' [32].

The model presented here is part of a longer-term study to evaluate how data could be collected to calibrate and validate the model. Further work to collect data to estimate ranges for threshold values, staring identity distributions, and more realistic environments for agent interactions will be required to advance this work. Additionally, future models may consider alternative social forces on refugee identity including media or culture during the hosting and assimilation experiences of forced migrants.

Acknowledgements. This work is ongoing and represents collaborative efforts with David C. Earnest. The authors thank Hamdi Kavak for assistance with data visualization.

References

1. Edwards, A.: Global forced displacement hits records high. UNHCR News (2016)
2. UNHCR: UNHCR Global Trends: Forced Displacement in 2014 (2015)
3. Executive Committee of the High Commissioner's Programme Standing Committee: Protracted refugee situations. UNHCR, Geneva, Switzerland (2004)
4. Loescher, G., Milner, J.: Protracted Refugee Situations: Domestic and International Security Implications. Routledge for the International Institute of Strategic Studies, Abingdon (2005)
5. https://www.state.gov/j/prm/policyissues/issues/protracted/
6. Betts, A., Bloom, L., Kaplan, J., Omata, N.: Refugee Economies: Rethinking Popular Assumptions. University of Oxford, Oxford (2014)
7. Lischer, S.K.: Dangerous Sanctuaries: Refugee Camps, Civil war, and the Dilemmas of Humanitarian Aid. Cornell University Press, Ithaca (2006)
8. Shacknove, A.E.: Who is a refugee? Ethics **95**, 274–284 (1985)
9. Agamben, G.: We refugees In: Symposium, vol. 49 (1995)
10. Barnett, M.: Humanitarianism with a sovereign face: UNHCR in the global undertow. Int. Migr. Rev. **35**, 244–277 (2001)
11. Anderson, B.: Imagined Communities: Reflections on the Origin and Spread of Nationalism. Verso, New York (1991)
12. Malkki, L.: Purity and Exile: Violence, Memory, and National Cosmology Among Hutu Refugees in Tanzania. University of Chicago Press, Chicago (1995)
13. Zetter, R.: Labelling refugees: forming and transforming a bureaucratic identity. J. Refug. Stud. **4**, 39–62 (1991)
14. Lustick, I.S., Miodownik, D.: Abstractions, ensembles, and virtualizations simplicity and complexity in agent-based modeling. Comput. Polit. **41**, 223–244 (2009)
15. Rousseau, D., van der Veen, A.M.: The emergence of a shared identity: an agent-based computer simulation of idea diffusion. J. Confl. Resolut. **49**, 686–712 (2005)
16. Smaldino, P., Pickett, C., Sherman, J., Schank, J.: An agent-based model of social identity dynamics. J. Artif. Soc. Soc. Simul. **15**, 7 (2012)
17. Kuran, T.: Now out of never: the element of surprise in the east European revolution of 1989. World Polit. **44**, 7–48 (1991)
18. Ring, J.: An Agent-Based Model of International Norm Diffusion. University of Iowa, Iowa City (2014)

19. Lustick, I.S.: Agent-based modeling of collective identity: testing constructivist theory. J. Artif. Soc. Soc. Simul. **3** (2000)
20. Lustick, I.S.: Secession of the center: a virtual probe of the prospects for Punjabi secessionism in Pakistan and the secession of Punjabistan. J. Artif. Soc. Soc. Simul. **14** (2011)
21. Miodownik, D.: Cultural differences and economic incentives: an agent-based study of their impact on the emergence of regional autonomy movements. J. Artif. Soc. Soc. Simul. **9** (2006)
22. Miodownik, D., Cartrite, B.: Does political decentralization exacerbate or ameliorate ethnopolitical mobilization? A test of contesting propositions. Polit. Res. Q. **63**, 731–746 (2010)
23. Yamamoto, K.: Mobilization, flexibility of identity, and ethnic cleavage. J. Artif. Soc. Soc. Simul. **18**, 8 (2015)
24. Edwards, S.: Computational tools in predicting and assessing forced migration. J. Refug. Stud. **21**, 347–359 (2008)
25. Frydenlund, E., Padilla, J.J., Earnest, D.C.: A theoretical model of identity shift in protracted refugee situations. In: Spring Simulation Multi-Conference 2017. Springer (Year)
26. Turner, S.: Politics of Innocence: Hutu Identity, Conflict and Camp Life. Berghahn Books, New York (2010)
27. Jager, W., Amblard, F.: Uniformity, bipolarization and pluriformity captured as generic stylized behavior with an agent-based simulation model of attitude change. Comput. Math. Organ. Theory **10**, 295–303 (2005)
28. Wilensky, U.: NetLogo. In: Modeling, C.f.C.L.a.C.-B. (ed.). Northwestern University, Evanston, IL (1999)
29. Axtell, R., Axelrod, R., Epstein, J.M., Cohen, M.D.: Aligning simulation models: a case study and results. Comput. Math. Organ. Theory **1**, 123–141 (1996)
30. Miller, J.H.: Active nonlinear tests (ANTs) of complex simulation models. Manag. Sci. **44**, 820–830 (1998)
31. Stonedahl, F., Wilensky, U.: BehaviorSearch. Northwestern University (2013)
32. Bauman, Z.: Wasted Lives: Modernity and its Outcasts. Polity, Malden (2004)

Linking Twitter Sentiment and Event Data to Monitor Public Opinion of Geopolitical Developments and Trends

Lucas A. Overbey, Scott C. Batson(✉), Jamie Lyle, Christopher Williams, Robert Regal, and Lakeisha Williams

Space and Naval Warfare Systems Center Atlantic,
P.O. Box 190022, North Charleston, SC 29419-9022, USA
{lucas.overbey,scott.batson}@navy.mil

Abstract. Readily observable communications found on Internet social media sites can play a prominent role in spreading information which, when accompanied by subjective statements, can indicate public sentiment and perception. A key component to understanding public opinion is extraction of the aspect toward which sentiment is directed. As a result of message size limitations, Twitter users often share their opinion on events described in linked news stories that they find interesting. Therefore, a natural language analysis of the linked news stories may provide useful information that connects the Twitter-expressed sentiment to its aspect. Our goal is to monitor sentiment towards political actors by evaluating Twitter messages with linked event code data. We introduce a novel link-following approach to automate this process and correlate sentiment-bearing Twitter messages with aspect found in connected news articles. We compare multiple topic extraction approaches based on the information provided in the event codes, including the Goldstein scale, a simple decision tree model, and spin-glass graph clustering. We find that while Goldstein scale is uncorrelated with public sentiment, graph-based event coding schemes can effectively provide useful and nuanced information about the primary topics in a Twitter dataset.

1 Introduction

The use of publicly accessible Internet social networking sites (social media) now occupy a key role in the spread of information. Often this spread of information is paired with affective statements indicating sentiment [1]. General public sentiment within a region can change with events that effect its population. In particular, recent international events such as natural disasters, the cascade of protests and revolutions in the Arab world, and terror attacks have uncovered the utility of social networking sites for understanding social and political unrest. Effectively monitoring changes in sentiment toward political events may reveal trends in public opinion that indicate social and political instability.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 223–229, 2017.
DOI: 10.1007/978-3-319-60240-0_26

A key component to understanding public opinion is extraction of the aspect toward which a given sentiment is directed. A metric to quantify the sentiment that social media users express towards political actors and/or politically relevant events would be useful to measure how much these actors or events influence affected populations. Twitter is a micro-blogging service that limits messages to 140 characters in length. Because of this limitation, the aspect of directed sentiment is often contextually embedded in Uniform Resource Locator (URL) links to other sources of information. For example, Twitter users commonly share an opinion on events described in linked news stories.

Open source automated political event coding schemes and datasets provide information about interactions between state and non-state actors in online news stories. The Open Event Data Alliance's Phoenix dataset [2] relies on the open source Python Engine for Text Resolution and Related Coding Hierarchy (PETRARCH) project, an automated coding library utilizing natural language processing (NLP), a lexical reference consisting of event/actor ontologies and verb/noun phrase dictionaries, and the Conflict and Mediation Event Observations (CAMEO) [3] coding scheme, to output source/target actors, an event that took place between actors, a location, and date of the event. These datasets have been studied to forecast political instability, identify trends such as escalation of conflict between actors, and monitor interaction between countries.

We develop an innovative link-following approach to automate the process of identifying sentiment and aspect in contextually ambiguous tweets by correlating sentiment-bearing Twitter messages with aspect found in shared news articles. We compare multiple topic identification approaches based on information provided in the event codes, including the Goldstein scale [4], a simple decision tree model [5], and spin-glass graph clustering [6]. We find that while Goldstein scale is uncorrelated with public sentiment, graph-based topic modeling does yield aspects directly tied to the tweets themselves, with information that cannot be gleaned from automated content analysis of the tweets alone. Correlating extracted topics with corresponding measures of public opinion within the tweets themselves may help identify trends for studying and predicting social and political changes.

2 Methodology

To develop an approach for correlating social media and political event data, we used a self-collected dataset from the Twitter streaming Application Programming Interface (API). We used twenty-five broad English hashtag-based search terms related to European governmental organizations and known events to collect Twitter data. Using broad, geopolitically relevant terms allows us to collect a large dataset likely containing a substantial number of URL links to politically relevant news stories. The search terms were chosen to facilitate manual grouping of tweets into one of four manually chosen topics: (1) EU Referendum in the UK (Brexit), (2) Migrant Crisis, (3) the North Atlantic Treaty Organization (NATO), and (4) Russia. We also allowed for an "Other" topic, encompassing

Table 1. Event distributions by ground truth topic

	Brexit	Migrant crisis	NATO	Russia	Other
Tweets with extracted events	1,582	410	338	981	744
Unique actor/event dyads	409	218	186	475	434
Unique actor dyads	166	132	85	202	267

anything not easily classified in the previous categories. 123,649 total tweets were collected from April 26, 2016 through May 23, 2016. To identify sentiment of the tweets, we employed the lexical approach introduced by Musto et al. (details found in [7]). As this approach relies on English, we used a logistic regression model across infinite-length character n-grams as language detection [8] to filter out non-English tweets, leaving 94,570 of the 123,649 collected tweets.

We then determined English tweets contained URL links, and followed these links to trusted news outlets with a web scraper. PETRARCH [2] was applied to the retrieved articles to extract geopolitical events with automated coding. From the linked URLs, we identified 4,323 unique events. Several tweets referenced the same news articles, yielding 10,287 English tweets with a corresponding event extracted from the linked URL. Of those, we found 4,055 were identified as subjective messages (a non-neutral sentiment score, heuristically $>|0.5|$). The event coding scheme yields actor dyads, CAMEO-based event type codes [3], and a corresponding Goldstein scale [4]. The Goldstein scale represents an ordinal measure ranging from extreme conflict (-10.0) to extreme cooperation (8.3). If sentiment were directly tied to the degree of conflict or cooperation, one could expect that negative sentiment corresponds to a high degree of conflict and positive sentiment corresponds to a high degree of cooperation.

The actor dyads represent the two primary state or non-state actors involved in the extracted event, each containing a country code with up to two role codes. We used the combined country code and secondary role code of the CAMEO actor dictionary to define distinct actors. The set of dyads can be represented by a directed graph, where the edge weight can be either sentiment or Goldstein scale. We utilized multiple reviewers to form a consensus manual, "ground truth" label for each tweet by one of the five topics (ignoring event codes but reviewing tweets and corresponding links). We then compared these labels to those generated by automated topic modeling approaches using the event graph dyads.

First, we employed a simple, supervised decision tree model [5]. We used five-fold cross-validation to perform training and testing, with individual actors and actor dyads as features, and a conditional inference based framework [9] for training. Alternatively, we applied an unsupervised graph clustering approach to identify individual topics based on the graph constructed from actor dyads. We employed the spin-glass graph clustering algorithm [6] based on the Potts model [10] of interacting spins on a crystalline lattice. Advantages to this graph clustering approach are that it naturally leads to known modularity measures [11] and allows for both directed and weighted networks.

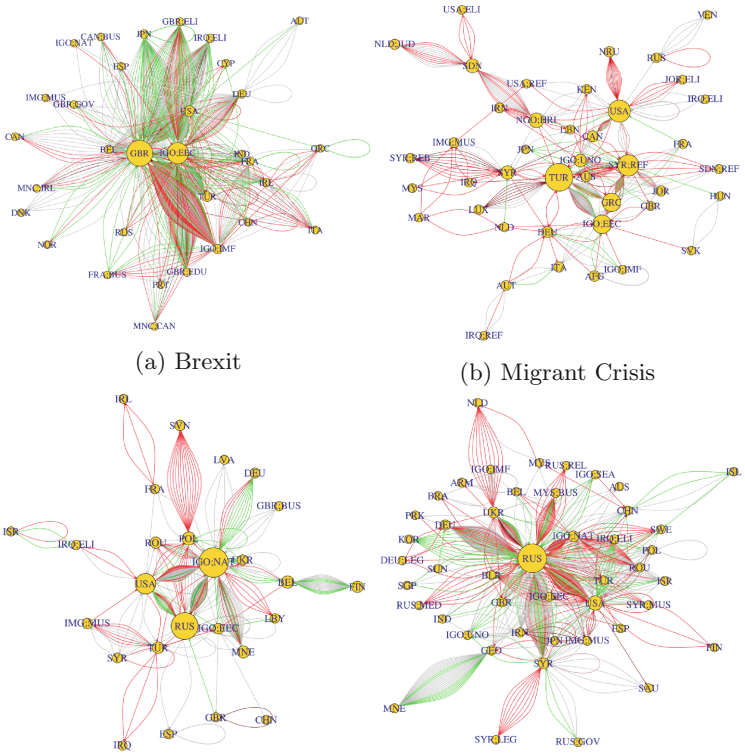


Fig. 1. Actor graphs by topic, with edges colored by five sentiment categories from very negative (dark red) to very positive (dark green) (Color figure online)

3 Results

A breakdown of the 4,055 tweets by manually labeled topic are shown in Table 1. The *Brexit* and *Russia* topics make up a majority of the messages (39.0% and 24.2%, respectively). The large number of *Brexit* tweets is a result of the immediacy of the Referendum vote taking place in the United Kingdom and the English language restriction. The *Russia* topic was fairly broad and covered a wide variety of subtopics, including the interaction between NATO and Russia, which made manual labeling of the "ground truth" data slightly subjective. Tweets categorized as *Migrant Crisis* also covered a range of subtopics corresponding to events taking place in Syria, Greece, Sudan, Iraq, and Nauru, among others. Table 1 also shows the unique actor/event dyads and actor dyads (irrespective of event type) in the event coded news articles from extracted tweet links.

The Pearson product-moment correlation between sentiment and Goldstein scale was 0.094, not indicative of a strong correlation. Hence, the amount of conflict or cooperation of the events does not reflect the public opinion regarding

how positively or negatively they feel about the event. Someone may be pleased about conflict and angry about cooperation, or vice versa. The two measures provide distinct and useful information; however, one cannot be substituted for the other for analyses.

The resulting event subgraphs (not including the *Other* category) are displayed in Fig. 1. The CAMEO codes for geopolitical actors are the node labels, with node size reflecting the frequency that an actor appeared in the dataset. Edges represent individual tweets with links carrying a resultant event code and are colored by the sign of the sentiment. Several observations can be made from these graphs, including the centralization of *Brexit* (around the United Kingdom and European Union), decentralization of *Migrant Crisis* (composed of several subtopics), complexity of *Russia*, and the interdependence of *Russia* and *NATO*. Precision, recall, and f-measure for the supervised decision tree model are provided in Table 2a. The total accuracy for the model was 82.4%. The model is fairly accurate at identifying topics using the actor dyad information from the extracted event codes alone, without using any tweet content.

Table 2. Precision, recall, and f-measure metrics for topic modeling

(a) Decision tree classification

Class	Precision	Recall	f_1
Brexit	0.88	0.96	0.91
Migrant Crisis	0.75	0.81	0.78
NATO	0.72	0.56	0.63
Russia	0.81	0.84	0.83
Other	0.79	0.62	0.70
Macro-Avg Total	0.79	0.76	0.77

(b) Spin-glass graph clustering

Class	Precision	Recall	f_1
Brexit	0.92	0.94	0.93
Migrant Crisis	0.70	0.50	0.58
NATO / Russia	0.89	0.95	0.92
Macro-Avg Total	0.84	0.72	0.81

Spin-glass clustering resulted in ten total topics found with the maximum number of communities set at 20. Four of the generatively derived topics could be easily identified as representative of a *Brexit* cluster, a *Russia/NATO* combined cluster, and two separate clusters around *Migrant* issues. The remaining clusters could all be classified as part of the manually labeled *Other* category, with node sizes no greater than 11. As mentioned previously, the *Russia* and *NATO* labeled topics were interdependent, so it is not surprising that they were clustered together. Furthermore, the subtopics within the *Migrant Crisis* could easily be part of different topic areas if not forced into the same one through pre-determined classes. This unsupervised approach may actually produce topical breakdowns that are more indicative of the data, as the chosen topic areas were broad and with unrepresented complexities in the labels.

To evaluate the accuracy of the unsupervised algorithm, we distilled the topic classes into three categories (*Brexit*, *NATO/Russia*, and *Migrant Crisis*), as the former *NATO* and *Russia* topics are intertwined and the nuance of the individual

clusters found by the spin-glass algorithm are lost in the manually labeled *Other* category. The total accuracy for the model was 88.7%. Precision, recall, and f-measure metrics are provided in Table 2b. The supervised and unsupervised event-topic modeling approaches yield somewhat similar prediction capabilities. The spin-glass method provides the added advantage of using the natural structure of the manifolds in the data itself rather than forced manifolds based on the broad manual labels.

4 Conclusion

The brevity of tweets makes identifying sentiment aspect in Twitter data difficult. However, many geopolitically-relevant tweets expressing opinion have links to news articles with more information about the topic and/or event(s). Event coding combined with topic modeling can be used to inform public sentiment aspect. This form of sentiment aspect may be used to identify the political actors involved as identified through event coding. An unsupervised graph clustering approach provides accurate results, with the added benefits of not requiring training data and an ability to capture nuance relationships between topics that may be missed by subjective manual labels. Both of these approaches were more successful than traditional topic modeling approaches such as latent Dirichlet allocation (LDA) [12], which was unable to achieve a macro-averaged f_1 score greater than 0.5 using the tweet content, the actor dyads, or the original linked news article content.

References

1. Ringsquandl, M., Petkovic, D.: Analyzing political sentiment on Twitter. In: AAAI Spring Symposium (2013)
2. Schrodt, P., Beieler, J., Idris, M.: Three's a Charm? Open event data coding with EL: DIABLO, PETRARCH, and the open event data alliance. In: International Studies Association Meeting (2014)
3. Schrodt, P.A.: CAMEO: conflict and mediation event observations event and actor codebook. Pennsylvania State University, March 2012
4. Goldstein, J.S.: A conflict-cooperation scale for WEIS events data. J. Confl. Resolut. **36**(2), 369–385 (1992)
5. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
6. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**, 016110 (2006)
7. Musto, C., Semararo, G., Polignano, M.: A comparison of lexicon-based approaches for sentiment analysis of microblog posts. In: 8th International Workshop on Information Filtering and Retrieval (2014)
8. Shuyo, N.: Short text language detection with infinity-gram. NAIST Seminar (2012)
9. Hothorn, T., Hornik, K., Zeileis, A.: Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. **15**(3), 651–674 (2006)
10. Wu, F.-Y.: The Potts model. Rev. Mod. Phys. **54**, 235 (1982)

11. Newman, M.E.J., Girvan, M.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. **99**, 7821–7826 (2002)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(4–5), 993–1022 (2003)

Identifying Smoking from Smartphone Sensor Data and Multivariate Hidden Markov Models

Yang Qin(✉), Weicheng Qian, Narjes Shojaati, and Nathaniel Osgood

University of Saskatchewan, Saskatoon, Canada
{yang.qin,weicheng.qian,narjes.shojaati,nathaniel.osgood}@usask.ca

Abstract. Smoking is one of the foremost public health threats listed by the World Health Organization, and surveillance is a key to informing effective policies. High smartphone penetration and mature smartphone sensor data collecting techniques make smartphone sensor data based smoking monitoring viable, yet an effective classification algorithm remains elusive. In this paper, we sought to classify smoking using multivariate Hidden Markov models (HMMs) informed by binned time-series of transformed sensor data collected with smartphone-based Wi-Fi, GPS, and accelerometer sensors. Our model is trained on smartphone sensor time series data labeled with self-reported smoking periods. Two-fold cross-validation shows A_z (area under receiver operating characteristic curve) for HMMs using five features = (0.52, 0.84). Comparison of univariate HMMs and multivariate HMMs, suggests a high accuracy of multivariate HMMs for smoking periods classification.

Keywords: Hidden Markov model · Smartphone sensor data · Tobacco · Smoking monitoring

1 Introduction

Smoking is one of the biggest public health threats listed by the World Health Organization. Effective tools for smoking recognition can ensure public health surveillance for policy making [11], provide early detection before addiction [3], and aid former smokers to avoid relapse. Detection of smoking status has long relied on biomedical assays based around detection of substances such as cotinine, nicotine [8], carbon monoxide [6] and respiration [2]. Application of such assays normally requires mildly to moderately invasive measurements, from breath tests to provision of saliva to clippings of hair, and many test results are available only after delays measured in days or more.

Studies using wearable sensors for recognition of smoking [1,9] have showed potentials to avoid the invasiveness of measurements and delays of test results to perform seamless online detection. These studies predominately based on hand-to-mouth gestures and breathing pattern and using specialized hardware to collect data, which can be costly and hard to comply continuously, informativeness from other aspects correlates of smoking have not yet been considered, such as

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 230–235, 2017.
DOI: 10.1007/978-3-319-60240-0_27

presence outdoors (as required by regional regulations) or designated smoking areas, activity levels, and characteristic length of dwelling period correlated to the burning time of cigarettes.

In recent years, and paralleling their rapid penetration across diverse strata of society worldwide, smartphone have become an attractive platform for sensor-based data collection on human behaviour. The use of such techniques has been enhanced by the growing maturity of data collection apps (such as the iEpi system [5], UPenn's DREAM project) that make smartphone sensor data a highly available and easily accessible data source for many studies [7]. Feasibility studies on using accelerometer sensor to detect smoking behaviors has been initiated [10], but published studies on fusing sensor data available on smartphone remains absent.

In this paper, we fused various types of sensor data commonly available on smartphone, after considering data completeness, accuracy and informativeness, examined the effects of five transformations for GPS, Wi-Fi and accelerometer sensor data, and applied multivariate hidden Markov model (HMMs) to classify periods to recognize whether smoking was taking place. Finally, we investigated the performance of univariate HMMs and multivariate HMMs, and the impact of tailoring the training and test set to preserve entire smoking cycles.

2 Data Processing and Algorithm

Dataset Description. Data used in the project came from a previously conducted Behavioural Ethics Board-approved study that collected multiple types of sensor data together with self-reported ground truth on smoking behavior by four participants (one did not complete) who were self-reported smokers. The dataset contains labeled data on segments of intervals of smoking and non-smoking periods. The sensor data was collected with a five-minute duty cycle by Ethica system [4,5] for three participants over one month from April 04, 2015 to May 12, 2015. There are 36 million records from gyroscope, 0.3 million records for location, 1.9 million records for Wi-Fi and 36 million records for accelerometer.

Data Processing. We found each participant has an extremely long smoking period at the end of their self-labeled smoking-nonsmoking periods, which are apparently outliers and therefore we preserved only a period at the head of each of those extremely long periods, whose length equals to the average length of previous smoking periods of this very person, and trimmed the rest smoking periods at the end.

For accelerometer data, we applied a high-pass filter, using a standard deviation of norms of readings on X, Y, Z accelerometer axes in the 30 s timeslot, to separate out the dominant invariant gravitational component. For Wi-Fi data, Received signal strength indication (RSSI) in 30-second timeslots was considered. ECDF of the counts of unique MAC address during 30 s timeslots showed better difference between two states than that of maximum RSSI. Maximum

RSSI indicated the strongest Wi-Fi signal, and counts of unique MAC address represented number of accessible networks for the smartphone. The source of location data, either GPS (using satellite) or network (using cell tower and Wi-Fi based location), indicates whether the participant is indoor or outdoor. So the count of GPS readings across 30 s timeslots specifically drawn from satellite (as opposed to network) sources was used.

Multivariate HMM. Using the transformed data described above, a multivariate HMMs was employed to classify smoking and non-smoking intervals based on real world labeled observations. In this model, each state has multiple observations corresponding to readings from Wi-Fi, accelerometer and GPS sensors. The probabilities of observations follow empirical distributions, and observations are assumed to be independent from each other, conditional on being in a given state. So the likelihood of observing a given vector of observed quantities was approximated as the product of independent probability density functions as given by kernel density estimates.

Two-Fold Cross Validation. Hidden Markov models expect sequential observations, therefore when choosing training set and test set, we can not simply sample at random time intervals from data sequence, but rather need to divide the data sequence into disjoint contiguous sequences. Firstly, we made use of a two-fold cross-validation approach, where we cut the sequence from the head to 50%, 55%, 60%, 65%, 70% and 75% of the sequence to ensure sequential observations (including NAs) for training, and used the balance of the observations for testing. Second, We swapped the training set and test set in first step to feed the HMMs.

3 Results

Structured learning was used in this project. This work was conducted in several phases. In phase 1, maximum RSSI, counts of unique MAC address during 30 s timeslots for Wi-Fi, average norms, standard deviation of norms for accelerometer and counts of GPS reading from GPS source in 30 s timeslots were considered as a single feature, respectively. Each of the five features was used to train univariate HMMs, which were then evaluated. In phase 2, Multivariate HMMs using three features and five features were trained and evaluated.

The HMMs were found to yield favorable results in the multivariate cases and in univariate cases considering accel sensor data as feature. As shown below, multivariate HMMs exhibit accuracy over 0.9, and an area under ROC curve (AUC) above 0.8 when collected data is representative. The results of HMMs with a single feature are less favorable than those for multivariate HMMs.

3.1 Results with Single Feature

For using average of norm of the accelerometer as a feature, the AUC for training set and test set with different size of training set range from 0.69 to 0.94 and

from 0.60 to 0.79, respectively. And the error rates of the training set and test set range from 0.096 to 0.27 and from 0.057 to 0.357, respectively. For using of standard deviation of norm of the accelerometer as a feature, the AUC for training set and test set with different size of training set range from 0.76 to 0.92 and from 0.63 to 0.86, respectively. And the error rates for the training set and test set range from 0.05 to 0.12 and from 0.06 to 0.2, respectively, as shown in Fig. 1.

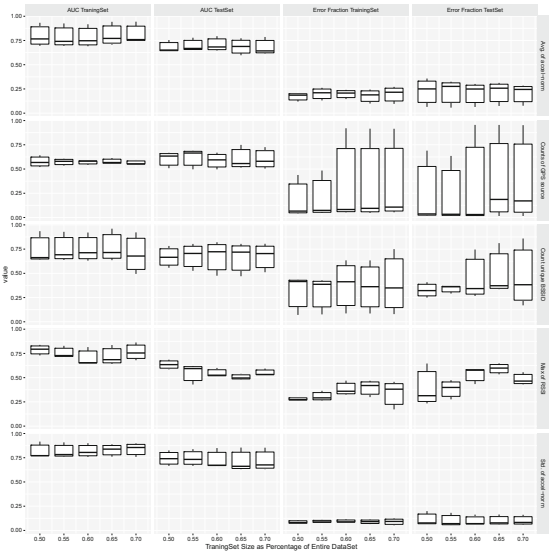


Fig. 1. AUC and error rate of HMMs using avg. and std. of accel-norm, count of GPS source, count of unique BSSID and max of RSSI as single feature

For using readings from Wi-Fi sensor as single feature (maximum RSSI or counts of unique MAC address), for maximum RSSI, the range of AUC for test set is from 0.43 to 0.69, and the range of error rate for test set is from 0.23 to 0.65. For counts of unique MAC address, the range of AUC and error rate for test set ranges from 0.47 to 0.82 and from 0.17 to 0.86, respectively.

In this model, using only one feature derived from the GPS sensor, the range of AUC for training set is from 0.52 to 0.64, for test set is from 0.50 to 0.75. Error rates for training set and test rate are from 0.04 to 0.92, and from 0.015 to 0.96, respectively.

3.2 Results with Three Features

Counts of unique MAC address for Wi-Fi, standard deviation of accel norm and counts of GPS sourced signal were employed together to train the HMMs. The results offer AUC and error rates of the HMMs were shown in Fig. 2. The range

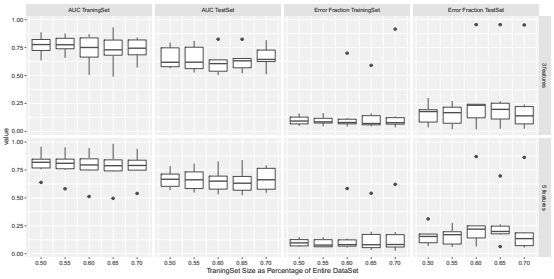


Fig. 2. AUC and error rate of HMMs with three features and five features

of AUC for the test set is from 0.5 to 0.83 with an average of 0.65, and error rate for test set ranged from 0.017 to 0.96 with average of 0.23.

3.3 Results with Five Features

All five features derived from sensors were employed together to train HMMs. The results for AUC and error rates of the five feature HMMs were also shown in Fig. 2. The range of AUC for training set is from 0.5 to 0.98 with average 0.79, and for test set, is from 0.52 to 0.84 with average 0.66. The Wi-Fi and GPS data are location-based features, while the components of the accelerometer data are associated with the body gestures and orientation features of participant. The combination of five features can capture a larger set of information on the current smoking activity of participants. This enlarged information can in turn enhance performance of the HMMs.

4 Related Work

Sazonov et al. (2013) developed a wearable sensor system based on hand-to-mouth smoking gestures and breathing pattern [9], Lopez-Meyer et al. (2013) further applied a support vector machine and achieved 87% and 80% of average user-independent precision and recall and 90% in user-independent precision and recall [1]. Scholl and Laerhoven (2012) used wearable accelerometer device to collect data and applied basic Gaussian classifier to detect smoking gestures with a precision of 51.2% and 70% of user specific recall [10]. We have not yet found papers about smoking detection using multivariate HMM based on various types of sensor data available from commodity smartphones.

5 Limitations and Future Work

Despite high granularity data for each participant, our study is limited by the number of participants, as a future work, we will experiment with a larger participant size. We will also consider covariance among sensor observations for performance boost and whether commonalities among personal empirical distributions can be extracted and reused on other persons.

6 Conclusions

The results of multivariate HMMs demonstrated classification and detection of smoking activity with high accuracy. Compared to single feature HMMs, the multivariate HMMs had higher accuracy, because additional types of sensor data can help us better describe smoking gesture and activity.

This work further suggests that tailoring training set and test set close to entire smoking cycles can improve the performance of HMMs. The large variation in results across participants further raises the possibility that significant components of remaining error rates may be due to limitations in the accuracy of self-reporting of ground truth data on smoking behaviour.

References

1. Lopez-Meyer, P., Tiffany, S., Patil, Y., Sazonov, E.: Monitoring of cigarette smoking using wearable sensors and support vector machines. IEEE Trans. Biomed. Eng. **60**(7), 1867–1872 (2013)
2. Ali, A., Hossain, S., Hovsepian, K., Rahman, M., Plarre, K., Kumar, S.: mPuff: Automated detection of cigarette smoking puffs from respiration measurements. In: IPSN 2012 - Proceedings of the 11th International Conference on Information Processing in Sensor Networks, pp. 269–280 (2012)
3. Community Preventive Services Task Force: Reducing tobacco use and secondhand smoke exposure: mobile phone-based cessation interventions (2013)
4. Ethica Data. https://www.ethicadata.com/
5. Hashemian, M., Knowles, D., Calver, J., Qian, W., Bullock, M.C., Bell, S., Mandryk, R.L., Osgood, N., Stanley, K.G.: iEpi: an end to end solution for collecting, conditioning and utilizing epidemiologically relevant data. In: Proceedings of the 2nd ACM International Workshop on Pervasive Wireless Healthcare, pp. 3–8. ACM (2012)
6. Meredith, S.E., Robinson, A., Erb, P., Spieler, C.A., Klugman, N., Dutta, P., Dallery, J.: A mobile-phone-based breath carbon monoxide meter to detect cigarette smoking. Nicotine Tob. Res. **16**(6), 766–773 (2014)
7. Qian, W., Stanley, K.G., Osgood, N.D.: The impact of spatial resolution and representation on human mobility predictability. In: Liang, S.H.L., Wang, X., Claramunt, C. (eds.) W2GIS 2013. LNCS, vol. 7820, pp. 25–40. Springer, Heidelberg (2013). doi:10.1007/978-3-642-37087-8_3
8. Raja, M.: Diagnostic methods for detection of cotinine level in tobacco users: a review. J. Clin. Diagn. Res. **10**(3), 4–6 (2016)
9. Sazonov, E., Lopez-Meyer, P., Tiffany, S.: A wearable sensor system for monitoring cigarette smoking. J. Stud. Alcohol Drugs **74**(6), 956–964 (2013)
10. Scholl, P.M., van Laerhoven, K.: A feasibility study of wrist-worn accelerometer based detection of smoking habits. In: 2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 886–891 (2012)
11. WHO: Tobacco Factsheet (2016). http://www.who.int/mediacentre/factsheets/fs339/

Is Word Adoption a Grassroots Process? An Analysis of Reddit Communities

Jeremy R. Cole(✉), Moojan Ghafurian, and David Reitter

The Pennsylvania State University, University Park, PA, USA
{jrcole,moojan,reitter}@psu.edu

Abstract. This study examines how novel words originate in and disperse through online communities. It asks whether larger numbers of people or closer social ties are better environments to foster the adoption of new words. The data stem from Reddit, a large sample of web-mediated, asynchronous conversations. Reddit communities are divided by size in this study: larger communities are based on discussing general topics and have weak social ties, whereas small communities are based on discussing specific topics and have strong social ties. The analysis shows that the majority of new words are created/first adopted in larger communities.

Keywords: Word adoption · Dispersion · Reddit · Community organization

1 Introduction

Language is a communication system that varies among speakers and is constantly changing. English is a particularly productive language: new words are invented frequently. In fact, the rate of new word formation has increased in the past century [6]. Newly introduced words might be used for only a short period of time or may last longer and contribute to large-scale language change. This process relies on speakers taking liberties with their word choice and on speaker communities that facilitate and accept the use of novel words.

Experimentally, lexical change has been studied in relatively small groups: for example, using the *naming game* paradigm. Such experiments show how new words can be created and will emerge as an agreed-upon standard in these small, artificial and temporary communities that are created for that purpose. The naming game can be seen as a model of how communities reach consensus about a communication system, naming, or generalized linguistic systems [3]. In most communities, the members successfully reach consensus in such games [4].

There are individual and group differences in the task. For instance, such differences could lead to some migratory speakers adopting new words more than others. This result was found by comparing old and new profession names in census data [9]. Further, mixing up group composition can increase the quality of the communication systems produced [4]. However, even relatively stationary members of these communities can adopt new ideas and words.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 236–241, 2017.
DOI: 10.1007/978-3-319-60240-0_28

Group size also influences the convergence of communication systems. As Dall'Asta et al. [3] discuss, the time of convergence increases as the population size increases. According to this model, smaller groups are expected to converge faster (thus agree on new words faster) than larger ones. However, in the context of a social network, more centralized distribution leads to faster adoption [10]. Can both be true?

Twitter has also been used to study word adoption in the context of larger networks. The dispersion dynamics of hashtags can be surprising in that they appear to be different depending on the topic [8]. Still, the interaction of those broadcasting tweets in a public forum is not necessarily a good model of the directed communication of communities: users have weak social ties with each other, and they are not naturally partitioned in such a way as to study the nature of the communities that facilitate innovation and early adoption of new ideas or words.

In naming games the focus is on language as a shared community contract (e.g., [4,7]); however, it does not discuss interactions across communities and its framework limits the possible number of participants. In this study, we focus on the conditions that facilitate the adoption of novel words within and across larger communities. As we will see, group size in the communities we study is, surprisingly, correlated with more rather than less innovation. As we address the features of communities that facilitate creativity or early adoption of new concepts, we examine data source stemming from delineated, but connected communities.

To explore the effect of community size on the rate of adoption of novel words, we analyze the *Reddit* dataset, as it provides us with a variety of groups of different sizes. The question we ask regarding the communities relate to their size and specificity: are words first adopted in more specific, smaller communities on Reddit and dispersed to the larger, more general communities, or does dispersion take place in the opposite direction? We think that answering this question could give valuable cues as to whether and how group size promotes linguistic and thematic innovation.

2 Methods

2.1 Data: Reddit Corpus

Our data set consists of approximately 426 GB of Reddit data, ranging from the year 2012 to the year 2014.

Reddit.com is a community-driven news aggregation website that mostly contains discussions and ratings [2]. Communities on Reddit are divided into different categories (such as Education, Technology, image sharing, etc.), and each of these general categories are then divided into several subcategories, which are called *subreddits*. For example, Educational category can have science as a subreddit. Each of these subreddits can be about any given topic, general or specific. For instance, it is possible to have subreddits about science, biology,

and genetics. These three subreddits can be completely independent, as they are not organized hierarchically.

Before applying any of our analysis, we filter out comments in subreddits with a small number of users. As anyone can make a subreddit and invite their friends to join, we wanted to avoid small subreddits that may more closely resemble social networks than communities.

2.2 Defining New Words

As discussed, our data spans 2012–2014. In this sense, we came up with a simple way to determine if a word is new: its time of first use was more than a year after the first comment for the data we used. We also excluded words with a small number of uses or that did not consist entirely of alphabetic characters. Some example words can be found in Table 1. In total, we found 3550 words matching these criteria.

Table 1. On the left, there are sample words with their originating subreddit, along with an example destination subreddit. On the right, there are example pairs of sub-subreddits and supersubreddits.

Word	First adopter	Later adopter	Subsubreddit	Supersubreddit
dogetips	dogecoin	funny	UniversityOfHouston	houston
isanderkirby	AdviceAnimals	AskReddit	justneckbeardthings	fatpeoplestories
peshka	gaming	Warthunder	simpleios	iOSProgramming
gamecribs	leagueoflegends	counterstrike	DotaCR	DotA2
squarecash	economy	Bitcoin	lisp	programming

These words are not entirely recognizable to those not in the culture, and they vary in their novelty. For instance, Square Cash, a financial product, is frequently referred to as *squarecash* by Reddit users. Others, however, are strictly adoptions, such as Chromecast. Thus, we will refer to these as *first adoption* events.

While some of the first adoption events are origination events, all of them are a discussion of something new. The first discussion of a new idea has social consequence. In Reddit, people receive both explicit and implicit rewards for social acceptance, through the curation mechanism. Thus, first adoptions will likely occur in communities that maximize this payoff. Thus, we seek to determine what type of community that is. We will use *adoption* to refer to any usage of a new word by a subreddit, using origination or first adoption for the first subreddit to adopt it, and *later adoption* for later usages.

2.3 Inducing the Structure of Reddit Subforums

Reddit's inherent organization is shallow, rather than hierarchical. Underneath the top-level hierarchy, subreddits are not formally organized. Still, the topics

have a range of specificities. For instance, there could be a subreddit focused on board games in general, with a separate subreddit for specific board games, such as Settlers of Catan or Monopoly. Our intuition suggests then, that people passionate about Monopoly are also passionate about board games in general.

We then define a *subsubreddit* and a *supersubreddit* as such: A is a subsubreddit of B if at least $N\%$ of A's members are also members of B. B is a supersubreddit of A if and only if A is a subsubreddit of B. We used $N = 25\%$, because for this value, for any subsubreddit, A of B, A was not also a supersubreddit of B. This resulted in approximately four thousand pairs of subreddits. In general, the supersubreddits will cover more general topics and involve more users, while the subsubreddits will cover more narrow topics and have fewer users. Some example pairs can be found in Table 1. As an important point, this relationship is ultimately relative. Lastly, this pairing is somewhat conservative: not all subreddits are in any pairing. Even larger subreddits were only in a relationship with intuitive relations: AskReddit's only relationship was with TrueAskReddit.

Still, these pairs are largely interesting in the context of our research question: which types of communities lead to first adoption, rather than later adoption?

2.4 Results

The results of all of the Wilcoxon Signed-Rank Tests are found in Table 2. Our first analysis focused on first adoptions that were later adopted by the other in the pair (Paired Adoptions). For origination events, this could be an adoption from the previous subreddit, rather than an external source. Regardless, it reflects words that are of interest to both subreddits in the pair, to examine how such words move through subreddits around that topic. We find that supersubreddits have significantly more first adoptions than the subsubreddits.

Our second analysis focuses on the total numbers (Full). While more words may originate in any given super-subreddit than a sub-subreddit, it is possible that the totals tell a different story. In this analysis, we maintain the pairings, but instead look at the total number of words that originated in that subreddit and

Table 2. The results of the Wilcox Signed Rank Tests, Norm(U) is normalized by number of users, while Norm(C) is normalized by number of comments

	V	Mean-sub	Mean-super	p-value
Paired adoptions	8263.5	0.5083	4.9472	<0.0001
Full first adoptions	32158	0.3632	6.1997	<0.0001
Full later adoptions	121350	6.9978	41.9935	<0.0001
Norm(U) first adoptions	95613	0.0001	0.0004	<0.0001
Norm(U) later adoptions	850780	0.0052	0.0030	0.7078
Norm(C) first adoptions	114840	0.00001	<0.00001	<0.0001
Norm(C) later adoptions	1233900	0.0004	0.0001	<0.0001

the total number of words that were adopted by that subreddit. In this analysis, there are once again reliably more first adoption events in the supersubreddits. Further, there are more later adoption events as well. This is possibly because there is simply more of all types of events, since supersubreddits are larger.

Therefore, we wanted to disentangle the effect of different population sizes for subsubreddits and supersubreddits (Norm(U)). In other words, given that more specific communities have fewer users on average, do they still originate more words per user? In short: they do not. Even normalized for the number of users, there are more first adoptions on supersubreddits, though there are significantly fewer later adoption events. This also suggests that it's not simply the result of there being more total events per user.

Nonetheless, we examine the results normalized for the total number of comments, in case there is a non-linear effect of more users on more comments (Norm(C)). Indeed, there is, and in this analysis, the effect reverses. While the numbers are small and still fairly close, there's a robust effect where subsubreddits now have both more first adoptions and later adoptions. This means that every comment is more likely to have a first or later adoption event.

3 Discussion

The story on the dispersion of new words in online communities we present is somewhat complicated by the different direction of the effects based on the type of normalization. We think it can still be disentangled. It relies on two basic social principles. The first is that more people leads to more diverse, varied, and rich conversation. However, specialized subcommunities, such as those in the subsubreddits, have very focused conversation about specific topics.

As we can see from the results, the supersubreddits have more first adoption events in the majority of the analyses we ran. There are likely some social phenomena at play here: with more people, conversation is more varied. In this sense, in Reddit at least, adding people has a greater than linear increase in the amount of conversation. Indeed, when it comes to brainstorming more generally, which word origination could be a subset of, larger groups come up with more and better ideas [5]. Furthermore, electronic brainstorming does not suffer from production blocking with large groups [5]. In Reddit, where communication is possibly asynchronous and does not rely on a shared communication channel, we could expect that effect to be enhanced.

A possible explanation for that is that varied conversation lowers the chance that new content is considered controversial. As the subreddit's topic is broader, the range of allowable discussion topics is also broader. Alternatively, it could be that these communities rely on more centralized sources, perhaps due to the up-vote system. This would be in line with the idea that centralization causes faster dispersion [10]. On the other hand, in a more specific community, reaching outside the norm may cause disagreement in the community.

In fact, this could extend to the point where discussion in these communities, rather than be broadened, is extensively narrowed. This leads to a very small amount of focused discussion. Thus, even though fewer new words are adopted,

there are more new words adopted when normalized for the number of comments. Each comment is more likely to contain a new word, due in part to the small number of new comments.

Indeed, other irregularities have been observed around the effects of large group size before, such as cooperation in social dilemmas [1]. In certain situations, large group size facilitates cooperation, while in others, it seems to hinder it. It is possible that our events are likewise split into two categories: for instance, word originations and first adoptions of external words. Nonetheless, definitively answering that question is beyond the scope of this paper.

4 Conclusion

The adoption and dispersion of new words have been studied from the perspective of social networks, small groups, and cultures, but more rarely at the level of small but overlapping communities. While studies with naming games have suggested smaller groups are more productive, social network analysis has suggested centralization is faster at innovation. We now provide evidence that both of these observations can generalize to the level of small communities, depending by what scale you measure it, corroborating previous research on both word adoption and cooperation.

References

1. Barcelo, H., Capraro, V.: Group size effect on cooperation in one-shot social dilemmas. Sci. Rep. **5**, 7937 (2015)
2. Bergstrom, K.: "Don't feed the troll": shutting down debate about community expectations on Reddit.com. First Monday **16**(8) (2011)
3. Dall'Asta, L., Baronchelli, A., Barrat, A., Loreto, V.: Nonequilibrium dynamics of language games on complex networks. Phys. Rev. E **74**(3), 036105 (2006)
4. Fay, N., Garrod, S., Roberts, L.: The & fitness and function-ality of culturally evolved communication systems. Phil. Trans. Roy. Soc. Lond. B: Biol. Sci. **363**(1509), 3553–3561 (2008)
5. Gallupe, R.B., Dennis, A.R., Cooper, W.H., Valacich, J.S., Bastianutti, L.M., Nunamaker, J.F.: Electronic brainstorming and group size. Acad. Manag. J. **35**(2), 350–369 (1992)
6. Lehrer, A.: Neologisms. In: Brown, K. (ed.) Encyclopedia of Language & Linguistics, 2nd edn, pp. 590–593. Elsevier, Oxford (2006)
7. Reitter, D., Lebiere, C.: How groups develop a specialized domain vocabulary: a cognitive multi-agent model. Cogn. Syst. Res. **12**(2), 175–185 (2011). doi:10.1016/j.cogsys.2010.06.005
8. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 695–704. ACM (2011)
9. Urbatsch, R.: Movers as early adopters of linguistic innovation. J. Sociolinguistics **19**(3), 372–390 (2015)
10. Vilpponen, A., Winter, S., Sundqvist, S.: Electronic word-of-mouth in online environments: exploring referral networks structure and adoption behavior. J. Interact. Advert. **6**(2), 8–77 (2006)

Understanding Discourse Acts: Political Campaign Messages Classification on Facebook and Twitter

Feifei Zhang, Jennifer Stromer-Galley(✉), Sikana Tanupabrungsun, Yatish Hegde, Nancy McCracken, and Jeff Hemsley

School of Information Studies, Syracuse University, Syracuse, NY, USA
{fzhang09,jstromer,stanupab,yhegde,njmccrac,jjhemsle}@syr.edu

Abstract. To understand political campaign messages in depth, we developed automated classification models for classifying categories of political campaign Twitter and Facebook messages, such as calls-to-action and persuasive messages. We used 2014 U.S. governor's campaign social media messages to develop models, then tested these models on a randomly selected 2016 U.S. presidential campaign social media dataset. Our classifiers reach .75 micro-averaged F value on training sets and .76 micro-averaged F value on test sets, suggesting that the models can be applied to classify English-language political campaign social media messages. Our study also suggests that features afforded by social media help improve classification performance in social media documents.

Keywords: Automated classification · Political campaign · Social media · Supervised learning · Text mining

1 Introduction

Since U.S. political campaigns have incorporated social media like Twitter and Facebook into their strategic messaging, it has become more challenging to have a full appreciation for the substance and style of campaigning. To understand campaign messages in depth, we built models to classify each campaign-generated message into a category based on what the message is trying to do: urging people to act, changing their opinions through persuasion, informing them about some activity or event, featuring an endorsement, honoring or mourning people or holidays, or on Twitter having a conversation with members of the public. In this way, we can provide a more expansive and comprehensive lense for understanding political discourse.

To develop classifiers that automatically categorize political campaign message type, we used 2014 governor's campaign data to develop a codebook to categorize campaign message categories, generate training data, and build initial models. Then, we applied these to the 2016 presidential campaign messages to test their generalizability. Finally, we used combined governor's and presidential social media campaign datasets to rebuild more generalizable models, aimed at predicting other political campaign messages. Our classifiers reach .75 micro-averaged F value or better on combined training sets and .76 micro-averaged F value or better on randomly selected presidential test sets. These results suggest that our models can be applied to categorize political campaign social

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 242–247, 2017.
DOI: 10.1007/978-3-319-60240-0_29

media messages written in English. Our study also suggests that considering characteristics of social media messages and adding features afforded by social media help improve classification performance in social media documents.

2 Relevant Literature

Previous studies suggest adding features afforded by social media to the common bag-of-words/ngrams models helps improve classification performance. Conover et al. [1] classified the political alignment of tweets as either politically left, right or ambiguous. They developed two Support Vector Machines (SVM) classifiers with different sets of features. The first classifier focused on representing messages with bag-of-words, but removing common stop words, hashtags, mentions and URLs. The second classifier was trained with features consisting of a bag-of-hashtags. This classifier performed better, with 83.5% accuracy over the first classifier, which was 72.6% accurate. The results suggest that the hashtags of tweets contain more significant semantics indicating the political alignment of the users than the message words. Sriram et al. [8] classified tweets by author's intentions - daily chatter, conversations, sharing information and reporting news. They compared three models: bag-of-words, bag-of-words plus author, and bag-of-words plus author and tweet specific features (e.g. the presence of shortened words and slang, time-event phrases, emphasis on words, and mentions at the beginning and within the tweets). Results showed that the third model performed better than the first two models by 10% to 23%.

3 Data and Content Analysis

We used two open source toolkits [3, 4] to collect social media data covering all 36 states during the 2014 U.S. gubernatorial elections (78 candidates) and all major party candidates (26 campaigns) of the 2016 U.S. presidential election. We collected 34,275 gubernatorial campaign tweets and 9,133 Facebook posts between September 14th and November 11th, 2014. We collected 79,102 presidential campaign tweets and 29,503 Facebook posts from when they declared their presidential bids until election day.

We developed a codebook for categorizing 2014 gubernatorial social media campaign messages through deductive analysis following prior political campaign studies [5]. We developed additional categories through inductive analysis of our corpus. Our final codebook contains 5 message categories for Facebook and 6 categories for Twitter: calls-to-action (CTA), persuasive (PER), informative (INF), endorsements (END), ceremonial (CER), and conversational (CON), in Twitter only.

Four annotators were trained to apply the codebook to sub-datasets of gubernatorial election data. The final inter-coder agreement on a random sample of 648 messages reaches .79 agreement on message categories. When Krippendorff's alpha is .75 or higher, the coding of a variable is considered reliable [6].

We randomly selected 4,147 tweets and 2,494 Facebook posts from gubernatorial data for annotators to code and adjudicate. The adjudicated data is called gold standard data and used for model training. The distribution of the gold standard data is skewed

on both Twitter and Facebook, with more messages of types CTA, PER, INF than of types END, CER and CON.

4 Automated Text Classification

4.1 Model Building

Because Facebook posts and tweets are different in terms of length and other characteristics, we trained multi-class classifier models separately for each application. We built models using the Scikit-learn toolkit [7]. Before training, we pre-processed tweets and Facebook posts by parsing each message to tokens using the ARK Twitter Tokenizer [2], and converted all tokens to lowercase to avoid superfluous features.

We performed many experiments using different multi-class classification algorithms, e.g. SVM with a linear kernel, Naïve Bayes (NB), and Multinomial Logistic Regression (MaxEnt). Among the three, SVM performs best, followed by MaxEnt and NB. As such, we chose SVM for our study.

We tested different combinations of widely used document and feature representation techniques to identify the optimal combination. For document representation, a combination of unigrams and bigrams gives us the highest performance. For four feature representations we compared, results shows that normalized frequency performs worst and the performances of Boolean, term frequency and term frequency–inverse document frequency make little difference. We ended up representing our data with a combination of unigrams and bigrams using Boolean features.

To avoid bloating feature space and over-fitting models, we set a threshold of N-grams representations and filtered out low-frequency features. Our experiments suggest we keep the 3,000 most frequent unigrams when their frequency is higher than 2, and keep the 1,000 most frequent bigrams.

Given the skewed data distribution in our study, we used micro-averaged F value (Micro-F1) to measure overall performance of models. It weights raw scores based on the number of instances in each class [9], which makes it possible to compare averages across result sets. We evaluated classification tasks with 5-fold cross validation.

Using the settings noted above, the F1 scores of our first models are over .70 for all categories except CER and CON (see Model 1 in Table 1). We then developed the second version of models by including the characteristics of social media data based on Model 1. We canonicalized some word tokens by replacing numbers, emoticons, and URLs (e.g. http://abc) with the general tokens. This improves the F1 score of CON category from .48 to .66. We also tried to canonicalize hashtags, but it degrades performance of all categories. We noted that many CON messages started with @mention, and thus added this as a feature. This improves the F1 score of CON from .66 to .76 (see Model 2 in Table 1). Model 2 is our best Twitter model, with .72 Micro-F1, and Model 1 is our best Facebook model, with .74 Micro-F1.

Table 1. Machine learning experiments and performance on governor's data (M: model; P: precision; R: recall; N: Number of training data)

M	Features	Category	Twitter				Facebook			
			P	R	F1	N	P	R	F1	N
1	Lowercase TF = 3 Unigram = 3000 Bigram = 1000	CTA	.78	.78	.78	991	.81	.79	.80	999
		PER	.74	.73	.73	1393	.74	.73	.73	755
		INF	.73	.71	.72	1342	.67	.73	.70	526
		END	.76	.75	.76	166	.79	.73	.76	106
		CER	.41	.40	.40	135	.46	.38	.42	108
		CON	.42	.57	.48	120	n/a	n/a	n/a	n/a
		Micro-F1	.72	.72	.72	4147	.74	.74	.74	2494
2	Lowercase TF = 3 Unigram = 3000 Bigram = 1000 Canonical_form @mention	CTA	.78	.79	.78	991	.81	.80	.80	999
		PER	.74	.74	.74	1393	.74	.74	.74	755
		INF	.74	.73	.73	1342	.67	.70	.68	526
		END	.71	.73	.72	166	.78	.78	.78	106
		CER	.50	.48	.49	135	.52	.45	.49	108
		CON	.76	.77	.76	120	n/a	n/a	n/a	n/a
		Micro-F1	.72	.72	.72	4147	.74	.74	.74	2494

4.2 Codebook and Model Testing

We tested whether the codebook developed and the models trained and validated on governor's data were generalizable to presidential campaign messages. We applied the codebook and models against a randomly selected subset from the early stage of 2016 presidential campaign, 2,989 tweets and 2,638 Facebook posts. We did another round of model testing using a random sample over the campaign period when all the candidates finished their campaigns (see details below).

We used our best classification models to predict message categories of the selected subset. Two annotators who had achieved good intercoder agreement (.79) separately corrected machine predictions. Annotators did not find differences on message categories between governor's and presidential data, suggesting our codebook is applicable for presidential data.

We then compared differences between machine-predicted and human-corrected categories to evaluate the generalizability of models. Our models achieve at .70 Micro-F1 value for both tweets and Facebook posts. This suggests that the gubernatorial and presidential datasets share many features in common. As such, we constructed a new gold standard dataset by combining the gubernatorial gold standard data and the human-corrected presidential data for each social media platform. These new gold datasets were used for re-building the more generalizable models for prediction. The new set comprises of 7,136 tweets and 5,132 Facebook posts.

For the new gold standard data, the optimal sets of features are the same as the gubernatorial models. Specifically, the basic features still give the highest performance for Facebook (Micro-F1 of .76), and adding canonical form and @mention features to the base set is best for Twitter (Micro-F1 of .75), as shown in Table 2.

Table 2. Machine learning experiments and performance on combined governor's and presidential data (M: model; P: precision; R: recall; N: Number of training data)

M	Features	Category	Twitter				Facebook			
			P	R	F1	N	P	R	F1	N
1	Lowercase TF = 3 Unigram = 3000 Bigram = 1000	CTA	0.78	0.81	0.8	1575	0.85	0.82	0.83	2058
		PER	0.76	0.76	0.76	2780	0.75	0.75	0.75	1660
		INF	0.72	0.68	0.7	2187	0.68	0.72	0.7	1065
		END	0.73	0.73	0.73	181	0.73	0.76	0.75	127
		CER	0.41	0.44	0.43	219	0.68	0.72	0.7	222
		CON	0.42	0.51	0.46	194	n/a	n/a	n/a	n/a
		Micro-F1	0.73	0.73	0.73	7136	0.76	0.76	0.76	5132
2	Lowercase TF = 3 Unigram = 3000 Bigram = 1000 Canonical_form @mention	CTA	0.83	0.8	0.82	1575	0.84	0.83	0.84	2058
		PER	0.78	0.76	0.77	2780	0.76	0.75	0.75	1660
		INF	0.72	0.73	0.72	2187	0.66	0.71	0.69	1065
		END	0.7	0.7	0.7	181	0.72	0.7	0.71	127
		CER	0.4	0.48	0.44	219	0.54	0.49	0.51	222
		CON	0.68	0.75	0.71	194	n/a	n/a	n/a	n/a
		Micro-F1	0.75	0.75	0.75	7136	0.76	0.76	0.76	5132

We tested the reliability of these more generalizable models in Dec 2016, when all the candidates finished their campaigns. We randomly selected 1,000 tweets (994 written in English) and 1,000 Facebook posts (987 written in English) over the course of the campaign as test sets. After comparing differences between machine-predicted and human-corrected categories, results suggest that our models work well on the test sets, especially for CTA, PER and INF with Micro-F1 scores of .81, .81, .73 for Twitter and .84, .77, .74 for Facebook, as shown in Table 3.

Table 3. Model testing on presidential data over the campaign period (P: precision; R: recall; N: Number of training data)

Category	Twitter				Facebook			
	P	R	F1	N	P	R	F1	N
CTA	.78	.84	.81	173	.86	.82	.84	331
PER	.85	.77	.81	574	.76	.77	.77	342
INF	.71	.74	.73	189	.67	.82	.74	215
END	.71	.50	.59	10	.67	.38	.48	21
CER	.62	.41	.49	32	.69	.44	.54	78
CON	.85	.69	.76	16	n/a	n/a	n/a	n/a
Micro-F1	.80	.78	.78	994	.77	.77	.76	987

It is not surprising that the CER category is still not predicted accurately, given the limited number of messages in training data and the lack of obvious patterns in text. The END category is also not predicted well. The good performance from cross-validation on training data but poor with the test data indicates that the model might over-fit the training data for this category. We expect that more training data would be helpful to improve the performance of this category.

5 Conclusion and Future Work

To better understand political discourse on social media, this paper built classification models to categorize campaign messages on Twitter and Facebook. The good model performance on training data (Micro-F1 of .75) and randomly selected test data (Micro-F1 of .76) suggests that the models are applicable to categorize other political campaign social media messages. Our study supports prior research [8] that considering social media messages characteristics and including features afforded by social media platforms help improve model performance. In our experimentation, using canonical forms and Twitter's @mention is helpful for classifying conversational tweets. We also found that the classifier trained with a feature space of hashtags performs better, since they include important information [1]. For future research, external aspects of the political campaign might be beneficial, such as including the mentions of the candidate's opponent or self might improve the performance of persuasive category.

Acknowledgements. We thank Dr. Bei Yu's helpful feedback on this paper. The project was supported by the Tow Center for Digital Journalism at Columbia University and the Center for Computational and Data Sciences at the School of Information Studies at Syracuse University.

References

1. Conover, M.D., Gonçalves, B., Ratkiewicz, J., Flammini, A., Menczer, F.: Predicting the political alignment of twitter users. In: Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 192–199. IEEE (2011)
2. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2, pp. 42–47. Association for Computational Linguistics (2011)
3. Hegde, Y.: fb-page-scraper: version 1.33 (2016). https://doi.org/10.5281/zenodo.55940
4. Hemsley, J., Ceskavich, B., Tanupabrungsun, S.: Syracuse Social Media Collection Toolkit (2014). https://github.com/bitslabsyr/stack
5. Jamieson, K.H., Waldman, P., Sherr, S.: Eliminate the negative? Categories of analysis for political advertisements. In: Crowded Airwaves: Campaign Advertising in Elections, pp. 44–64 (2000)
6. Lombard, M., Snyder-Duch, J., Bracken, C.C.: Content analysis in mass communication: assessment and reporting of intercoder reliability. Hum. Commun. Res. **28**(4), 587–604 (2002)
7. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
8. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842. ACM (2010)
9. Van Asch, V.: Macro-and micro-averaged evaluation measures [basic draft] (2013)

Improving the Efficiency of Allocating Crowd Donations with Agent-Based Simulation Model

Chi-Hsien Yen(✉), Yi-Chieh Lee, and Wai-Tat Fu

Department of Computer Science,
University of Illinois Urbana-Champaign, Urbana, USA
{cyen4,wfu}@illinois.edu, be341341@gmail.com

Abstract. Crowdfunding platforms are emerging as an important online social platform to raise capital and awareness for innovative projects. When considered as a general online social system, the goal of a crowdfunding platform is to efficiently allocate a large number of small funds to promising new projects. However, the efficiency of donation allocation and the success rate of projects can be influenced by the behavior of donors, such as how they evaluate each project and choose the projects to donate. To understand how such behavior could impact crowdfunding market, we developed an agent-based model of crowdfunding to investigate three factors, i.e., project visibility, noise of perceived project quality, and donor strategies. These factors may impact the efficiency of a crowdfunding platform.

Keywords: Crowdfunding · Fundraising · Algorithm · Simulation · Social

1 Introduction

Crowdfunding platforms such as DonorsChoose and GiveForward have been receiving growing numbers of projects and donations as they become more popular. Crowdfunding platforms have also been used by philanthropic organizations to raise money from a community of potential donors and promote various campaigns [7]. Similar to other forms of social media platforms, the success of crowdfunding depends critically on how well the platforms support interactions among the online community of donors and project creators. In fact, research has shown that enhancing interactions between creators and donors can increase the success rates of crowdfunding campaigns [5] because donors will more likely engage with the community and donate to the projects.

In addition to fostering an online community, a crowdfunding platform also serves as a marketplace that allows a large number of donors to collectively use their small donations to "vote" for high projects, in ways such that resources from the community can be efficiently allocated. Recent studies on existing crowdfunding platforms, however, showed that the process of matching donations to projects is not always efficient. For example, Solomon et al. [1] argued

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 248–253, 2017.
DOI: 10.1007/978-3-319-60240-0_30

that these superstar projects might have attracted too much attention from the crowd, in the sense that the amount of attention (and donations) is not proportional to their quality relative to other projects. In fact, research has shown that donations dynamics, social media activities, and project updates could influence donors' behavior, which eventually impact the successes of crowdfunding campaigns [1,4]. Consistent with previous research on similar online marketplaces, these results suggest that the collective "voting" by donors may be inefficient in selecting high quality projects [3].

In regard to the efficiency of donation allocation, recently a study [6] proposed a donation method, which allows a donor to put multiple projects in a donation, and the crowdfunding system can reallocate money based on the donor's preference. They used an agent-based model to prove their donation method can efficiently distribute donations and increase overall success rate of crowdfunding. However, it was no clear how different donation dynamics would affect donors' donation behavior in their work. Therefore, we developed an agent-based model based on their model [6] and aim to understand further how different donors with using different strategies to choose projects may play a role impacting outcomes of crowdfunding campaigns.

1.1 Research Question

This paper is motivated by one main research question: *How do donation dynamics interact impact the success rates of low and high perceived quality projects as donors use different strategies to choose projects to donate?*

Specifically, we are studying the effects of three factors: (1) *visibility*, how many projects a donor could review before he or she makes a decision, (2) *noise of perceived quality*, how sensitive a donor is when evaluating the quality of projects, and (3) *donating strategy*, whether a donor is looking for high quality projects to donate or high funded projects.

The three factors are chosen because they reflect the behavior characteristics of the donors in a real crowdfunding platform. First, donors may have limited time and effort to go through all of the projects on a website, which limit their visibility and thus may impact the overall donation efficiency. Second, donors may have difficulties to accurately distinguish high quality and low quality projects, introducing noise into their perceived quality. In the simulation, we defined the quality of a project as an objective measurement of how much a project should be funded based on its presentation quality, motivation, educational benefits, etc. We then controlled the noise of the perceived quality when each donor is evaluating a project to simulate how quality sensitivity could impact market outcome. Third, donors may use different strategies to decide which projects to donate, such as high quality project seekers or high funded project seekers. All of the factors may influence the overall efficiency and impact the success rate of high quality and low quality projects differently.

2 Agent-Based Model

2.1 Model Description

Three main components: agents, crowdfunding projects, and mechanism. In our proposed system, each donor can select multiple projects at the same time with consideration of different factors.

Agents (Donors). Based on the model [6], a donor to put multiple projects in a donation and group them into different preference levels, within which levels all projects are treated as equally preferred. This is a general way to structure their preference toward the selected projects. We investigated two kinds of choosing project strategies as follows,

1. **Quality seeker:** Donors prefer to donate to projects which have high project quality. For example, a project has comprehensive information and convincing contents.
2. **Success seeker:** Donors prefer to choose projects which are close to their project deadline and have already received high proportion of their donation goals.

To test the influence of different strategies to choose projects on the success rate of crowdfunding, we ran multiple simulations with donors using these two different strategies, and controlled the ratio of donors who have these two strategies.

Given that donors cannot review all projects in reality, we also controlled the visibility (V) of each donor. Because our current focus is not on how projects are presented on the platform (e.g., they could be ranked by popularity, dates, etc.), we randomly selected V ongoing projects from the set of all possible projects in each decision cycle, from which the donors would consider and choose to donate using either the quality or success strategy. Given that the perception of project quality is likely noisy, we added a noise error function (E) to each project when it is evaluated by each agent. We assume that the function follows a Normal (Gaussian) distribution $N(0, E_\sigma)$ (i.e., mean of 0 and standard deviation E_σ). We tested the effects of E_σ as it varies from 0 to 5.

In our simulation, 2000 donors per month are randomly generated. Each donor will donate once and the donation amount is drawn from a uniform distribution from \$10 to \$150, where the average (\$80) is close to the average donation amount in real world.

Crowdfunding Projects. In the simulation, we randomly created 1000 new projects per month for 12 months, with a total of 12000 projects. Each generated project was randomly assigned a quality score drawn from a standard normal distribution $N(0, 1)$. Each project had a donation goal of between US\$100 and US\$5000 and a duration of 7 to 30 days (both following uniform distribution).

Donation Distribution. Because this model allows a donor (agent) to select multiple projects in a donation, we designed a donation distribution algorithm to assign donation to crowdfunding projects [6].

3 Simulation Result and Discussion

We ran three simulation experiments to understand how success rate will be impacted by visibility, perceived quality noise, and strategies the donors use. In the first simulation, there were only donors who are quality seekers with $E_\sigma = 1$. We controlled visibility (V) from 10 to 100 in order to see how the number of projects each donors could see would influence the project success rate. In the second simulation, we controlled the perceived quality of the donors ($E_\sigma = 0$ to 4) and set visibility as 10 to investigate how sensitive of quality seeker would have an impact on the crowdfunding success rate. In the final simulation experiment, we set a fixed visibility ($V = 10$) and quality noise ($E_\sigma = 1$). We added success seekers into the simulation and controlled the percentage of quality seekers from 0% to 100% to help us understand the impact of different strategies on the market.

Each data point shown in the figures is an average of 100 repetitive simulation runs, with an error bar of the standard deviation. In each simulation run, 1200 projects and 24000 donors were generated.

1. Visibility: Figure 1 shows the success rate of all projects, high-quality projects (projects that has a quality score higher than average by one standard deviation), and low-quality projects (quality score is lower than average by one standard deviation). The figure also shows the coefficient of quality in a logistic regression model, where the independent variable is quality and the dependent variable is whether a project succeeds or not (with respect to the right y axis). We found that if agents (quality seeker) are able to view more projects and pick out the high-quality ones, the overall success rate will decrease from 64% to 48%. The reason may be that when visibility increases, the donors

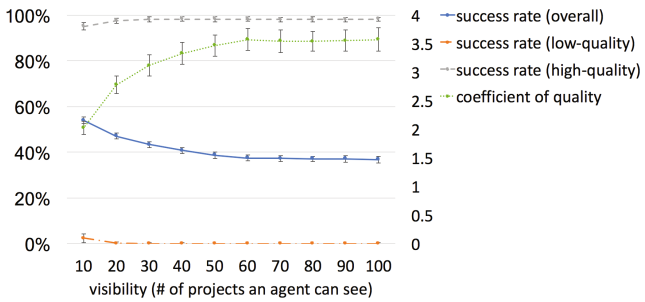


Fig. 1. The trend of success rates and coefficient of quality with respect to visibility (from $10 to $100).

will be more likely to view the same high-quality projects and donate to them. In addition, because our simulation model only allowed an agent to select up to five projects and all agents are using quality strategy, donations may be easily concentrated on a small set of high-quality projects. However, the advantage of increased visibility is that more high-quality projects can be successful.

These findings may be used to improve project recommendation systems in a crowdfunding platform. Because the advantage of increased visibility is that more high-quality projects can be successful, the platform can recommend proper number of projects to the donors and further control frequency of exposure of each project, which could prevent to generate many super star projects, and donations can be allocated more efficiently.

2. Noise of Perceived Quality: Figure 2 presents the trend when E_σ is increased from 0 to 4. When E_σ is low, the coefficient of quality is high and the success rate of high-quality projects is much higher than that of low-quality projects, because the agents can select the good projects accurately. On the other hand, when noise is higher (higher E_σ), more and more low-quality projects are selected, resulting in a decreased coefficient. However, the overall success rate is slightly higher. Once again, the reason is that the projects with highest quality scores are more easily to be overfunded when the noise is low.

3. Ratio of Different Seekers: In this simulation, we set visibility (V) as 10, because it is closer to the real world situation (e.g., most users do not review more than 10 projects before making decision). Also, we set E as 1 because it balances between the success rate of all projects and high-quality projects. In Fig. 3, the overall success rate only slightly decreases when the proportion of quality seekers increases, while the success rates of good and bad projects deviate fast and significantly, which is a desirable outcome for a crowdfunding platform. If half of the donors are quality seekers, more than 85% of high-quality projects will succeed while the overall success rate only decreases 3%, compared to the case that everyone is a success seeker. This result may suggest practitioners of

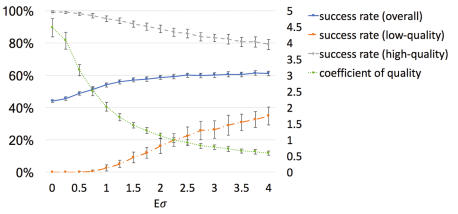


Fig. 2. The trend of success rates and coefficient of quality with respect to perceived quality noise (E_σ from 0 to 4).

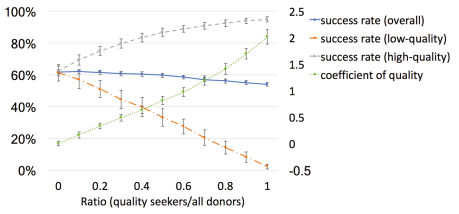


Fig. 3. The trend of success rates and co-efficient of quality with respect to the ratio of quality seekers to success seekers.

crowdfunding websites can encourage donors to focus on the high quality projects or give them some rules to find out high quality projects.

4 Conclusion

In this study, we have presented an agent-based model for crowdfunding platforms, and we investigate the impact of different strategies to choose projects on the efficiency of crowdfunding. The efficiency of a crowdfunding platform is crucial, which depends on how a crowdfunding platform increase overall successful rate and help high quality projects achieve their donation goals in the same time. Based on the findings of the simulations, we found that encouraging donors to choose projects based on the project quality may improve the efficiency of allocating donations. In addition, the visibility and sensitivity of the high quality projects would also impact the successful rate of crowdfunding. The results of this research may provide an indicator to future crowdfunding platforms to redesign their donation methods for improving efficiency of crowdfunding.

References

1. Solomon, J., Ma, W., Wash, R.: Don't wait! how timing affects coordination of crowdfunding donations. In: CSCW 2015, pp. 547–555. ACM Press (2015)
2. Beltran, J.F., Siddique, A., Abouzied, A., Chen, J.: Codo: Fundraising with conditional donations. In: UIST 2015, pp. 213–222. ACM Press (2015)
3. Salganik, M.J., Dodds, P.S., Watts, D.J.: Experimental study of inequality and unpredictability in an artificial cultural market. Science **311**, 854–856 (2006). American Association for the Advancement of Science
4. Xu, A., Yang, X., Rao, H., Fu, W.T., Huang, S.W., Bailey, B.P.: Show me the money!: an analysis of project updates during crowdfunding campaigns. In: ACM SIGCHI (2014)
5. Kim, J.G., Kong, H.K., Karahalios, K., Fu, W.T., Hong, H.: The power of collective endorsements credibility factors in medical crowdfunding campaigns. In: SIGCHI (2016)
6. Lee, Y.C., Yen, C.H., Fu, W.T.: Improving donation distribution for crowdfunding: an agent-based model. In: Xu, K., Reitter, D., Lee, D., Osgood, N. (eds.) SBP-BRiMS 2016. LNCS, vol. 9708, pp. 3–12. Springer, Cham (2016). doi:10.1007/978-3-319-39931-7_1
7. Gerber, E.M., Hui, J.: Crowdfunding: motivations and deterrents for participation. ACM TOCHI (2013)
8. Budak, C., Rao, J.M.: Measuring the efficiency of charitable giving with content analysis and crowdsourcing. In: ICWSM (2016)
9. Kuppuswamy, V., Bayus, B.L.: Crowdfunding creative ideas: the dynamics of project backers in Kickstarter. UNC Kenan-Flagler Research Paper (2015)

Temporal Analysis of Influence to Predict Users' Adoption in Online Social Networks

Ericsson Marin(✉), Ruocheng Guo, and Paulo Shakarian

Arizona State University, Tempe, AZ, USA
Ericsson.Marin@asu.edu

Abstract. Different measures have been proposed to predict whether individuals will adopt a new behavior in online social networks, given the influence produced by their neighbors. In this paper, we show one can achieve significant improvement over these standard measures, extending them to consider a pair of time constraints. These constraints provide a better proxy for social influence, showing a stronger correlation to the probability of influence as well as the ability to predict influence.

1 Introduction

Research has shown that measures which leverage the people's ego network correlate with influence - the confidence at which their neighbors adopt a new behavior [1]. In this paper, we introduce two time constraints to improve these measures: *Susceptible Span* and *Forgettable Span*. *Susceptible Span* (τ_{sus}) refers to the interval when people receive social signals from their neighbors (possible influencing actions), blinding individuals to no more interesting connections. *Forgettable Span* (τ_{fos}) refers to the interval before an influencer's action is forgotten by his neighbors, due to human brain limitation. These constraints define evolving graphs where influence is better measured, as illustrated in Fig. 1.

The contributions of this paper are: we introduce a framework to consider τ_{sus} and τ_{fos} in social influence; we examine the correlation of 10 social network measures to influence under different conditions; we compare the adoption prediction performance of our method with others [1–3], showing relevant improvements. For instance, we obtained up to 92.31% gain in correlation of a simple count of the "active" neighbors with the probability of influence. Considering adoption prediction, F1 score improves from 0.606 (using the state-of-the-art [1]) to 0.689 for active neighbors. Similar results are found for the other measures analysed.

This paper is structured as follows: Sect. 2 presents the related work. Section 3 formalizes a framework to consider the time constraints in our networks. Section 4 presents the experimental setup to produce samples. Section 5 introduces the social influence measures and their corresponding gains in correlation coefficient. Section 6 details the classification experiments and results for the adoption prediction problem. Finally, Sect. 7 concludes the paper.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 254–261, 2017.
DOI: 10.1007/978-3-319-60240-0_31

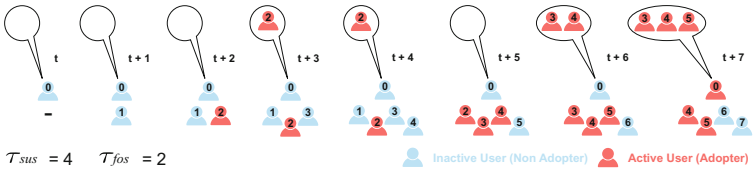


Fig. 1. At time t, the ego node 0 has no neighbors. At $t+1$, he has 1 neighbor (node 1 at the bottom). From this moment, node 0 will be aware about node 1' actions. At $t+2$, node 0 has 2 neighbors, nodes 1, 2. This cumulative process continues until $t+5$ when node 1 is no more a neighbor of node 0, since τ_{sus} was defined as 4. After this time limit, node 0 cannot visualize actions of node 1 anymore. The illustration also shows the node 0's memory inside the balloons. As we made $\tau_{fos} = 2$, node 2 (activated at $t+2$) fades away from node 0's memory after $t+4$, when node 0 is no longer influenced by him. Therefore, at $t+7$, node 0 is activated only by nodes 3, 4 and 5.

2 Related Work

Many works have been proposed to measure social influence and predict users' adoption. For instance, the seminal work of Kempe et al. [4] describes two popular models for diffusion in social networks that were generalized to the General Threshold Model. In this model, the collective influence from a node's infected neighbors will trigger his infection once his threshold is exceeded. Later, Goyal et al. [3] leveraged a variety of models based on pair-wise influence probability, finding the probability of adoption increases with more adopters amongst friends. With an alternative approach, Zhang et al. [1] proposed the *influence locality*, developing two instantiated functions based on pair-wise influence but also on structural diversity to predict adoptions. Comparing two different perspectives, Fink et al. [2] proposed probabilistic contagion models for simple and complex contagion, with the later producing a superior fit for themed hashtags.

In these works, the authors slightly explored the dynamic aspect of social influence. Here, we take the next steps to apply a pair of time constraints to our networks, finding that influence is better measured and predicted dynamically.

3 Framework for Consideration of Time Constraints

In this section, we describe the notations of this work. We denote a set of users V, as the nodes in a directed network $G = (V, E)$, a set of topics (hashtags) Θ, and a set of discrete time points T. We will use the symbols v, θ, t to represent a specific node, topic and time point. With nodes being active or inactive w.r.t θ, an active node (adopter) is a user who retweeted a tweet with θ. We denote an activity log $\mathcal{A}$ (containing all retweets) as a set of tuples of the form $\langle v_1, v_2, \theta, t\rangle$, where $v_1, v_2 \in V$. It describes that "v_1 adopted θ retweeting v_2 at time t", creating a directed edge $(v_1, v_2) \in E$. The intuition behind this edge is that v_1 can be influenced by v_2 with respect to θ', if v_2 eventually adopts θ' after t.

Finally, we integrate into our model the two proposed time constraints τ_{sus} and τ_{fos}. Due to them, the neighborhood of a user can change over time, affecting the social influence measures that result in his decision to adopt a topic. This way, we define the set of neighbors of a node v at time t as:

$$\eta_{v,t} = \{v' | \exists \langle v, v', \theta, t' \rangle \in \mathcal{A},\ s.t.\ t' \leq t\ and\ t - t' \leq \tau_{sus}\}$$

$\eta_{v,t}$ is the set of users whose adoptions since $t - \tau_{sus}$ until t will be presented to v. After $t' + \tau_{sus}$, the adoptions of v' will not influence v. Then, we introduce $\eta^{\theta}_{v,t}$ as the set of users that can influence v to adopt θ at time t as:

$$\eta^{\theta}_{v,t} = \{v' \in \eta_{v,t'} | \exists \langle v', v'', \theta, t'' \rangle \in \mathcal{A},\ s.t.\ t' \leq t'',\ t'' - t' \leq \tau_{sus},\ t'' \leq t\ and\ t - t'' \leq \tau_{fos}\}$$

Consequently, after $t'' + \tau_{fos}$, the fact that v' adopted θ is forgotten by v, with v' no more influencing v in terms of θ. Using these generated dynamic networks, we want to measure the influence produced by the individuals' active neighbors.

4 Experimental Setup

This section details our dataset, how we collect samples using different values for the time constraints, which filters of users' activity are applied, and how we measure correlation of our features with probability of adoption.

Dataset Description. The dataset we use is provided by [5]. It contains 1,687,700 retweets (k), made by 314,756 users (the histogram fits a power-law with $p_k \approx k^{-1.8}$), about 226,488 hashtags on Twitter, from March 24 to April 25, 2012.

Sampling. Following previous works [1], we create balanced sets of samples for our experiments. For a given activity $\langle v, v'', \theta, t \rangle$ which corresponds to a positive sample, we create a negative sample uniformly getting a user v' from the set: $\{v' | v \in \eta^{\theta}_{v',t} \wedge \langle v', u, \theta, t' \rangle \notin \mathcal{A}, \forall u \in \eta^{\theta}_{v',t}\}$. This set includes all users under influence of v w.r.t θ at t, who did not adopted θ in our dataset. Then, we create $\langle v', v, \theta, t \rangle$ as the related negative sample for $\langle v, v'', \theta, t \rangle$, keeping the same timestamp for both users to have similar intervals to accumulate influence.

Filters. In addition, we apply 4 filters to exclude users with less actions than a given threshold, as their behaviors are hardly explainable by influence measurements [2]. We label *R30*, *R60* for users who retweeted at least 30 or 60 times, and *H20*, *H40* for users who retweeted at least 20 or 40 hashtags respectively. This also enables us to test the robustness of τ_{sus}, τ_{fos} under a variety of conditions.

Correlation Between Measures and Adoption Probability. Here, we study how each time constraint correlates with probability of adoption using the Pearson correlation coefficient. The idea is to identify the values for τ_{sus} and $\tau_{fos} \in \{8, 16, 24, 48, 72, 96, 120, 144, 168, 336, 504, 720\}$ $(hours)$, that produce high quality influence measurements (high positive correlation with adoption probability).

5 Influence Measures

Table 1 describes the 10 measures (in 7 categories), which we use to estimate the influence in users' active neighborhood. We define the measures based on the activity $a = \langle v, v', \theta, t \rangle$ by which we create samples. Then, we show the gain (or loss) of correlation coefficient by heat maps, plotting the 144 combinations of τ_{sus} and τ_{fos}. Cells in the right lower corner have values = 0, as $(\tau_{sus}, \tau_{fos}) = (720, 720)$ equals to applying no time constraints (data comprises of 720 h).

Figure 2(a) shows the heat maps for filters $R60$ and $H40$ with the gain (or loss) of correlation coefficients between *Number of Active Neighbors (NAN)* and probability of adoption. Previous work with no time constraints [2,3] argue that a positive correlation is expected here. Even so, many cells present gains for both filters, where combinations of τ_{sus} and τ_{fos} boost NAN's ability to explain users' behaviors under influence. Moreover, hot red cells dominate the left lower region of both heat maps where τ_{sus} and τ_{fos} are relatively high and low respectively, especially when $\tau_{sus} \geq 168$ and $\tau_{fos} \leq 48$, with gains in [9.84%, 92.31%]. Figure 2(b) presents the heat maps for Personal Network Exposure (PNE). Similar to NAN, hot red cells are mainly distributed where $\tau_{sus} >= 168$ and $\tau_{fos} <= 24$. Although the gains are smaller, in [1.18%, 11.76%], we show how PNE obtains high gains in classification performance. From this moment on, we plot the heat maps only for filter *R60*, since we get similar results for $H40$.

Figure 2(c) shows the heat map for Continuous Decay of Influence (CDI). Highest gains in [1.72% 60.34%] are observed when $\tau_{sus} \geq 168$ and $\tau_{fos} \leq 96$.

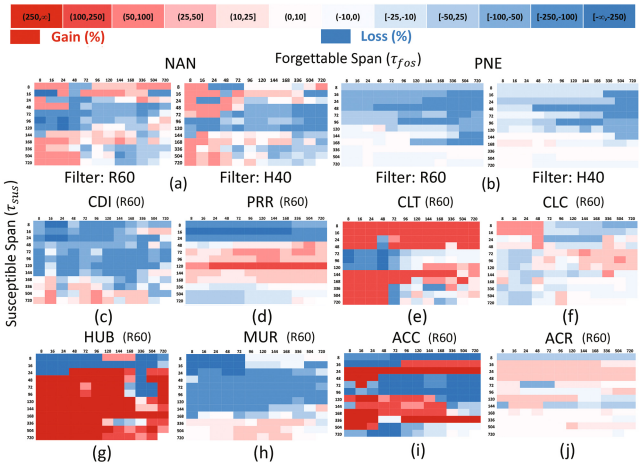


Fig. 2. Gains (red) or losses (blue) of correlation between each social influence measure and adoption probability when the time constraints are applied.

Table 1. Categories of Features.

Category	Feature	Formula
Connectivity	Number of Active Neighbors [3]	$NAN^{\theta}_{v,t} = \lvert\eta^{\theta}_{v,t}\rvert$
	Personal Network Exposure [6]	$PNE^{\theta}_{v,t} = \frac{\lvert\eta^{\theta}_{v,t}\rvert}{\lvert\eta_{v,t}\rvert}$
Temporal	Continuous Decay of Influence [3]	$CDI^{\theta}_{v,t} = \sum_{u \in \eta^{\theta}_{v,t}} \exp(\frac{-(t_l - t_u)}{\sigma})$ where t_l is the time when the latest neighbor in $\eta^{\theta}_{v,t}$ adopted θ, and σ is the globally longest identified time-delay for adoption.
Recorrence	Previous Reposts [7]	$PRR^{\theta}_{v,t} = \sum_{\theta'} \sum_{u \in \eta^{\theta}_{v,t}} \sum_{t' \leq t} \lvert\langle v, u, \theta', t'\rangle\rvert$
Transitivity	Closed Triads [7]	$CLT^{\theta}_{v,t} = \sum_{\{u,z\} \in \eta^{\theta}_{v,t}, u \neq z} f((u,z)^{\theta}_{t})$ and $f((u,z)^{\theta}_{t}) = \begin{cases} 1, \ if \ \langle u, z, \theta, t'\rangle \in \mathcal{A} \wedge t' \leq t \\ 0, \ otherwise \end{cases}$
	Clustering Coefficient [7]	$CLC^{\theta}_{v,t} = \sum_{\{u,z\} \in \eta^{\theta}_{v,t}, u \neq z} \frac{g((u,z)^{\theta}_{t})}{\lvert\eta^{\theta}_{v,t}\rvert^2}$ and $g((u,z)^{\theta}_{t}) = \begin{cases} 1, \ if \ \langle u, z, \theta, t_z\rangle \in \mathcal{A} \wedge t_z \leq t \\ 0, \ otherwise \end{cases}$
Centrality	Hubs [7]	$HUB^{\theta}_{v,t} = \sum_{u \in \eta^{\theta}_{v,t}} h(u,t)$ and $h(u,t) = \begin{cases} 1, \ if \sum_{\theta'} \sum_{x \in V} \sum_{t' \leq t} \lvert\langle x, u, \theta', t'\rangle\rvert >= \gamma \\ 0, \ otherwise \end{cases}$ where γ being the minimal number of messages retweeted. Upon some analysis, we made $\gamma = 104$, corresponding to 0.042% of all retweets. To reach this value, users should be retweeted at least.
Reciprocity	Mutual Reposts [7]	$MUR^{\theta}_{v,t} = \sum_{u \in \eta^{\theta}_{v,t}} i(u,t)$ and $i(u,t) = \begin{cases} 1, \ if \ \langle u, v', \theta, t'\rangle \in \mathcal{A} \wedge t' \leq t \\ 0, \ otherwise \end{cases}$
Structural diversity	Active Strong Connected Components Count [8]	$ACC^{\theta}_{v,t} = \lvert P(\eta^{\theta}_{v,t})\rvert$ where the function $P(V') : V' \rightarrow C$ maps the set of nodes V' to the set of strongly connected components C
	Active Strongly Connected Components Ratio [8]	$ACR^{\theta}_{v,t} = \frac{\lvert P(\eta^{\theta}_{v,t})\rvert}{\lvert P(\eta_{v,t})\rvert}$

Figure 2(d) shows the heat map for Previous Reposts (PRR). There is a slight tendency that higher values to τ_{fos} and lower values τ_{sus} result in higher when compared to the previous features, with gains in [4.76%, 156%].

Figure 2(e) presents the heat map for Closed Triads (CLT). Higher gains in [3.33%, 233.33%] are spread through the majority of cells. However, we can still observe best gains distributed over the area where high values of τ_{sus} and low values of τ_{fos} are found, specially when $\tau_{sus} >= 144$ and $\tau_{fos} <= 48$. Figure 2(f) shows the heat map for Clustering coefficient (CLC). Small values for both time constraints produce higher correlation gains in [4.84%, 66.67%].

Figure 2(g) presents the heat map for Hubs (HUB). We again observe the hot spots where τ_{sus} and τ_{fos} are relatively high and low respectively, with $\tau_{sus} >= 96$ and $\tau_{fos} <= 48$. Gains in [21.05%, 488.24%] are the highest.

Figure 2(h) presents the heat map for Mutual Reposts (MUR), another cumulative measurement whose correlation increases with both time constraints. However, we can observe higher gains in [1.61%, 35.48%] found in the area where the values of τ_{sus} are relatively high and the values of τ_{fos} are intermediate.

Figure 2(i) shows the heat map for Active Strong Connected Components Count (ACC). Hot cells are found where τ_{sus} and τ_{fos} are relatively high and low respectively, mainly when $\tau_{sus} >= 168$ and $\tau_{fos} <= 24$, with gains in [9.09%, 300%]. Figure 2(j) presents the heat map for Active Strong Connected Components Ratio (ACR). The gains in [1.47%, 45.59%] are spread through the cells, mainly where values of τ_{sus} and τ_{fos} are both relatively small.

6 Classification Experiments

This section presents our classification experiments and results for the adoption prediction task, detailing training and testing sets and baselines comparisons.

Training and Testing. Our 10 social influence measures are treated here as features in a machine learning task, such that we can measure their performance for adoption prediction individually and combined. We sort our samples chronologically, using the first 90% for training and the rest for testing (obeying causality which is neglected by some works). We use 2 classifiers, Logistic Regression [9] and Random Forest [9], but only report F1 score for Random Forest under the *R60* filter, since Logistic Regression and other filters produce comparable results.

Baselines. We compare our model with 3 baselines (Influence Locality (LRC-Q) [1], Static Bernoulli (SB) [3], Complex Probability Model (CPM) [2]), to check if our method outperforms them and if the baselines improve with τ_{sus} and τ_{fos}.

Individual Feature Analysis. Table 2 presents the individual classification performance of our 10 features. As done in the previous section, we run an experiment for each combination of τ_{sus} and τ_{fos}, sorting this table by the performance gain. The time constraints boosted the performance in all cases, with gains in F1 score in [7.22% and 23.2%]. In the great majority of cases, τ_{sus} shows values greater than τ_{fos}, repeating the correlation gain pattern.

Combined Feature Analysis. In Table 2, we also present the classification performance results when the 10 features are combined as "All", showing an

Table 2. Baselines, Individual and Combined Feature Performances.

PNE w/ time constraints				PNE w/o time constraints	CLC w/ time constraints				CLC w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Improv.	F1 score	τ_{sus}	τ_{fos}	F1 score	Improv.	F1 score
168	24	0.658	23.2%	0.534	72	16	0.652	22.0%	0.534
ACR w/ time constraints				ACR w/o time constraints	CLT w/ time constraints				CLT w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
336	120	0.632	18.7%	0.532	144	8	0.657	17.3%	0.560
NAN w/ time constraints				NAN w/o time constraints	CDI w/ time constraints				CDI w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
144	72	0.689	15.6%	0.596	144	16	0.677	13.5%	0.596
ACC w/ time constraints				ACC w/o time constraints	HUB w/ time constraints				HUB w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
144	16	0.675	13.2%	0.596	120	16	0.630	9.99%	0.573
PRR w/ time constraints				PRR w/o time constraints	MUR w/ time constraints				MUR w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
72	8	0.672	9.82%	0.612	336	72	0.712	7.23%	0.664
All w/ time constraints				All w/o time constraints	LRC-Q w/ time constraints				LRC-Q w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
336	48	0.755	10.54%	0.683	72	16	0.657	8.41%	0.606
SB w/ time constraints				SB w/o time constraints	CPM w/ time constraints				CPM w/o time constraints
τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score	τ_{sus}	τ_{fos}	F1 score	Impro.	F1 score
72	8	0.675	8.69%	0.621	72	8	0.689	12.58%	0.612

improvement of 10.54% when applying time constraints. The observed pattern for the individual features (social influence is better measured by the measures when $\tau_{sus} > \tau_{fos}$) is found again for the features combined, with performance achieving the best improvements when $\tau_{sus} = 336$, while $\tau_{fos} = 48$.

We interpret these results as: (1). users will start losing attention of their neighbors after 2 weeks, if they do not retweet them anymore; (2). users will no more remember the activations of their neighbors after approximately 2 days.

Performance of Baseline Methods. Finally, Table 2 includes the results of baselines. The time constraints boost all performances, with gains of 8.41% for LRC-Q, 8.69% for SB and 12.58% for CPM. These results highlight the effectiveness of τ_{sus} and τ_{fos}, also consolidating the pattern detected before: $\tau_{sus} > \tau_{fos}$. In addition, our model outperforms the baselines in both situations: when we use only an individual feature such as MUR, and when we use all features combined, with improvements in [3.33%, to 9.6%] (compared with CPM).

7 Conclusion

In this paper, we introduce a pair of time constraints to show how the dynamic graphs produced by them better capture the influence between users over time (specially when τ_{sus} and τ_{fos} are relatively high and low respectively). We validate our model under diverse conditions, detailing how it outperforms the state-of-the-art methods that aim to predict users' adoption. We also demonstrate how these constraints can be used to improve the performance of other approaches, enabling practical usage of the concepts for social influence prediction.

Acknowledgments. Some of the authors of this paper are supported by CNPq-Brazil, AFOSR Young Investigator Program (YIP) grant FA9550-15-1-0159, ARO grant W911NF-15-1-0282, and the DoD Minerva program.

References

1. Zhang, J., Liu, B., Tang, J., Chen, T., Li, J.: Social influence locality for modeling retweeting behaviors. In: Proceedings of 13th IJCAI 2013, pp. 2761–2767. AAAI Press (2013)
2. Fink, C., Schmidt, A., Barash, V., Kelly, J., Cameron, C., Macy, M.: Investigating the observability of complex contagion in empirical social networks. In: Proceedings of 10th ICWSM (2016)
3. Goyal, A., Bonchi, F., Lakshmanan, L.: Learning influence probabilities in social networks. In: 3rd ACM International Conference on Web Search And Data Mining (WSDM 2010), pp. 241–250 (2010)
4. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of 9th ACM SIGKDD, pp. 137–146. ACM, New York (2003)
5. Weng, L., Menczer, F., Ahn, Y.: Virality prediction and community structure in social networks. Sci. Rep. **3** (2013). Article no. 2522
6. Valente, T.W.: Network Models of the Diffusion of Innovations. Quantitative Methods in Communication. Hampton Press, Cresskill (1995). pp. 153–163
7. Zafarani, R., Abbasi, M., Liu, H.: Social Media Mining. Cambridge University Press, Cambridge (2014)
8. Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J.: Structural diversity in social contagion. Proc. Nat. Acad. Sci. **109**(16), 5962–5966 (2012)
9. Attewell, P., David, M., Darren, K.: Data Mining for the Social Sciences. UC Press, Berkeley (2015)

Ideology Detection for Twitter Users via Link Analysis

Yupeng Gu[1(✉)], Ting Chen[1], Yizhou Sun[1], and Bingyu Wang[2]

[1] University of California Los Angeles, Los Angeles, USA
{ypgu,tingchen,yzsun}@cs.ucla.edu
[2] Northeastern University, Boston, USA
rainicy@ccs.neu.edu

Abstract. The problem of ideology detection is to study the latent (political) placement for people, which is traditionally studied on politicians according to their voting behaviors. Recently, more and more studies begin to address the ideology detection problem for ordinary users based on their online behaviors that can be captured by social media, e.g., Twitter. As far as we are concerned, the vast majority of the existing methods on ideology detection on social media have oversimplified the problem as a binary classification problem (i.e., liberal vs. conservative). Moreover, though social links can play a critical role in deciding one's ideology, most of the existing work ignores the heterogeneous types of links in social media. In this paper we propose to detect *numerical* ideology positions for Twitter users, according to their *follow*, *mention*, and *retweet* links to a selected set of politicians. A unified probabilistic model is proposed that can (1) integrate heterogeneous types of links together in determining people's ideology, and (2) automatically learn the quality of each type of links in deciding one's ideology. Experiments have demonstrated the advantages of our model in terms of both ranking and political leaning classification accuracy.

1 Introduction

Ideology detection, i.e., ideal point estimation, dates back to early 1980s, where political scientists first studied politicians' political affiliation using their roll call voting data [12]. Recently, more and more studies pay attention to ideology detection for users on social media, which captures rich information for ordinary citizens in addition to political figures. However, there are two major limitations of the existing literature. First, most of these approaches oversimplify the ideology detection problem as a binary classification problem (liberal/conservative), while ignoring the fact that people's ideology lies in a very broad spectrum. Second, despite the successful utilization of link information in determining one's ideology, most of the works ignore the heterogeneous link types in social media, which leads to significant information loss.

In this paper, we propose a unified probabilistic model to detect *numerical* ideology positions for Twitter users, according to their *follow*, *mention*, and

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 262–268, 2017.
DOI: 10.1007/978-3-319-60240-0_32

retweet links. Although defined on Twitter network, our approach is very general to other social networks. Our approach is able to combine multiple types of links in determining people's ideology with different weights for each link type. In addition, the strength of each link type can be automatically learned according to the network. Experiments shows that (1) using multiple types of links is better than using any single type of links alone to determine one's ideology, and (2) the detected ideology for Twitter users aligns with our intuition quite well.

2 Approach

In this section, we introduce our solution to the proposed problem. We start from ideology model under a single link type, then introduce how to extend the model when multiple types of links exist, and finally introduce the learning algorithm.

2.1 Ideology Estimation Model via Single Link Type

As in traditional ideal point models, each user has an intrinsic position in a K-dimensional space $\boldsymbol{p}_i \in \mathbb{R}^K$, which represents his/her ideology. For a politics-related network, ideology can help explain the reason for link generation, which is a reflection of people's online behaviors. Take *follow* link as an example: the proximity of two users' positions in the latent ideology space indicates a high probability that they have many politician friends (followees) in common, and vice versa. Inspired by [2], we analogize the action of following others to one's voting behavior, and define the probability that user u_i follows user v_j as $p(i \rightarrow j) = \sigma(\boldsymbol{p}_i \cdot \boldsymbol{q}_j + b_j) := \sigma_{ij}$, where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. $\boldsymbol{q}_j$ can be interpreted as the image or impression vector of user v_j when viewed by others, and b_j can be regarded as a bias term for v_j which denotes her popularity. The dot product between two feature vectors can be regarded as a similarity measure in the vector space. Treating each link as generated by a Bernoulli distribution with parameter σ_{ij}, we are able to write down the log-likelihood of observing a network G as

$$l(G) = \log\Big(\prod_{(i,j)} \sigma_{ij}^{I_{[i\rightarrow j]}}(1-\sigma_{ij})^{1-I_{[i\rightarrow j]}}\Big) = \sum_{(i,j):i\rightarrow j} \log\sigma_{ij} + \sum_{(i,j):i\nrightarrow j} \log(1-\sigma_{ij}) \quad (1)$$

where $I_{[\cdot]}$ is the indicator function. We denote the set of existing links as S_+ and sampled set of non-existing links as S_- in the remaining of this paper.

2.2 Ideology Estimation Model via Multiple Link Types

We now address the challenge of utilizing multiple types of links for ideology detection. In a heterogeneous network, nodes can be connected via different types of relations. On Twitter, people can *follow*, *mention* or *retweet* others. It naturally forms three different types of links, and different link types certainly

have different interpretations. According to our previous assumption, the intrinsic ideology $\boldsymbol{p}_i$ will be consistent across all link types. However we posit that the images of users will change when observed by different types of behaviors. For example, u_i can easily decide to *follow* v_j but hesitates to *retweet* from v_j. Therefore, $\boldsymbol{q}_j$ and b_j will be changed to relation-specific parameters $\boldsymbol{q}_j^{(r)}$ and $b_j^{(r)}$. In consideration of the heterogeneity in different types of links, we also add a relation weight w_r which represents the relative importance of the links in the corresponding relation type r. Besides, we will use the average log likelihood of each link in each type of relation in order to balance the scale of different link types. An l_2 regularization term is added on parameters to avoid overfitting.

Denoting $\boldsymbol{P} = \{\boldsymbol{p}_i\}, \boldsymbol{Q} = \{\boldsymbol{q}_j^{(r)}\}$ and $\boldsymbol{B} = \{b_j^{(r)}\}$, we define our objective as

$$\begin{aligned} l(\boldsymbol{G}|\boldsymbol{P},\boldsymbol{Q},\boldsymbol{B}) = \sum_{r=1}^{R} w_r \cdot \frac{1}{N_r}\Big(\sum_{(i,j)\in S_+^{(r)}} e_{r,ij}^{+} \log \sigma_{r,ij} + \sum_{(i,j)\in S_-^{(r)}} e_{r,ij}^{-} \log\left(1-\sigma_{r,ij}\right)\Big) \\ -\frac{\mu}{2}\big(||\boldsymbol{P}||_F^2 + \sum_{r=1}^{R} ||\boldsymbol{Q}^{(r)}||_F^2 + \sum_{r=1}^{R} ||\boldsymbol{b}^{(r)}||_2^2\big) \end{aligned} \tag{2}$$

where $\sigma_{r,ij} = \sigma(\boldsymbol{p}_i \cdot \boldsymbol{q}_j^{(r)} + b_j^{(r)})$ for short, N_r is the total number of links in relation r, and $\mu > 0$ is a parameter that controls the effect of regularization terms. The constraint we put on w_r is $w_r > 0,\ r = 1, \cdots, R$ and $\prod_{r=1}^{R} w_r = 1$.

3 Experiments

3.1 Data Preparation

We first collect the list of all the members of the 113^{th} U.S. congress (2013–2015). Then we use Twitter's API to collect their followees and followers. We collect at most 5,000 followers and followees for every congressman. On one hand, in order to select politics-related users, we set a threshold where we keep users who follow at least t congressmen or are followed by at least t congressmen. We choose $t = 20$ in consideration of efficiency. On the other hand, we also include around 10,000 random users who follow 3–5 politicians as more peripheral (less politics-related) Twitter users. Our approach will be evaluated on this Twitter subnetwork with these users as vertices. Finally we collect their most recent tweets[1] up to Jan. 2016. Social networks for different relations (*follow*, *mention* and *retweet*) are built from the friend lists and extracted from one's tweets. In total, 46,477 users are involved in the dataset and the number of edges for *follow*, *mention* and *retweet* networks is 1.8 M, 2.4 M and 718 K, respectively.

3.2 Performance Evaluation

Baseline Methods. We compare our Multiple Link Types Ideal Point Estimation Model (*ML-IPM*) with the following baseline methods:

[1] Due to API limits, only the most recent 3,200 tweets for each user are available.

- *AVER*: the simplest baseline where the ideology of a user is the average score of her outgoing neighbors. Each Republican is assigned an ideology score of 1, and each Democrat is assigned a score of −1.
- *B-IPM* (Bayesian Ideal Point Estimation Model) [1]. Although the author does not mention their generalization to relations other than *follow*, we adopt the model for other types of links for comparison.
- *SL-IPM*: our Single Link Type Ideal Point Estimation Model where only one type of link is present in the social network, as introduced previously.
- *ML-IPM-fixed*: a special case of our model ML-IPM where the weights for different types of links are fixed. In this case the weights for different link types are uniformly distributed, namely $w_1 = w_2 = \cdots = w_R = 1$.

Evaluation Measures. In our experiments, we will evaluate the ranking and classification accuracy to demonstrate the effectiveness of our model.

Ranking. In order to evaluate the effect of continuous ideology, we design the ranking evaluation based on 100 manually labeled users, with integer labels from 1 (most liberal/left) to 5 (most conservative/right). The manual labels are obtained by reading their profile information and tweet content, which is never used in the training stage. Here we evaluate the pairwise accuracy between Twitter users, where a pair of users is considered correct if the order of their 1-dimensional ideologies aligns with the order of manual labels. The accuracy is defined as the fraction of correct pairwise arrangements between these users. We use five different sets of random initialization for the model parameters, and report the mean and standard deviation on a total of 3,857 pairs of users in Table 1, where the relation in the bracket represents the type of link used in the corresponding method.

Classification. In the classification task, we classify users as liberal or conservative based on the ideology we have inferred from the dataset. To obtain the ground truth of some users in our dataset, we collect congress people's party affiliation as well as the political leaning for 100 popular newspaper accounts[2], and we also take advantage of the labeled users in our previous task. These multi-dimensional ideal points are used to train a logistic regression classifier. The classification performance is measured by the Area Under ROC Curve (AUC), and is averaged over 10 different runs by different samples of training data. The mean AUC and standard deviation are reported in Table 1. We select the ideology dimension to be $K = 5$ in our method.

3.3 Case Studies

We visualize the latent ideology position of Twitter users. We collect all users in our dataset who claim themselves to live in one of the 50 states in the U.S. (or Washington, D.C.), and calculate the average ideology for each area. Then we are able to map the average score to a color between red and blue. As a

[2] Source: http://www.mondotimes.com/newspapers/usa/usatop100.html.

Table 1. Experimental results (test data)

Method	Ranking accuracy (%)	Classification AUC (%)
AVER (*follow*)	42.7	52.3
AVER (*mention*)	44.6	55.8
AVER (*retweet*)	47.4	58.7
B-IPM (*follow*)	44.3 ± 10.2	86.8 ± 2.1
B-IPM (*mention*)	43.3 ± 18.3	55.8 ± 6.4
B-IPM (*retweet*)	50.1 ± 12.7	56.1 ± 6.6
SL-IPM (*follow*)	62.6 ± 1.1	95.3 ± 1.5
SL-IPM (*mention*)	62.3 ± 2.7	95.1 ± 1.8
SL-IPM (*retweet*)	63.7 ± 0.5	95.8 ± 0.5
ML-IPM-fixed	65.5 ± 0.8	93.0 ± 3.5
ML-IPM	$\mathbf{66.3 \pm 0.7}$	$\mathbf{98.6 \pm 1.3}$

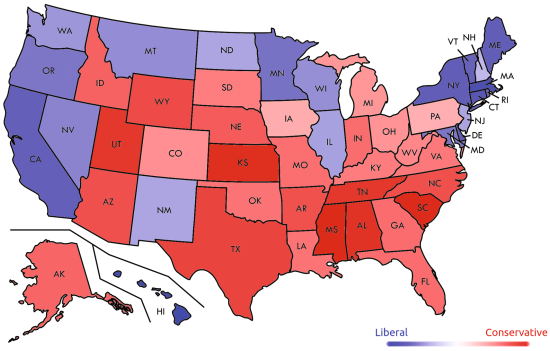


Fig. 1. Average ideology for Twitter users in each state. Darker red means more conservative, while darker blue means more liberal. (Color figure online)

result, 9,362 users are identified and 29 states are labeled as red (conservative), as shown in Fig. 1. We can see that the colors of most areas agree with recent election results: states along the west coast and new England area are mostly liberal; while most conservative states lie in the midwest and south region.

4 Related Work

4.1 Ideology Detection in Roll Call Voting Data

Ideal point models attempt to estimate the position of each lawmaker in the latent political space. Legislative voting is one of the sources for quantitative estimation of lawmakers' ideal points. Poole and Rosenthal [12] were among the first few researchers in political science domain to provide a thorough and rigorous approach for ideology estimation, which has been generalized by numerous

other political science scholars [2,7,9,10,13,14]. Researchers study the public voting record of lawmakers and model the probability of each vote, which is usually described as the interaction of the lawmaker's ideal point and the position of the bill. Along this line of research, computer science researchers extend the ideal point model to a variety of aspects, including applying natural language processing and topic modeling techniques on bills [3,4,6,8,11].

4.2 Ideology Detection in Social Networks

Apart from voting records, recently many approaches have been using information from social networks to analyze user's political leaning. Typically, inference of a user's ideal point is made by exploring her neighbors and her relationship with labeled users (e.g. politicians). Therefore, a simple yet intuitive approach would be calculating the ratio of Democrats and Republicans that a user befriends with [5]. Wong et al. [15,16] assume liberal people tend to tweet more about liberal events and the same for conservative users. Barberá [1] proposes a probabilistic model to describe the likelihood of the social network, where the probability of a link is defined as a function of ideal points of both users.

5 Conclusion

In this paper we present a novel approach for ideology detection on Twitter using heterogeneous types of links. Instead of predicting binary party affiliations of users, we focus on a more comprehensive task of detecting continuous ideal points for Twitter users. In addition, we improve over traditional ideology estimation models by integrating information from heterogeneous link types in social networks. Specifically, our model is able to automatically update the importance scores of various relations on Twitter. The experimental results on a subnetwork of Twitter show our advantage over the baseline methods.

References

1. Barberá, P.: Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit. Anal. **23**(1), 76–91 (2015)
2. Clinton, J., Jackman, S., Rivers, D.: The statistical analysis of roll call data. Am. Polit. Sci. Rev. **98**(02), 355–370 (2004)
3. Gerrish, S., Blei, D.M.: Predicting legislative roll calls from text. In: Proceedings of the 28th International Conference on Machine Learning (ICML 2011), pp. 489–496 (2011)
4. Gerrish, S., Blei, D.M.: How they vote: issue-adjusted models of legislative behavior. In: Advances in Neural Information Processing Systems (NIPS 2012), pp. 2762–2770 (2012)
5. Golbeck, J., Hansen, D.: Computing political preference among Twitter followers. In: Proceedings of the SIGCHI Conference on Human Factors, Computing Systems, pp. 1105–1108 (2011)

6. Gu, Y., Sun, Y., Jiang, N., Wang, B., Chen, T.: Topic-factorized ideal point estimation model for legislative voting network. In: Proceedings of the 20th International Conference on Knowledge Discovery and Data Mining (KDD 2014), pp. 183–192 (2014)
7. Heckman, J.J., Snyder Jr., J.M.: Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators. RAND J. Econ. **28**, S142–S189 (1997)
8. Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political ideology detection using recursive neural networks. In: Association for Computational Linguistics (2014)
9. Jackman, S.: Multidimensional analysis of roll call data via bayesian simulation: identification, estimation, inference, and model checking. Polit. Anal. **9**(3), 227–241 (2001)
10. Londregan, J.: Estimating legislators' preferred points. Polit. Anal. **8**(1), 35–56 (1999)
11. Nguyen, V.-A., Boyd-Graber, J., Resnik, P., Miler, K.: Tea party in the house: a hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In: Proceedings of ACL (2015)
12. Poole, K.T., Rosenthal, H.: A spatial model for legislative roll call analysis. Am. J. Polit. Sci. **29**, 357–384 (1985)
13. Poole, K.T., Rosenthal, H.: Patterns of congressional voting. Am. J. Polit. Sci. **35**(1), 228–278 (1991)
14. Poole, K.T., Rosenthal, H.: Congress: A Political-Economic History of Roll Call Voting. Oxford University Press, Oxford (1997)
15. Wong, F., Tan, C., Sen, S., Chiang, M.: Quantifying political leaning from tweets and retweets. In: ICWSM (2013)
16. Wong, F., Tan, C., Sen, S., Chiang, M.: Media, pundits and the us presidential election: quantifying political leanings from tweets. In: Proceedings of the International Conference on Weblogs and Social Media (2013)

Methodology

Spread of Pathogens in the Patient Transfer Network of US Hospitals

Juan Fernández-Gracia[1,2(✉)], Jukka-Pekka Onnela[3], Michael L. Barnett[3], Víctor M. Eguíluz[2], and Nicholas A. Christakis[4]

[1] Department of Epidemiology,
Harvard T.H. Chan School of Public Health, Boston, USA
[2] Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB),
Palma, Spain
juanf@ifisc.uib-csic.es
[3] Department of Biostatistics,
Harvard T.H. Chan School of Public Health, Boston, USA
[4] Yale Institute for Network Science, Yale University, New Haven, USA

Abstract. Antibiotic-resistant organisms, an increasing source of morbidity and mortality, have a natural reservoir in hospitals, and recent estimates suggest that almost 2 million people develop hospital-acquired infections each year in the US alone. We investigate the temporal network of transfers of Medicare patients across US hospitals over a 2-year period to learn about the possible role of hospital-to-hospital transfers of patients in the spread of infections. We analyze temporal, geographical, and topological properties of the transfer network and show that this network may serve as a substrate for the spread of infections. Finally, we study different strategies for the early detection of incipient epidemics on the temporal transfer network as a function of activation time of a subset of sensor hospitals. We find that using approximately 2% of hospitals as sensors, chosen based on their network in-degree, with an activation time of 7 days results in optimal performance for this early warning system, enabling the early detection of 80% of the *C. difficile*. cases with the hospitals in the sensor set activated for only a fraction of 40% of the time.

Every year in the US alone, there are 1.7 million nosocomial infections and 99,000 associated deaths, imposing substantial clinical and financial costs to the US health care system [1–3]. The vast majority of these are due to antibiotic-resistant bacteria [4], which have a natural reservoir in hospitals, presenting a potentially lethal threat to already-sick patients. The annual cost of antibiotic-resistant infections in the US has been estimated to range from $21 billion to $34 billion [5–7]. A 2013 CDC (Centers for Disease Control and Prevention) report on antibiotic-resistant bacteria identified the lack of infrastructure to detect and respond to emerging resistant infections as a pressing gap.

Antibiotic-resistant organisms have a natural reservoir in hospitals. In our study, over a two-year period, there were nearly one million transfer events across US hospitals of Medicare patients alone. Given this large number of transfers,

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 271–280, 2017.
DOI: 10.1007/978-3-319-60240-0_33

the network of patient transfers could plausibly act as a conduit for antibiotic-resistant bacteria from hospital to hospital. There are, however, only a few existing studies that have investigated the possible role of hospital-to-hospital transfers of patients for the spread of infections. Some studies have focused on the structure of the nationwide transfer network associated with critical care [8–11], while others have had a more restricted scope, limited to smaller geographical units, such as counties [12,14].

Local containment of antibiotic-resistant bacteria at the level of individual hospitals is a difficult but manageable task given that interactions between hospital wards are relatively structured and confined spatially [15,16]. But controlling a larger epidemic of antibiotic-resistant bacteria or responding to new mass outbreaks is much more challenging. This is in part related to the complex pattern of patient movements between hospitals, which gives rise to a broad, distributed network. To better understand the role of patient transfers for the spread of infections, we pursue three interconnected aims. First, we investigate the structure of the hospital-to-hospital patient transfer network in the US; second, we correlate the incidence of nosocomial infections on a national scale with properties of this network; and third, we develop a scalable method for the efficient early detection of the spread of nosocomial infections.

1 Materials and Methods

1.1 Study Data

We study hospital-to-hospital transfers of the entire population of US Medicare patients over a two-year period. Medicare provides almost universal coverage to all Americans aged 65 and older, about 15% of the US population [17]; and about 37% of all hospital admissions in 2003 were for Medicare patients [18]. We used a 100% sample of the Medicare Provider Analysis and Review (MedPAR) files for calendar years 2006 and 2007. The MedPAR files contain diagnosis, procedure, and billing information on all inpatient and skilled nursing facility (SNF) stays. Our study cohort consisted of Medicare patients aged 65 or older with a hospital stay at an acute medical or surgical hospital with an active record in the American Hospital Association (AHA) 2005 database [19]. Before applying these exclusion criteria, we identified 26.4 million stays of 12.5 million patients in 6,278 different hospitals. After the exclusions, our final cohort consisted of 21.0 million inpatient stays of 10.4 million patients in 5,667 different hospitals.

1.2 Hospital-to-hospital Transfers

According to our definition, a hospital-to-hospital transfer occurs whenever a patient is discharged from one hospital and admitted to another hospital on the same calendar day. Note that a minority of transfers as defined here may not correspond to actual formal transfers of patients. For example, a patient could be discharged from hospital A and then be re-admitted to hospital B on

the same day for a reason that is unrelated to her stay at hospital A. From an epidemiological point of view, however, these are essentially equivalent to formal patient transfers. Using this definition of transfer, we identified 936,101 transfer events taking place between 76,003 pairs of hospitals.

1.3 Constructing the Transfer Network

We consider a network representation of the patient transfers across hospitals. Hospitals are represented as nodes and a transfer of a total of x patients on day d from hospital i to hospital j is represented as a directed edge from node i to node j with weight x on day d. The longitudinal sequence of patient transfers forms a directed, weighted, temporal network. We consider a static representation of the network that retains no temporal information of patient transfers by aggregating the data for the two-year period, where the weight of the edge from node i to node j is the mean daily number of patient transfers through that edge, *i.e.*, the total number of transfers from hospital i to hospital j during the study period divided by the number of days in the period (730 days).

1.4 C. difficile Incidence on the Transfer Network

The MedPAR files contain diagnosis codes for each patient. We investigated the incidence of *Clostridium difficile* (*C.difficile*) infections and its correlation with properties of the transfer network. *C. difficile* is an anaerobic, gram-positive, spore-forming bacteria that occurs frequently in health care settings, found in over 20% of patients hospitalized for more than one week. The disease is spread by ingestion of *C. difficile* spores, which are very hardy and can persist on environmental surfaces for months without proper hygiene [20]. *C. difficile* associated infections kill an estimated 14,000 people a year in the US as a result of institutional infections [21]. We ascertained incident cases of *C. difficile* infection by identifying any hospital admissions with ICD-9 diagnostic code 008.45. The sensitivity and specificity of using ICD-9 codes to identify *C. difficile* infections have been reported by multiple groups to be adequate for identifying overall *C. difficile* burden for epidemiological purposes [22–24].

1.5 Sensor Placement on the Hospital Network

To set up a real-time surveillance system for infections, such as a new strain of antibiotic-resistant *C. difficile*. It is unlikely that exhaustive data would be available for all hospitals all the time, and this limitation calls for a parsimonious approach where only a subset of hospitals needs to be monitored at any given time. We call these monitored hospitals "network sensors" in the sense that they could be used to sense incipient epidemics. We consider three different prescriptions for sensor placement: (1) choose sensor hospitals in proportion to their in-degree rank in the static network; (2) choose sensor hospitals in proportion to their out-degree rank in the static network; and (3) choose sensor

hospitals uniformly at random from the set of all hospitals. In our simulations, we assume that a monitored hospital is able to detect every infected patient who is present either in the hospital itself or in any of its network neighbors to which it is connected via patient transfers. To learn about the potential of the hospital sensor framework to detect epidemics, we investigate its best-case performance by determining the optimal sensor set for the observed data.

1.6 Determining the Optimal Sensor Set

We define the relative efficacy of the sensor E_N set as $E_N = D_N/ND_1 - (M - D_N)/M$ where N is the number of sensors in the sensor set, D_N the number of infected patients detected by a sensor set of N sensors, and M is the total number of *C. difficile* cases in the network. While adding sensors to the system always improves its overall performance, any sensor set exhibits diminishing marginal returns in the sense that the per-sensor increment in performance declines with each added sensor. The first term in the definition corresponds to the number of detected cases normalized by the number of cases that would be detected if all sensors were as efficacious as the first sensor in the sensor set. The second term is a penalty term that corresponds to the fraction of undetected cases. High relative efficacy is therefore a combination of selecting a set of sensors that are as close as possible to the efficaciousness of the first sensor in the set and having these sensors miss as small a proportion of cases as possible. Note that the two terms in the definition of the relative efficacy could be assigned different weights; however, here, we opted for the simplest approach and only ensured that the two contributions are measured on the same scale.

1.7 Implementation of Network Sensors

The sensor hospitals can be either passive or active. When a sensor is passive, it can only detect infections in the hospital itself. Whenever an infection is detected, the sensor either transitions from the passive state to the active state for a period of T days or, if already in the active state, remains in that state for another T days. In addition to the efficacy of the sensor sets, for both implementations, we keep track of the fraction of *C. difficile* cases that are detected in order to assess the performance of the sensor system.

We monitor the admission times of *C. difficile* patients at each hospital, and whenever such a patient is admitted, we incorporate the hospital in the sensor set for T days, the activation time, following the admission. Once added to the sensor set, the hospital can detect the *C. difficile* cases present in the hospital itself and its network neighbors for a total of T days. The efficacy of the sensor system therefore depends on the value of T, and we compute the efficacy of the sensors for T from 0 to 100 days (shown from 0 to 30 days in Fig. 2 left). For each combination of parameter values, the number of sensors and the activation time, and for each strategy of prescribing sensors, we perform 1,000 independent realizations of the sensor selection process. We also track the average time each sensor stays in the active state. An optimal sensor set is one that has maximal

efficacy for activation time T, minimizes the average time the sensors stay active, and maximizes the fraction of detected cases.

2 Results

2.1 Properties of the Transfer Network

The topology of the network and the geography of patient transfers are closely related, with 90% of transfers between hospitals less than 200 Km apart (Fig. 1 left). On average, over the 2-year period, a hospital sent patients to 13.55 ± 0.15 (SE) hospitals and received patients from 13.55 ± 0.25 hospitals (note that the two means necessarily coincide in a directed network). The average number of patients transferred per edge in the 2-year period was 12.3 ± 0.63 (SE). Although the degree distributions (in-degree and out-degree) have fat tails (more so the in-degree), comparisons of the average clustering coefficient and the average shortest path length to randomized versions of the network show that the network closely resembles a spatial network. In particular, it is much more clustered than a random network and has a high average shortest path length. Finally, the network shows no significant assortativity by degree.

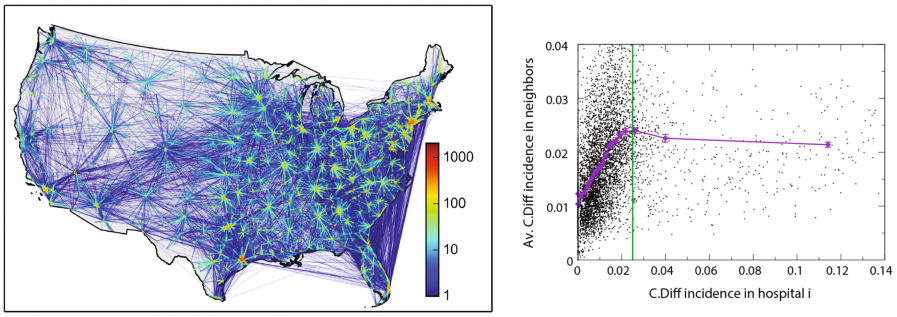


Fig. 1. Left: hospital transfer network of US Medicare patients. The network consists of hospitals connected by daily transfers of patients, aggregated over the two-year period. Edge color encodes the number of patients transferred through each connection. **Right: correlation between *C. difficile* incidence and transfer network structure.** The x-axis represents the temporal *C. difficile* incidence at the focal hospital over time and the y-axis is the mean *C. difficile* incidence in its network neighborhood (the mean taken first over time and then over all network neighbors). We exclude hospitals with fewer than 100 patients from subsequent correlation analyses, leading to exclusion of 7.5% (428) of all hospitals. The Pearson correlation coefficients are 0.47 and −0.01 for the low and high incidence regimes, respectively, which are separated by the vertical line. (Color figure online)

2.2 Spread of C. Diff. Infections

Over the two-year period, there were a total of 313,214 *C. difficile* infections in the 5,677 hospitals included in the study. The mean *C. difficile* incidence for each hospital and the mean *C. difficile* incidence for its network neighbors show two distinct regimes, one for low *C. difficile* incidence and another for high incidence (Fig. 1 right). The incidence of the pathogen in a given hospital is correlated to the incidence of the pathogen in its network neighborhood if the incidence at the focal hospital is relatively low; this correlation appears to vanish for hospitals displaying higher *C. difficile* incidence. One explanation for this phenomenon is that, if there were only very few cases of *C. difficile* in the low incidence regime, the transfers of infected patients might go undetected, therefore inducing correlations among pathogen incidences across the network. Conversely, if pathogen incidence were high and local, such that hospital outbreaks are detected, patient transfers might be restructured to curb the further spread of the infection. We determine the boundary between the two regimes based on the strength of correlation in pathogen incidence and assign the value for the crossover between the two regimes (shown as the vertical line in Fig. 1 right). For *C. difficile* incidence below this threshold, the Pearson correlation coefficient $R \approx 0.47$ (95% CI: 0.44, 0.49) whereas above the threshold $R \approx -0.01$ (95% CI: -0.08, 0.07), where the confidence intervals for the correlation coefficients where estimated using the Fisher z-transformation [25]. This finding on the correlation of *C. difficile* incidence across hospitals that are neighbors in the transfer network supports the use of the transfer network as a substrate for the spread of nosocomial infections.

2.3 Monitoring the System for Hypothetical Outbreaks

We used three different strategies for selecting the sensor nodes based on their properties in the static network, choosing them based on their in-degree rank, out-degree rank, or choosing them at random. Nodes with a high in-degree are expected to be efficient at funneling in pathogens from their network environment, whereas nodes with a high out-degree are expected to rapidly funnel out their pathogens.

Except for very low activation times of the order of a few days, the efficacy and the fraction of detected cases are almost unaffected by this parameter (Fig. 2 left). The optimal sensor set of a strategy stabilizes after $T = 5$ days (Fig. 2 right). These results corroborate that choosing sensors based on in-degree is the best overall strategy, followed by out-degree, and then the random strategy. All of the strategies result in similar sizes for the most efficient sensor sets as in the static case. In terms of the fraction of detected cases, all three strategies perform similarly, each covering about 80% of the cases. We find that the average time a sensor spends in the active state increases as a function of the activation time T. Therefore, an optimal approach is to choose the smallest activation time T that does not deteriorate performance of the sensor system in terms of the fraction of detected cases. For an activation time $T = 5$, the average fraction of time sensors spend in the active state is 0.51 for in-degree based selection, 0.47 for out-degree based selection, and 0.46 for the random strategy.

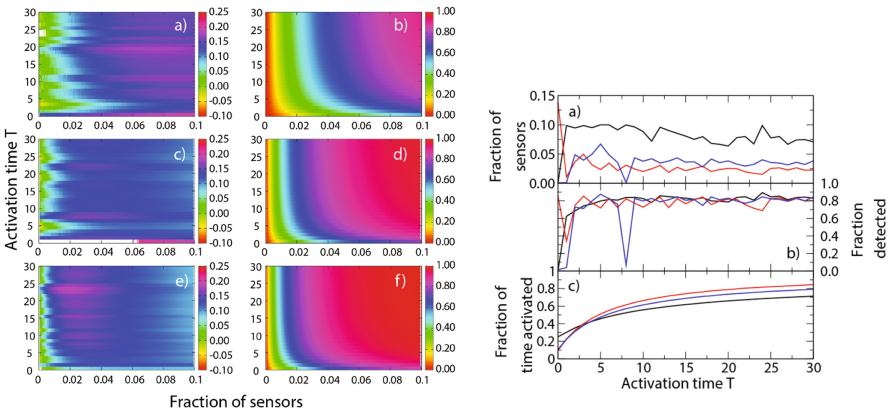


Fig. 2. Left: the optimal sensor set. Heatmaps showing the efficacy (left column) and fraction of detected cases (right column) on the temporal transfer network, as a function of the fraction of hospitals acting as sensors (horizontal axes) and the activity time that they implement (vertical axes). The rows of panels correspond to choosing the sensors randomly (top row), proportional to out-degree (middle row) and proportional to in-degree (bottom row). **Right: efficacy of temporal sensor sets. a)** Fraction of sensors for the most efficient sensor set from the temporal network for sensors chosen at random (black), proportional to in-degree (red), and proportional to out-degree (blue). We have smoothened the efficacy curves by averaging the results using a window of 5 sensors. **b)** Fraction of detected cases for the most efficient sensor set. **c)** Average fraction of time that a sensor stays in the active state (same color code as on the left) (Color figure online)

3 Conclusions

We studied a network defined by the transfer of 12.5M Medicare patients across 5,667 US hospitals over a 2-year period. The network is strongly geographically embedded, with 90% of all transfers spanning a distance less than 200 km. The transfer network could plausibly be used as a substrate for the spread of pathogens: we observed a positive correlation for *C. difficile* incidence between hospitals and their network neighbors, identifying two qualitatively distinct regimes corresponding to low and high *C. difficile* incidence. Finally, selecting hospitals as sensors based on their in-degree in the static network was able to detect a large fraction of infections. Furthermore, an activation time of just 5 to 7 days using the dynamic sensor implementation is sufficient to achieve this surveillance with just 2% of the hospitals acting as sensors. These results support our conceptual model that the structure of the nationwide hospital patient transfer network is important for the spread of health-care associated infections, likely well beyond the illustrative case of *C. difficile* considered here. In particular, our work highlights the need to monitor the network of transfers not just individual hospitals in order to track infectious outbreaks.

Other pathogens might need a different number of sensor hospitals, a different set of sensor hospitals, or different surveillance windows. Nevertheless, the health of the entire hospital system, from the perspective of nosocomial infections or other outbreaks, could be monitored by leveraging the network structure of patient transfers.

Our study has several limitations. First, the data we used to map the hospital networks are from 2006 and 2007. However, given that hospital transfer patterns are strongly embedded in the geography of the country, as we also demonstrated here, we do not expect the age of the data to affect our results substantially. Second, we cannot assess the extent to which unobserved policies or commercial constraints might have affected the flow of patients from one hospital to another; however, these policies merely affected patient transfers, which are, in any case, observable in the current and similar future data. Third, our analyses and models assume that patient transfers are the only mechanism responsible for the spread of infections. There are, of course, other vectors or means that might result in hospitals being infected, such as the movement of physicians, nurses, and other health care staff between hospitals. Finally, in this analysis, we did not make use of the fine-scale temporal information available in transfer data; future work could evaluate how bursts of infected patients, perhaps on particular days of the week, might contribute to an epidemic.

Understanding the structure and dynamics of the hospital transfer network for the spread of real infections has a number of important implications. Empirical data could be used, either periodically or perhaps even in real time to map networks of patient movement in the US health care system, and this network could then be used monitor the spread of nosocomial and other infections in the network. In our estimation, such a system could detect 80% of *C. difficile* cases using just 2% of hospitals as network sensors. Our methods suggest practicable strategies for identifying which hospitals should serve a surveillance function for the whole system and, in the dynamic implementation, how long the sensors should retain a higher level of alertness after each index case. These tools would be useful not only for public health interventions in the case of natural epidemics, but also in the case of deliberate ones, such as those due to a possible bioterror attack. In conclusion, the actual structure and flow pattern of patients across US hospitals confers certain specific vulnerabilities and defenses, regardless of the biology of the pathogen per se, placing theoretical bounds on any effective containment strategy directed at a contagious pathogen.

Acknowledgments. We thank Laurie Meneades for the expert assistance required to build the dataset. JFG and JPO are joint first authors of this article.

References

1. Zimlichman, E., Henderson, D., Tamir, O., et al.: Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. JAMA Intern. Med. **173**(22), 2039–2046 (2013)

2. Threat Report 2013-Antimicrobial Resistance - CDC. http://www.cdc.gov/drugresistance/threat-report-2013/
3. Klevens, R.M., Edwards, J.R., Richards, C.L., et al.: Estimating health care-associated infections and deaths in U.S. hospitals, 2002. Publ. Health Rep. **122**(2), 160–166 (2007)
4. Infectious Diseases Society of America (IDSA): Combating antimicrobial resistance: policy recommendations to save lives. Clin. Infect. Dis. **52**(S5) (2011)
5. Roberts, R.R., Hota, B., Ahmad, I., et al.: Hospital and societal costs of antimicrobial-resistant infections in a Chicago teaching hospital: implications for antibiotic stewardship. Clin. Infect. Dis. **49**(8), 1175–1184 (2009)
6. Mauldin, P.D., Salgado, C.D., Hansen, I.S., Durup, D.T., Bosso, J.A.: Attributable hospital cost and length of stay associated with health care-associated infections caused by antibiotic-resistant gram-negative bacteria. Antimicrob. Agents Chemother. **54**(1), 109–115 (2010)
7. Filice, G.A., Nyman, J.A., Lexau, C., et al.: Excess costs and utilization associated with methicillin resistance for patients with Staphylococcus aureus infection. Infect. Control Hosp. Epidemiol. **31**(4), 365–373 (2010)
8. Karkada, U.H., Adamic, L.A., Kahn, J.M., Iwashyna, T.J.: Limiting the spread of highly resistant hospital-acquired microorganisms via critical care transfers: a simulation study. Intensive Care Med. **37**(10), 1633–1640 (2011)
9. Iwashyna, T.J., Christie, J.D., Kahn, J.M., Asch, D.A.: Uncharted paths: hospital networks in critical care. Chest **135**(3), 827–833 (2009)
10. Iwashyna, T.J., Christie, J.D., Moody, J., Kahn, J.M., Asch, D.A.: The structure of critical care transfer networks. Med. Care **47**(7), 787–793 (2009)
11. Unnikrishnan, K.P., Patnaik, D., Iwashyna, T.J.: Spatio-temporal structure of US critical care transfer network. AMIA Summits Transl. Sci. Proc. **2011**, 74–78 (2011)
12. Lee, B.Y., McGlone, S.M., Song, Y., et al.: Social network analysis of patient sharing among hospitals in Orange County. California. Am. J. Public Health **101**(4), 707–713 (2011)
13. Lee, B.Y., McGlone, S.M., Wong, K.F., et al.: Modeling the spread of Methicillin-Resistant Staphylococcus Aureus (MRSA) outbreaks throughout the hospitals in Orange County. California. Infect. Control Hosp. Epidemiol. **32**(6), 562–572 (2011)
14. Huang, S.S., Avery, T.R., Song, Y., et al.: Quantifying interhospital patient sharing as a mechanism for infectious disease spread. Infect. Control Hosp. Epidemiol. **31**(11), 1160–1169 (2010)
15. Obadia, T., Silhol, R., Opatowski, L., Temime, L., Legrand, J., et al.: Detailed contact data and the dissemination of Staphylococcus aureus in hospitals. PLoS Comput. Biol. **11**(3), e1004170 (2015)
16. Isella, L., Romano, M., Barrat, A., Cattuto, C., Colizza, V., Van den Broeck, W., Gesualdo, F., Pandolfi, E., Rav, L., Rizzo, C., Tozzi, A.E.: Close encounters in a pediatric ward: measuring face-to-face proximity and mixing patterns with wearable sensors. PLoS ONE **6**(2), e17144 (2011)
17. Medicare beneficiaries as a percent of total population. http://kff.org/medicare/state-indicator/medicare-beneficiaries-as-of-total-pop/
18. Overview of hospital stays in the United States (2010). http://www.hcup-us.ahrq.gov/reports/statbriefs/sb144.jsp
19. American Hospital Association. http://www.aha.org/

20. Gerding, D.N., Johnson, S.: Harrisons principles of internal medicine. In: Fauci, A.S., Braunwald, E., Kasper, D.L., et al. (eds.) 17th Editi. McGraw-Hill, New York (2008)
21. Bajardi, P., Barrat, A., Savini, L., Colizza, V.: Optimizing surveillance for livestock disease spreading through animal movements. J. R. Soc. Interface **9**(76), 2814–2825 (2012)
22. Schmiedeskamp, M., Harpe, S., Polk, R., Oinonen, M., Pakyz, A.: Use of international classification of diseases, ninth revision, clinical modification codes and medication use data to identify nosocomial clostridium difficile infection. Infect. Control Hosp. Epidemiol. **30**(11), 1070–1076 (2009)
23. Scheurer, D.B., Hicks, L.S., Cook, E.F., Schnipper, J.L.: Accuracy of ICD-9 coding for Clostridium difficile infections: a retrospective cohort. Epidemiol. Infect. **135**(6), 1010–1013 (2007)
24. Dubberke, E.R., Butler, A.M., Yokoe, D.S., et al.: Multicenter study of surveillance for hospital-onset Clostridium difficile infection by the use of ICD-9-CM diagnosis codes. Infect. Control Hosp. Epidemiol. **31**(3), 262–268 (2010)
25. Fisher, R.A.: Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika **10**(4), 507–521 (1915)
26. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E **64**(2), 026118 (2001)
27. Newman, M.: Assortative mixing in networks. Phys. Rev. Lett. **89**(20), 208701 (2002)
28. Newman, M.: Mixing patterns in networks. Phys. Rev. E **67**(2), 026126 (2003)
29. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms, chap. 16. MIT Press, Cambridge (2001)
30. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Science **220**, 671–680 (1983)

On Predicting Geolocation of Tweets Using Convolutional Neural Networks

Binxuan Huang and Kathleen M. Carley(✉)

School of Computer Science, Carnegie Mellon University,
5000 Forbe Ave., Pittsburgh, USA
{binxuanh,kathleen.carley}@cs.cmu.edu

Abstract. In many Twitter studies, it is important to know where a tweet came from in order to use the tweet content to study regional user behavior. However, researchers using Twitter to understand user behavior often lack sufficient geo-tagged data. Given the huge volume of Twitter data there is a need for accurate automated geolocating solutions. Herein, we present a new method to predict a Twitter user's location based on the information in a single tweet. We integrate text and user profile meta-data into a single model using a convolutional neural network. Our experiments demonstrate that our neural model substantially outperforms baseline methods, achieving 52.8% accuracy and 92.1% accuracy on city-level and country-level prediction respectively.

1 Introduction

Recently, there is growing interest in using social media to understand social phenomena. For example, researchers have shown that analyzing social media reveals important geospatial patterns for keywords related to presidential elections [23]. People can use Twitter as a sensor to detect earthquakes in real-time [22]. Recent research also has demonstrated that Twitter data provides real-time assessments of flu activity [1].

Using Twitter's API[1], a keyword search can be done and we can easily get tweet streams from across the world containing keywords of interest. However, we cannot conduct a fine-grained analysis in a specific region using such a keyword-based search method. Alternatively, using the same API tweets with geo-information can be collected via a bounding box. Since less than 1% of tweets are tagged with geo-coordinates [7], using this location-based search means we will lose the majority of the data. If we can correctly locate those ungeotagged tweets returned from a keyword search stream, that would enable us to study users in a specific region with far more information.

With this motivation, we are aiming to study the problem of inferring a tweet's location. Specifically, we are trying to predict on a tweet by tweet basis, which country and which city it comes from. Most of the previous studies rely

[1] https://dev.twitter.com/docs.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 281–291, 2017.
DOI: 10.1007/978-3-319-60240-0_34

on rich user information(tweeting history and/or social ties), which is time-consuming to collect because of the Twitter API's speed limit. Thus those methods could not be directly applied to Twitter streams. In this paper, we study a global location prediction system working on each single tweet. One data sample is one tweet JSON object returned by Twitter's streaming API. Our system utilizes location-related features in a tweet, such as text and user profile metadata. We summarize useful features that can provide information for location prediction in Table 1.

Table 1. Feature table

Feature	Type
Tweet content	Free text
User personal description	Free text
User name	Free text
User profile location	Free text
Tweet Language (TL)	Categorical
User Language (UL)	Categorical
Timezone (TZ)	Categorical
Posting Time (PT)	UTC timestamp

Recent research has shown that using bag-of-words and classical machine learning algorithms such as Naive Bayes can provide us a text-based location classifier with good accuracy [9]. Different from previous research, we intend to use the convolutional neural network(CNN) to boost prediction power. Inspired by the success of convolutional neural network in text classification [12], we are going to use CNN to extract location related features from texts and train a classifier that combines high-level text feature representations with these categorical features. To benchmark our method, we compared our approach with a stacking-based method. Experimental results demonstrate that our approach achieves 92.1% accuracy on country-level prediction and 52.8% accuracy on city-level prediction, which greatly outperforms our baseline methods on both tasks.

2 Related Work

Identifying demographic details of Twitter users [16] has been widely studied in previous literature including inferring users' attributes like age, gender [5], and personalities like openness and conscientiousness [18]. Among these research, there is increasing interest in inferring Twitter user's location, which is largely driven by the lack of sufficient geo-tagged data [7]. In many situations, it is important to know where a tweet came from in order to use the information in the tweet to effect a good social outcome. Key examples include: disaster relief [14], earthquake detection [6], and predicting flu trends [1].

A majority of previous works either focus on a local region e.g. United States [4], Sweden [2], or using rich user information like a certain number of tweets for each user [4], user's social relationship [11,17]. Different from these works, this paper works on worldwide tweet location prediction. We only utilized features in one single tweet without any external information. Thus this method could be easily applied to real-time Twitter stream.

For fine-grained location prediction, there are several types of location representation methods existing in literature. One typical method is to divide earth into small grids and try to predict which cell one tweet comes from. Wing and Baldridge introduced a grid-based representation with fixed latitude and longitude [24]. Based on the similarity measured by Kullback-Leibler divergence, they assign each ungeotagged tweet to the cell with most similar tweets. Because cells in urban area tend to contain far more tweets than the ones in rural areas, the target classes are rather imbalanced [8]. To overcome this, Roller et al. further proposed an adaptive grid representation using K-D tree partition [20]. Another type of representation is topic region. Hong et al. proposed a topic model to discover the latent topic words for different regions [10]. Such parametric generative model requires a fixed number of regions. However, the granularity of topic regions is hard to control and will potentially vary over time [9].

The representation we choose is city-based representation considering most tweets come from urban area. One early work proposed by Cheng et al. using a probabilistic framework to estimate Twitter user's city-level location based on the content of tweets [4]. Their framework tries to identify local words with probability distribution smoothing. However, such method needs a certain number of tweets(100) for each user to get a good estimation. Han et al. proposed a stacking-based approach to predict user's city [8]. They combine tweet text and meta-data in user profile with stacking [25]. Specifically, they train a multinomial naive Bayes base classifier on tweet text, profile location, timezone. Then they train a meta-classifier over the base classifiers. More recently, Han et al. further did extensive experiments to show that using feature selection method, such as information gain ratio [19] could greatly improve the classification performance.

3 Location Prediction

In this section, we will introduce our location prediction approach. We first briefly describe the useful features in a tweet JSON object. After that, we will further explain how we utilize these features in our prediction model.

3.1 Feature Set

We have listed all useful information we want to utilize in Table 1. Tweet content, user personal description, user name and profile location are four text fields that we will use. Twitter users often reveal their home location in their profile location and personal description. However, location indicating words are often mixed with informal tweet text (e.g. chitown for Chicago). It is unrealistic to

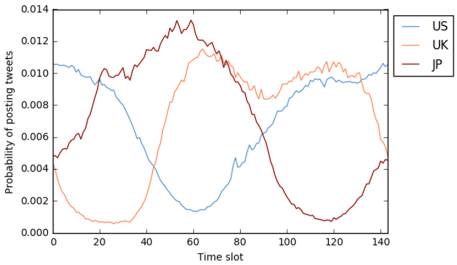


Fig. 1. The probability of an user posting a tweet in different time slot in three different countries: United States, United Kingdom, Japan.

use a gazetteer to find these words. In this work, we choose to apply CNN on these four text fields to extract high-level representations.

In addition to these four text fields. There are another three categorical features: tweet language, user language, and timezone. Tweet language is automatically determined by Twitter's language detection tool. User language and timezone are selected by the user in his/her profile. These three categorical features are particularly useful for distinguishing users at the country-level.

The last feature is UTC posting time. Using posting timestamp as a discriminative feature is motivated by the fact that people in a region are more active on Twitter at certain times during the day. For example, while people in United Kingdom start to be active at 9:00 am in UTC time, most of the people in United States are still asleep. We transform the posting time in UTC timestamp into discrete time slots. Specifically, we divide 24 h into 144 time slots each with a length of 10 min. Thus each tweet will have a discrete time slot number in the range of 144, which can be viewed as a categorical feature. In Fig. 1, we plotted the probability distribution of an user posting tweets in each time slot in three different countries. As expected, there is a big variance between these three countries.

3.2 Our Approach

Our approach is based on the convolutional neural network for sentence classification proposed by Kim [12]. Different from traditional bag-of-words method, such convolutional neural networks take the word order into consideration. Our model architecture is shown in Fig. 2. We use this CNN architecture to extract high-level features from four text fields in a tweet. Let $x_i^t \in R^k$ be the k-dimensional word vector corresponding to the i-th word in the text t, where $t \in \{$tweet content, user description, profile location, user name$\}$. As a result, one text of length n can be represented as a matrix

$$X_{1:n}^t = x_1^t \oplus x_2^t \oplus .... \oplus x_n^t \tag{1}$$

where $\oplus$ is concatenation operator. In the convolutional layer, we apply each filter $w \in R^{hk}$ to all the word vector matrices, where h is the window size and

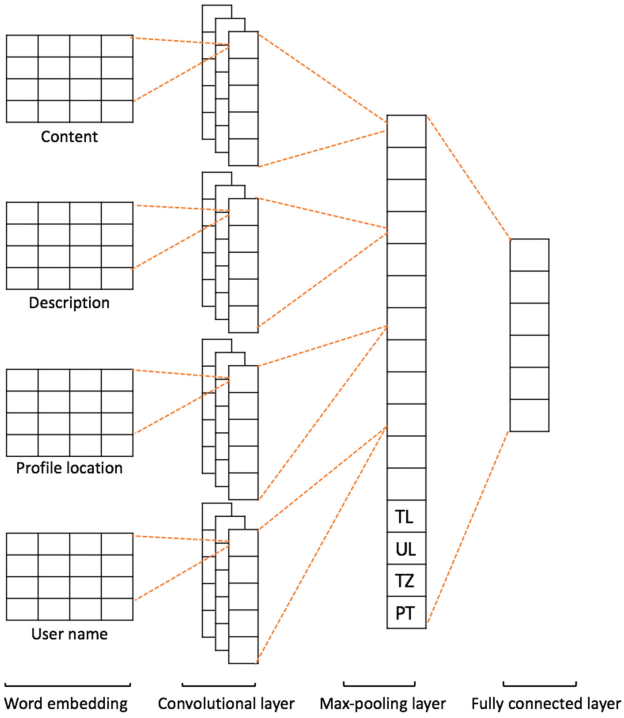


Fig. 2. A diagram of the architecture of our neural model.

k is the length of a word vector. For example, applying filter w to a window of word vectors $x^t_{i:i+h-1}$, we generated $c^t_i = f(w \cdot x^t_{i:i+h-1} + b)$. Here $b \in R$ is a bias term and we choose $f(x)$ as a non-linear ReLU function $max(x, 0)$. Sliding the filter window from the beginning of a word matrix till the end, we generated a feature vector $c^t = [c^t_1, c^t_2, ..., c^t_{n-h+1}]$ for each text t. If we have m filters in the convolutional layer, then we can produce m feature vectors for each text field and $4m$ vectors in total.

In the max-pooling layer, we apply a pooling operation over each feature vector generated in the convolutional layer. Each pooling operation takes a feature vector as input and outputs the maximum value $\hat{c}^t = max(c^t)$. $\hat{c}^t$ can be viewed as the most representative feature generated by a filter on text t. Hence we finally got a long vector $\theta \in R^{4m}$ after the max-pooling layer. To avoid the co-adaptation of hidden units, we apply dropout on the max-pooling layer that randomly set elements in θ to zero in the training phase. After that, we append four categorical features tweet language (TL), user language (UL), timezone (TZ) and posting time (PT) with one-hot encoding at the end of θ and get $\hat{\theta}$. In the last fully connected layer, we use a softmax function over this long vector $\hat{\theta}$ to generate the probability distribution over locations. Specifically, the probability of one tweet coming from location l_i is

$$P(l_i|\hat{\theta}) = \frac{exp(\beta_i^T \hat{\theta})}{\sum_{j=1}^{L} exp(\beta_j^T \hat{\theta})} \tag{2}$$

where L is the number of locations and β_i are parameters in softmax layer. The output predicted location is just the location with highest probability.

The minimization objective in the training phase is the categorical cross-entropy loss. The parameters to be estimated include word vectors, weight vectors w for each filters, the weight vectors β in softmax layer, and all the bias terms. The optimization is performed using mini-batch stochastic gradient descent and back-propagation [21].

4 Data

We used geo-tagged tweets collected from Twitter streaming API[2] for training and evaluation. In this study, we set the geographic bounding box as [-180, -90, 180, 90] so that we could get these geo-tagged tweets from the whole world. Our collection started from January 7, 2017 to February 1, 2017. Because it is very common for one user to post tweets from the same city, we randomly chose one tweet for each city that one user has visited. This could ensure that there is no strong overlap among our data samples. We only used tweets either with specific geo-coordinates or a geo-bounding box smaller than [0.1, 0.1]. For the latter case, we used the center of one tweet's bounding box as its coordinates. No other filtering was done. There are 3,321,194 users and 4,645,692 tweets in total. For test data we used all tweets from 10% of the users who were randomly selected. For the remaining 90% users, we picked tweets from 50,000 of them as a development set and used the remaining tweets as training data.

There are two location prediction tasks we consider in this paper. The first task is country-level location prediction. We adopted the country code in the geo-tagged tweet as the label we want to predict. In our dataset, there are 243 countries and regions in total. The second task is city-level location classification. We adopt the same city-based representation as Han et al. [3]. The city-based representation consists of 3,709 cities throughout the world and was obtained by aggregating smaller cities with the largest nearby city. We assigned the closest city for each tweet based on orthodromic distance. Table 2 contains basic statistics about our dataset. It is worth mentioning that this dataset is rather imbalanced, where a majority of tweets are sent from a few countries/cities.

5 Experiments

5.1 Evaluation Measures

Following previous work of tweet geolocation prediction [8], we used four evaluation measures listed below. One thing to note is that when we calculated the error distance we used distance between predicted city and the true coordinates in the tweet rather than the center of assigned closest city.

[2] https://dev.twitter.com/streaming/reference/post/statuses/filter.

Table 2. Summaries about the dataset. Numbers in brackets are standard deviation.

# of tweets	# of users	# of time-zones	# of lang	# of coun-tries (or regions)	Tweets per country	# of cities	Tweets per city
4645692	3321194	417	103	243	19118.0 (99697.1)	3709	1252.5 (4184.5)

- Acc: The percentage of correct location predictions.
- Acc@Top5: The percentage of true location in our top 5 predictions.
- Acc@161: The percentage of predicted city which are within a 161 km (100 mile) radius of the true coordinates in the original tweet to capture near-misses. This measure is only tested on city-level prediction.
- Median: The median distance from the predicted city to the true coordinates in the original tweet. This measure is only tested on city-level prediction.

5.2 Baseline Method

We compared our approach with one commonly used ensemble method in previous research works [8,9]. We implemented an ensemble classifier based on stacking [25] with 5-fold cross validation. The training of stacking consists of two steps. First, five multinomial naive Bayes base classifiers are trained on different types of data(tweet content, user description, profile location, user name and the remaining categorical features). The outputs from the base classifiers are used to train a multinomial naive Bayes classifier in the second layer. We call such method STACKING in this paper. Same as [9], we also use information gain ratio to do feature selection on text tokens. We call STACKING with feature selection STACKING+.

5.3 Hyperparameters and Training

We used a tweet-specific tokenizer provided by NLTK[3] to tokenize text fields. We built our dictionary based on the words that appeared in text, user description, and profile location. To reduce low-utility words and noise, we removed all words that had a word frequency less than 10. For our proposed approach, we used filter windows(h) of 3,4,5 with 128 feature vectors each, a dropout rate of 0.5 and batch size of 1024. We initialize word vectors using word2vec[4] vectors trained on 100 billion words from Google News. The vectors have dimensionality of 300 and were trained using the continuous bag-of-words architecture [15]. For those words that are not included in word2vec, we initialized them randomly. We also performed early stopping based on the accuracy over the development set. Training was

[3] http://www.nltk.org/api/nltk.tokenize.html.
[4] https://code.google.com/archive/p/word2vec/.

Table 3. Country prediction results.

	Acc	Acc@Top5
STACKING	0.868	0.947
STACKING+	0.871	0.950
Our approach	**0.921**	**0.972**

Table 4. City prediction results.

	Acc	Acc@161	Acc@Top5	Median
STACKING	0.389	0.573	0.595	77.5 km
STACKING+	0.439	0.616	0.629	47.2 km
Our approach	**0.528**	**0.692**	**0.711**	**28.0** km

done through stochastic gradient descent using Adam update rule with learning rate 10^{-3}[13]. For our baseline models, we applied additive smoothing with $\alpha = 10^{-2}$, which is selected on the development set. For STACKING+ method, we first ranked these words by their information gain ratio value, then selected the top n% words as our vocabulary. The tuning of n is based on accuracy over the development set. We selected n as 40%, 55% for city-level prediction and country-level prediction respectively.

5.4 Results

The comparison results between our approach and the baseline methods are listed in Tables 3 and 4. Our approach achieves 92.1% accuracy and 52.8% accuracy on country-level and city-level location prediction respectively. Our approach is consistently better than the previous model on the country-level location prediction task as shown in Table 3. It greatly outperforms our baseline methods over all the measures, especially on the city-level prediction task. It could assign more than half of the test tweets to the correct city and gain more than 20% relative improvement over the accuracy of the STACKING+ method.

Our approach performs better for countries with a large number of tweets. In Fig. 3, we plotted the precision and recall value for each country as a scatter chart. The dot size is proportional to the number of tweets that come from that country. Turkey appears to be the country with highest precision and recall. These results suggest that our approach works better with more data samples.

The same graph is also plotted for city prediction in Fig. 3. Because of the skewness of our data and the difficulty of city-level prediction, our classifier tends to generate labels towards big cities, which leads to high recall and low precision for cities like Los Angeles.

For real world applications, people may ask how we could set a threshold to get prediction results with high confidence. To answer this question, we further examined the relation between prediction accuracy and the output probability. Here the output probability is just the probability of our predicted location calculated by Eq. 2. Figure 4 shows the distribution of tweets in terms of output probability for two tasks. As expected, the prediction accuracy increases as the output probability increases. We get 97.2% accuracy for country-level prediction with output probability larger than 0.9. Surprisingly, the accuracy of city-level is as high as 92.7% for the 29.6% of the tweets with output probability greater

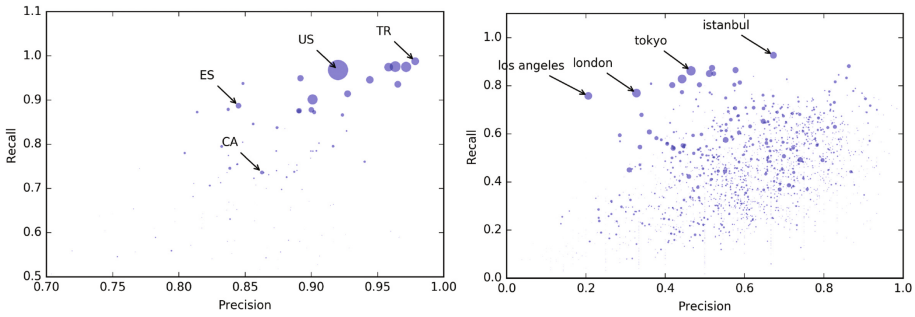


Fig. 3. Two scatter graphs that show the performances for each country and cities. The x-axis is precision, y-axis is recall. Each dot represents a country/city. The dot size is proportional to the number of tweets that comes from the corresponding location. Some tiny invisible country outside of the scope are not shown in the figure.

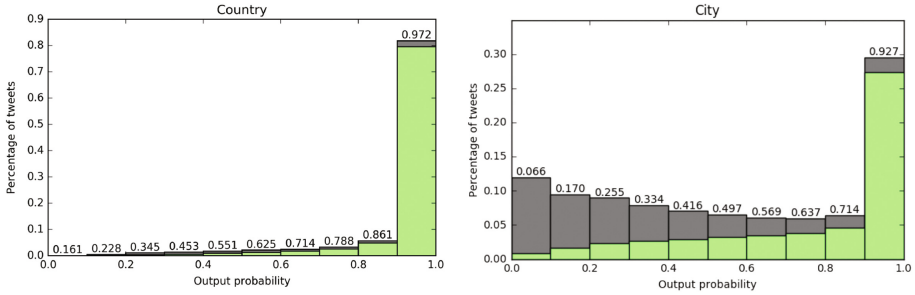


Fig. 4. Two bar charts that show the distribution of tweets in terms of the output probability. The x-axis is the output probability associated with each prediction, the y-axis is the percentage of tweets. The height of grey bar represents the percentage of test data that has certain output probability. The height of green bar represents the percentage of correctly predicted tweets in each probability range. We listed the accuracy for each probability range above the bar. Take the rightmost bar in country-level prediction for example, there are 81.8% tweets' country are predicted with output probability larger than 0.9. Among these 81.8% tweets, 97.2% are predicted correctly.

than 0.9. However, the city-level accuracy for the remaining tweets with output probability less than 0.9 is only 48.4%. Unlike country-level prediction, the number of tweets decreases as output probability increases, unless the output probability is larger than 0.9.

6 Discussion and Conclusion

These experiments demonstrate that our approach is consistently better than the prior method thus supporting more tweets to be accurately located by country, and city, of origin. At the country level, the more tweets that come from the country, the better the prediction. Regardless of the number of tweets per

country, we can predict the country location for most tweets with extremely high confidence and accuracy. At the city level the results are more mixed. For a small fraction of tweets we can get greater than 90% accuracy, but for the rest of tweets the accuracy is less than 50%. For about half the tweets it is difficult to infer the city location. This result is partially due to the fact that we base the prediction on only a single tweet. Future work may consider using collection of tweets per user. This result is also partially due to the fact that the data is highly skewed toward a few cities. Future work should develop a training set that is more evenly distributed across cities. Despite these limitations, this approach shows promise.

This paper presents a method for geo-locating a single tweet based on the information in a tweet JSON object. The proposed approach integrates tweet text and user profile meta-data into a single model. Compared to the previous stacking method with feature selection, our approach substantially outperforms the baseline method. We developed the approach for both city and country level and demonstrated the ability to classify tweets at both levels of granularity. The results demonstrate that using a convolutional neural network utilizes the textual location information better than previous approaches and boosts the location prediction performance substantially.

Acknowledgments. This work was supported in part by the Office of Naval Research (ONR) N000140811186, and the National Science Foundation (NSF) 00361150115291. 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 NSF. We want to thank tutors in the Global Communication Center at Carnegie Mellon for their valuable advice.

References

1. Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.H., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707. IEEE (2011)
2. Berggren, M., Karlgren, J., Östling, R., Parkvall, M.: Inferring the location of authors from words in their texts. arXiv preprint arXiv:1612.06671 (2016)
3. Bo, H., Cook, P., Baldwin, T.: Geolocation prediction in social media data by finding location indicative words. In: Proceedings of COLING, pp. 1045–1062 (2012)
4. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM (2010)
5. Culotta, A., Kumar, N.R., Cutler, J.: Predicting the demographics of twitter users from website traffic data. In: AAAI, pp. 72–78 (2015)
6. Earle, P.S., Bowden, D.C., Guy, M.: Twitter earthquake detection: earthquake monitoring in a social world. Ann. Geophys. **54**(6) 211 (2012)
7. Hale, S., Gaffney, D., Graham, M.: Where in the world are you? Geolocation and language identification in twitter. In: Proceedings of ICWSM 2012, pp. 518–521 (2012)
8. Han, B., Cook, P., Baldwin, T.: A stacking-based approach to twitter user geolocation prediction. In: ACL (Conference System Demonstrations), pp. 7–12 (2013)

9. Han, B., Cook, P., Baldwin, T.: Text-based twitter user geolocation prediction. J. Artif. Intell. Res. **49**, 451–500 (2014)
10. Hong, L., Ahmed, A., Gurumurthy, S., Smola, A.J., Tsioutsiouliklis, K.: Discovering geographical topics in the twitter stream. In: Proceedings of the 21st International Conference on World Wide Web, pp. 769–778. ACM (2012)
11. Jurgens, D.: That's what friends are for: inferring location in online social media platforms based on social relationships. In: ICWSM 2013, pp. 273–282 (2013)
12. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
13. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Landwehr, P.M., Carley, K.M.: Social media in disaster relief. In: Chu, W.W. (ed.) Data Mining and Knowledge Discovery for Big Data, pp. 225–257. Springer, Heidelberg (2014)
15. Mikolov, T., 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, pp. 3111–3119 (2013)
16. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.N.: Understanding the demographics of twitter users. In: 5th ICWSM 2011 (2011)
17. Qian, Y., Tang, J., Yang, Z., Huang, B., Wei, W., Carley, K.M.: A probabilistic framework for location inference from social media. arXiv preprint arXiv:1702.07281 (2017)
18. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our twitter profiles, our selves: predicting personality with twitter. In: 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust (PASSAT), pp. 180–185. IEEE (2011)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Elsevier, San Francisco (2014)
20. 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, pp. 1500–1510. Association for Computational Linguistics (2012)
21. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Cogn. Model. **5**(3), 1 (1988)
22. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
23. Tsou, M.H., Yang, J.A., Lusher, D., Han, S., Spitzberg, B., Gawron, J.M., Gupta, D., An, L.: Mapping social activities and concepts with social media (twitter) and web search engines (yahoo and bing): a case study in 2012 US presidential election. Cartography Geogr. Inf. Sci. **40**(4), 337–348 (2013)
24. Wing, B.P., Baldridge, J.: Simple supervised document geolocation with geodesic grids. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 955–964. Association for Computational Linguistics (2011)
25. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)

Stigmergy-Based Modeling to Discover Urban Activity Patterns from Positioning Data

Antonio Luca Alfeo[1(✉)], Mario Giovanni C.A. Cimino[1], Sara Egidi[1], Bruno Lepri[2], Alex Pentland[3], and Gigliola Vaglini[1]

[1] University of Pisa, largo Lazzarino 1, Pisa, Italy
luca.alfeo@ing.unipi.it, {mario.cimino,gigliola.vaglini}@unipi.it, s.egidi1@studenti.unipi.it
[2] Bruno Kessler Foundation, via S. Croce, 77, Trento, Italy
lepri@fbk.eu
[3] M.I.T. Media Laboratory, Cambridge 02142, USA
pentland@media.mit.edu

Abstract. Positioning data offer a remarkable source of information to analyze crowds urban dynamics. However, discovering urban activity patterns from the emergent behavior of crowds involves complex system modeling. An alternative approach is to adopt computational techniques belonging to the emergent paradigm, which enables self-organization of data and allows adaptive analysis. Specifically, our approach is based on stigmergy. By using stigmergy each sample position is associated with a digital pheromone deposit, which progressively evaporates and aggregates with other deposits according to their spatiotemporal proximity. Based on this principle, we exploit positioning data to identify high-density areas (hotspots) and characterize their activity over time. This characterization allows the comparison of dynamics occurring in different days, providing a similarity measure exploitable by clustering techniques. Thus, we cluster days according to their activity behavior, discovering unexpected urban activity patterns. As a case study, we analyze taxi traces in New York City during 2015.

Keywords: Urban mobility · Stigmergy · Emergent paradigm · Hotspot · Pattern mining · Taxi-GPS traces

1 Introduction

The increasing volume of urban human mobility data arises unprecedented opportunities to monitor and understand crowd dynamics. Identifying events which do not conform to the expected patterns can enhance the awareness of decision makers for a variety of purposes, such as the management of social events or extreme weather situations [1]. For this purpose GPS-equipped vehicles provide a huge amount of reliable data about urban human mobility, exhibiting correlation with people daily life, events, and city structure [2]. The majority of the methods approaching the analysis of vehicle traces can be grouped into three

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 292–301, 2017.
DOI: 10.1007/978-3-319-60240-0_35

categories: *cluster-based*, *classification-based*, and *pattern mining-based*; whereas the main application problems include the hotspot discovery, the extraction of mobility profiles, and the detection and monitoring of big events and crowd behavior [3]. For example, in [4] the impact of a social event is evaluated by analyzing taxi traces. Here, the authors model typical passenger flow in an area, in order to compute the probability that an event happens. Then, the event impact is measured by analyzing abnormal flows in the area via Discrete Fourier Transform. In [5] GPS trajectories are mapped through an Interactive Voting-based Map Matching Algorithm. This mapping is used for off-line characterization of normal drivers' behavior and real-time anomaly detection. Furthermore, the cause of the anomaly is found exploiting social network data. In [6] the authors use a Multiscale Principal Component Analysis to analyze taxi GPS data in order to detect traffic congestion.

One of the main issues concerning the analysis of this kind of data is their dimensionality. Many approaches handle it by focusing on specific areas (*hotspots*) whose high concentration of events and people can summarize mobility dynamics [7]. As an example, in [8] a density-based spatial clustering is employed to perform spatiotemporal analysis on taxi pick-up/drop-off to find seasonal hotspots. Authors in [9] use OPTICS algorithm in order to detect city hotspots as density-based clusters of taxi drop-off positions. Recently, in [10] an Improved Auto-Regressive Integrated Moving Average algorithm is proposed; it is aimed to detect urban mobility hotspots via taxi GPS traces and analyze the dynamics of pick-ups in dense locations of the city. However, due to the complexity of human mobility data, the modeling and comparison of their dynamics over time remain hard to manage and parametrize [11]. In this paper, we present an innovative approach based on *stigmergy* [12] that aims to handle both complexity and dimensionality of these data, providing an analysis of urban crowds dynamics by exploiting taxi GPS data. Specifically, our investigation covers the city hotspots identification, the characterization of their activity over time and the unfolding of unexpected activity pattern.

The paper is structured as follows. In Sect. 2 the architectural view of our approach is described. In Sect. 3 the experimental studies and results are presented. Finally, Sect. 4 summarizes conclusions and future work.

2 Approach Description

In this section, we present our approach, based on the principle of *stigmergy*. Stigmergy is an indirect coordination mechanism used in social insect colonies [12]. It is based on the release of chemical markers (*pheromones*), which aggregate when subsequently deposited in proximity with each other. This mechanism can be employed in the context of data processing, providing self-organization of data [13] while unfolding their spatial and temporal dynamics [14]. By exploiting stigmergy, we discover city hotspots, characterize their activity dynamics (i.e. presence of people over time) and assess unexpected activity patterns. In order to focus on activity dynamics, we employ New York City taxi positioning data,

considering the amount of passengers together with the GPS position of each pick-up/drop-off.

2.1 Hotspot Detection

At the beginning, data samples are transformed in digital pheromone deposits, allowing the progressive emergence of city hotspots (i.e. the most high-density areas within the city). Firstly, data are treated by the smoothing process (Fig. 1), in order to remove insignificant activity levels and highlight relevant dynamics. This process is implemented by applying a sigmoidal function to the samples. Then, a mark is released in correspondence of each smoothed sample in a three-dimensional virtual environment. Marks are defined by a truncated cone with a given width and intensity (height) equal to data sample value. The trailing process aggregates marks, forming a *stigmergic trail*, which is characterized by evaporation (i.e. temporal decay δ) and defined as $T_i = (T_{i-1} - \delta) + Mark_i$.

As an effect, isolated marks tend to disappear, whereas the arrival of new marks in a given region counteracts the evaporation. Thus, aggregation and evaporation can act as an agglomerative spatiotemporal clustering with historical memory. Hotspots are identified as the city areas corresponding to the overlapping of the most relevant trails obtained by processing data in early morning (i.e. 3am–8am), morning (i.e. 9am–2pm), afternoon/evening (i.e. 3pm–8pm), and night (i.e. 9pm–2am) time slots. As an example, Fig. 1 shows the hotspots identified in Manhattan (New York City). Their locations correspond to: East Harlem - Upper East Side (A), Midtown East (B), Broadway (C), East Village - Gramercy - MurrayHill (D), Soho - Tribeca (E), Chelsea (F) and Time Square - Midtown West - Garment (G).

2.2 Hotspot Activity Characterization

For each identified hotspot, we generate the activity time series, by periodically collecting the amount of activity occurred in the hotspot during a day. Let us consider an activity time series; what is actually interesting is not the continuous variation of the activity over time, but the transition from one type of behavior to another.

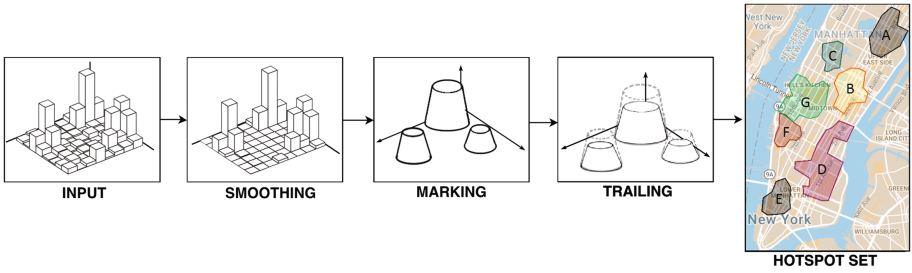


Fig. 1. The stigmergy-based process of hotspot discovery.

Generally, given a time window each hotspot behavior can be characterized by an ideal time series segment of hotspot activity representing that specific behavior. More formally, we define it as an *archetype*. An example of an archetype is *asleep* behavior, which usually occurs during the night, between the calming down of the nightlife and the arrival of the workers; here the city exhibits its lowest activity level.

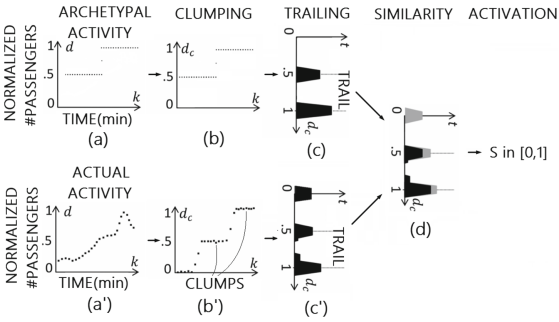


Fig. 2. The architecture of a SRF.

In order to detect an archetypal behavior in hotspot activity time series, we design a processing schema called Stigmergic Receptive Field (SRF), because it is receptive to a specific archetype and it processes samples employing the principle of stigmergy. Specifically, SRF computes a degree of similarity between a specific archetype (Fig. 2a) and an activity time series (Fig. 2a′), by subsequently processing their samples, which are assumed to be normalized between 0 and 1.

First, samples undergo the *clumping process* (Fig. 2b and b′), which acts as a sort of soft discretization creating clumps of samples. Clumps arrangement can be parametrized allowing to fit the analysis over the archetype's levels of interest. The clumping can be implemented as a double sigmoidal function. Second, the marking process (Fig. 2c and c′) enables the release of a mark in a bi-dimensional virtual environment in correspondence of the sample value. The mark can be implemented by a trapezoid with given intensity (height) and width ϵ. Third, the trailing process accumulates marks creating the trail structure, whose intensity decays (i.e. evaporates) of a given rate δ at each step of time. As an effect, evaporation rate and mark width allow the trail to capture coarse spatiotemporal structure in data, handling micro-fluctuations. Fourth, current T_{act} and archetypal trails T_{arc} are compared by the similarity process (Fig. 2d), by using the Jaccard coefficient $S = |T_{arc} \cap T_{act}|/|T_{arc} \cup T_{act}|$ [15]. This coefficient provides a measure of similarity between 1 (identical trails) and 0 (non-overlapping trails). Finally, the activation process is applied to enhance only relevant similarity values and remove insignificant values according to the activation thresholds α_a, β_a. This process can be implemented by using the already mentioned sigmoidal function, i.e. $f(x, \alpha_a, \beta_a) = 1/(1 + e^{-\alpha_a(x-\beta_a)})$.

In order to provide an effective similarity, the SRF's parameters have to be properly tuned. With this aim, the Adaptation process uses the Differential Evolution (DE) to adapt the structural parameters of the SRF: (i) the clumping inflection points α, β, γ, λ; (ii) the mark width ϵ; (iii) the trail evaporation δ; (iv) the activation thresholds α_a, β_a. The aim of DE is to minimize the mean square error (MSE), considering the error as the difference between the target $\hat{S}$ and the computed S similarity values over a set of M labeled time series, i.e. $Fitness = \sum_{i=1}^{M}(|S_i - \hat{S}_i|^2)/M$. The target similarity value is 1 if the current time series exhibits the archetypal behaviour, 0 otherwise.

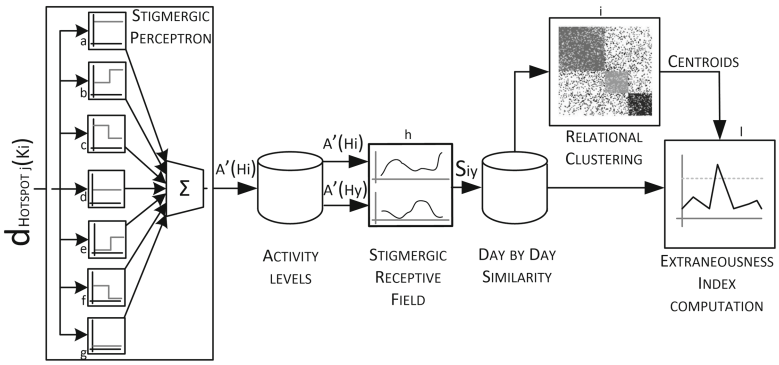


Fig. 3. The overall processing of activity samples.

Since any real signal is usually similar to more than one archetype, a collection of SRFs, specialized on different archetypes and ordered for increasing activity, is arranged in a connectionist topology to make a Stigmergic Perceptron (Fig. 3). Specifically, adopted archetypes are: Asleep (Fig. 3g), i.e. the hotspot at its lowest activity level; Falling (Fig. 3f), i.e. the flow just before the city activity calms down; Awakening (Fig. 3e), i.e. the waking up of urban life after a calm phase; Flow (Fig. 3d), i.e. the hotspot at its operating capacity, usually exhibited during working hours; Chill (Fig. 3c), which usually occurs after a rush hour, when people leave work and take taxis to return home; Rise (Fig. 3b), i.e. the hotspot transition to its most intense activity level; and Rush-Hour (Fig. 3a), which usually occurs in early morning and late afternoon, when people movement is at its highest rate. A perceptron computes a single output from multiple inputs, by forming a linear combination of them. Similarly, the stigmergic perceptron (SP) combines linearly SRFs' outcomes by computing their weighted mean, using the provided similarities S_i as weights, i.e. $ActivityLevel = \sum_{i=1}^{N}(S_i * i)/\sum_{i=1}^{N}(S_i)$. The resulting value is called activity level and is defined between zero and N, where N is the number of SRFs. An important aspect concerning hotspot activity level computation is to train each SRF inside a SP in order to prevent multiple activations of SRFs. Let us consider the most sensitive SRF parameter, i.e. the evaporation δ. High evaporation

prevents marks aggregation and pattern reinforcement, while low evaporation causes the saturation of the trail. In order to handle this sensitivity, the adaptation of each SRF inside a SP is twofold: (i) the Global Training phase is aimed to determine an interval for the evaporation rate of each SRF. The interval $[\delta_{min}, \delta_{max}]$ is obtained considering the narrowest interval including the fitness values above its 90th percentile, while the intervals for the other parameters can be statically assigned on the basis of application domain constraints; and (ii) the Local Training phase aims to find the optimum values for every module of each single SRF, by using the interval generated in the Global Training phase. As a result, a proper trained Stigmergic Perceptron provides the characterization of hotspot activity, by transforming a given time series of activity samples in a new time series of activity levels. In order to compute the overall similarity between hotspot activity levels gathered in two different days, we employ a further SRF (Fig. 3h) which uses one activity level time series just like it was an archetype. The adaptation in this specific SRF tunes mark width ϵ, trail evaporation δ, and activation thresholds α_a and β_a. As fitness function, we use the Mean Squared Error (MSE) between computed and ideal similarity over a set of labeled pairs of activity time series (i.e. the training set).

2.3 Unexpected Patterns Detection

Exploiting the mechanism described above, we generate the similarity matrix, that is the collection of similarities obtained by matching with each other the activity level time series of the training set. Provided similarity matrix can be processed by a fuzzy relational clustering technique, grouping days according to their daily activity similarity. Specifically, we employ Fuzzy C-Mean to compute the clusters centroid. The number of clusters corresponds to the number of daily activity behaviors taken into account in the analysis. Based on these centroids, the membership degrees of further daily activity level time series can be computed. The membership degrees are between 0 (not belonging to the cluster) and 1 (completely belonging to the cluster). By exploiting the membership degrees u_n as a distance, we measure the extraneousness of current activity level with respect to its expected cluster. The Extraneousness Index (EI) is defined as the Manhattan Distance between current daily activity level series d and the centroid of the cluster in which current day is assumed to belong. In Eq. 1, the computation case with 3 clusters is shown.

$$EI(d) = (|u_1(d) - u_1(C_2)| + |u_2(d) - u_2(C_2)| + |u_3(d) - u_3(C_3)|)/2 \quad (1)$$

We define as an Unexpected Pattern a day characterized by an activity level whose EI exceeds the maximum EI computed over the training set.

3 Experimental Studies and Results

We have analyzed a dataset of taxi traces provided by the Taxi and Limousine Commission of New York City, which contains information about all medallion taxi trips from 2009 to 2016 [16]. We focus our investigation on dynamics

occurred during 2015 in Manhattan considering that it attracts the most of the taxi trips in New York City. A pre-processing step has been performed to remove missing values and discretize data in spatiotemporal bins defined as a squared area 10-foot- wide with duration of 5 min. Then, the min-max normalization is applied. In order to search for hotspots characterizing every possible city routine (i.e. summer and winter ones), the hotspot discovery procedure has been performed comprising data gathered in working days and week-ends of February 2015 and June 2015.

Since archetypes are assumed to be general, the training set for the SP's global and local phases is generated by using the pure archetype time series as seeds and applying spatial noise and temporal shift.

In order to validate the SP archetypal behavior detection, a set of time series have been manually labeled and the difference with the actual results of the SP is used to evaluate detection error. Each label corresponds to the expected SP result according to the archetypal behavior visually detected in current time series (i.e. 1 if Asleep, 2 if Falling, and so on). To this purpose, 35 time series (i.e. 5 for each archetype) have been provided to the SP. The obtained MSE is shown in Table 1. By considering the activity level operative range (i.e. [1,7]) and the provided MSE values, the system shows good detection performances, proving the functional effectiveness of the SRF and the SP.

Table 1. Mean square error in archetypal behavior detection via SP.

Archetype	Asleep	Falling	Awakening	Flow	Chill	Rise	Rush-hour	TOT
MSE	0.215	0.029	0.029	0.028	0.166	0.020	0.143	0.633

In the next processing phase, a further SRF is aimed to assess the similarity between daily activity levels. It is provided with a training set obtained by selecting a set of pairs of daily activity levels. In order to supply a clustering process, such SRF is trained to distinguish similar and dissimilar signals, according to the behavioral class of daily activity levels, namely: (i) Working days (expected to fall between Monday and Tuesday), when crowd movements are mainly caused by working routines; (ii) Entertainment days (expected to fall on Friday and Saturday), in which people tend to spend the night out; (iii) Leisure days (expected to fall on Sunday), which are characterized by limited transportation usage. Their target similarity is 1 if days belong to the same behavioral class, 0 otherwise. Since the defined classes refer to the cyclical sequence of week days, our ground truth can be provided by the calendar itself. The 10% of computed daily activity levels have been used to create these pairs (i.e. 1296 pairs overall).

The Fuzzy C-Mean algorithm is used to group days according to their stigmergy-based similarity in order to arrange them among the three provided clusters, namely: Working, Entertainment and Leisure days. Upon this, we exploit the Extraneousness Index in Eq. 1 to assess unexpected patterns.

We show results obtained analyzing hotspot D, since it is characterized by multiple usages [17] allowing the displaying of every activity level behavioral class. Interestingly, this area is also found to be an hotspot by [9].

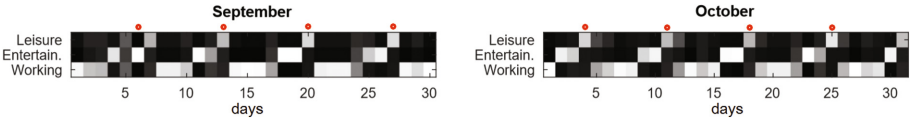


Fig. 4. Membership degrees of days in September and October. The whitest, the higher.

Figure 4 shows the computed membership degree for each cluster, obtained with days in September and October. The whitest the box, the higher the degree. Clearly, the stigmergy-based characterization of hotspot daily activity allows to cluster days according to their behavioral class which corresponds to the arrangement we assumed. Indeed, most of the Sundays (highlighted by a circle in Fig. 4) exhibit their highest membership degree with Leisure day cluster. The same happened with days in Entertainment and Working cluster. It is worth noting that provided approach allows the mapping of daily behaviors to emerge from data instead of being explicitly injected into the system.

However, some days does not confirm this behavior. Indeed, by comparing their EI with the maximum EI in the training set (red line in Fig. 5), they are recognized as an unexpected pattern (red spot in Fig. 5).

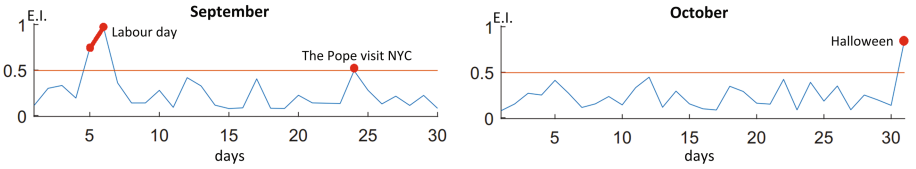


Fig. 5. Extraneousness Index computed over days in September and October. (Color figure online)

Table 2 shows the most relevant unexpected patterns detected by analyzing the whole year 2015. Each unexpected pattern date is shown together with their most probable cause, such as an occurred social event. EI provides a continuous measure of the magnitude of unexpected patterns, allowing the comparison of their impact on hotspot activity dynamics. As an example, Easter affects the activity in hotspot D much more than the NYC Half Marathon. Indeed, the greatest Easter celebrations in NYC are kept by the St. Patrick Cathedral, which is located in the area corresponding to hotspot D, whereas this area was not directly involved in the NYC Half Marathon 2015. By repeating the analysis in

Table 2. Most relevant unexpected patterns detected all over 2015.

EI	Date and occurred city event
0.96	06-Sep, Labour day
0.94	24-May, Memorial day
0.86	31-Oct, Halloween
0.83	26-Nov, Thanksgiving
0.83	28-Jun, Gay Pride
0.82	25-Dec, Christmas
0.81	01-Jan, New Year's Eve
0.80	04-Apr, Easter (holy Saturday)
0.79	27-Jan, Winter Storm Juno [18]
0.74	05-Sep, Labour day celebrations
0.63	03-Jul, Independence day
0.63	31-Dec, New Year's Eve
0.61	15-Mar, NYC Half Marathon
0.49	24-Sep, Pope Francis visit NYC

the same date on hotspot C, the computed EI results roughly 60% higher (i.e. 0.96); indeed the zone corresponding to hotspot C was directly crossed by NYC Half Marathon 2015.

4 Conclusion

In this paper, we proposed a novel approach aimed to provide knowledge discovery in the context of human urban mobility data. In contrast with the literature in the field, our approach does not require the in-depth modeling of the dynamics under investigation since it relies on data self-organization provided by employing the principle of stigmergy. Indeed, by using stigmergy, the spatiotemporal density in data has been exploited to identify city hotspots and characterize their dynamics, allowing to generate data-driven prototypes of typical daily activity. By treating them via a clustering technique, we were able to discern expected patterns from unexpected ones, which were found to be usually related to various events. One of the most promising improvements for this investigation can be achieved by cross-checking results obtained via vehicle GPS data with other data sources (e.g. social media or car crash data). Indeed, by employing a more detailed ground truth, the system can be specialized to model and detect patterns characterized by a timescale shorter than a daily one.

References

1. Sagl, G., Loidl, M., Beinat, E.: A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. ISPRS Int. J. Geo-Inf. **1**(3), 256–271 (2012)
2. Veloso, M., Phithakkitnukoon, S., Bento, C.: Urban mobility study using taxi traces. In: Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis, pp. 23–30. ACM (2011)
3. Mazimpaka, J.D., Timpf, S.: Trajectory data mining: a review of methods and applications. J. Spat. Inf. Sci. **2016**(13), 61–9 (2016)
4. Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., Wu, Z.: City-scale social event detection and evaluation with taxi traces. ACM Trans. Intell. Syst. Technol. **6**(3), 40 (2015)
5. Pan, B., Zheng, Y., Wilkie, D., Shahabi, C.: Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 344–353. ACM (2013)
6. Kuang, W., An, S., Jiang, H.: Detecting traffic anomalies in urban areas using taxi GPS data. Math. Prob. Eng. **2015**, Article ID 809582, 1–13 (2015). doi:10.1155/2015/809582
7. Hu, Y., Miller, H.J., Li, X.: Detecting and analyzing mobility hotspots using surface networks. Trans. GIS **18**(6), 911–935 (2014)
8. Lu, Y.: An intelligent system for taxi service monitoring, analytics and visualization. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (2016)
9. Keler, A., Krisp, J.M.: Is there a relationship between complicated crossings and frequently visited locations? A case study with boro taxis and OSM in NYC. In: 13th International Conference on Location-Based Services (2016)
10. Li, X., Pan, G., Wu, Z., Qi, G., Li, S., Zhang, D., Zhang, W., Wang, Z.: Prediction of urban human mobility using large-scale taxi traces and its applications. Front. Comput. Sci. **6**(1), 111–121 (2012)
11. Castro, P.S., Zhang, D., Chen, C., Li, S., Pan, G.: From taxi GPS traces to social and community dynamics: a survey. ACM Comput. Surv. **46**(2), 17 (2013)
12. Marsh, L., Onof, C.: Stigmergic epistemology, stigmergic cognition. Cogn. Syst. Res. **9**(1–2), 136–149 (2008)
13. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: implications for the autonomous development of mental capabilities in computational agents. IEEE Trans. Evol. Comput. **11**(2), 151–180 (2007)
14. Barsocchi, P., Cimino, M.G.C.A., Ferro, E., Lazzeri, A., Palumbo, F., Vaglini, G.: Monitoring elderly behavior via indoor position-based stigmergy. Pervasive Mob. Comput. **23**, 26–42 (2015). Elsevier Science
15. Niwattanakul, S., Singthongchai, J., Naenudorn, E., Wanapu, S.: Using of Jaccard coefficient for keywords similarity. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, vol. 1, p. 6 (2013)
16. NYC.gov: Taxi and Limousine Commission (TLC) Trip Record Data. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
17. Zola: interactive map for zone and land use of NYC. http://maps.nyc.gov/doitt/nycitymap/template?applicationName=ZOLA
18. Weather NYC: Thousands of transatlantic travellers face serious disruption caused by New York winter storm 'Juno'. The Independent, 26 January 2015

Prospective Detection of Foodborne Illness Outbreaks Using Machine Learning Approaches

Aydin Teyhouee[1(✉)], Sara McPhee-Knowles[1], Chryl Waldner[2], and Nathaniel Osgood[1]

[1] Department of Computer Science, University of Saskatchewan, Saskatoon, Canada
{ayt227,sam123}@mail.usask.ca, osgood@cs.usask.ca
[2] Western College of Veterinary Medicine,
University of Saskatchewan, Saskatoon, Canada
chryl.waldner@usask.ca

Abstract. Despite advances in food safety regulations, food-borne illness imposes a heavy health burden, with nearly 50 million estimated incident cases of illness each year. Having a prospective foodborne illness outbreak detection mechanism for more accurate and timely triggering of outbreak control measures would offer notable public health dividends, but is challenging due to the subclinical character of most foodborne illnesses. Within this work, collected synthetic datasets of incident illness cases and vendor contamination records from a previously contributed and empirically grounded model of foodborne illness, are used to study the efficacy of Hidden Markov Models (HMMs) for syndromic surveillance monitoring and disease outbreak detection under two data collection regimes, one involving a sentinel population using smartphone-based app for tracing location of food consumption and subclinical reporting. A support vector machine (SVM) approach was applied to compare the results to the HMM. Findings suggest that while reliance on clinical data offers poor potential for automatic outbreak detection, the use of HMMs offer excellent potential for detecting foodborne illness outbreak when informed by subclinical reporting by even a very small (4% of population) sentinel group. By contrast, SVM offers relatively poor prospects for detection. Furthermore, experiments with an empirically grounded agent-based model suggest that use of an HMM may be advantageous for triggering outbreak investigations among public health inspectors.

Keywords: Foodborne illness · Outbreak · Prospective detection · Machine learning · Hidden Markov model

1 Introduction

Each year, a large population worldwide suffer from foodborne illness.

While the public health inspection regime of food vendors successfully prevents many potential illnesses, the dynamic nature of restaurants kitchens, the

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 302–308, 2017.
DOI: 10.1007/978-3-319-60240-0_36

human resource constraints on carrying out consecutive inspections and the time-consuming character of the inspection process allow violations to remain undetected and limit the completeness of food illness prevention. Moreover, numerous food poisoned people who show mild to moderate symptoms of illness never show up at clinics and health care centers, but are greatly curtailed in their activity. While such subclinical cases impose stiff health, quality of life and economic costs, the absence of such data (of this kind of illness) in public health incidence records makes it almost impossible to figure out a potential outbreak occurrence. On the other hand, outbreak prediction methods mostly rely on telephone interviews of the clinical registered poisoned patients, days or weeks after their illness. This makes the situation even worse in two ways. First, the patient will be subject to forgetfulness about food vendors visited during a specified time, making it hard to prioritize the most probable contaminated restaurants in an investigation. Second, and a consequence, because of inaccuracies in the data collected and the prolonged investigation process, the adverse health and cost impacts of the outbreak will be magnified.

Lately, prospective detection of disease outbreaks in general using machine learning approaches has attracted the attention of researchers. The challenge on this new field is to diagnose the occurrence of an outbreak timely enough, helping the public health agents for taking quick outbreak controlling measurements. However, the applications of such works to foodborne illness outbreaks has been very limited [1]. Machine learning provides a set of tools which can be applied in different problem domains for data analysis. Given that the challenge of detecting foodborne illness outbreaks consists of identifying the evolution of the categorical latent state (outbreak vs. non-outbreak) of a system (municipality) over time in the light of noisy observations (incident cases) strongly influenced by state, Hidden Markov Models (HMMs) offer a particularly attractive analysis lens. Here we seek to distinguish between these outbreak and non-break states based on our observation of the number of reported illnesses. Although this is not the main goal of our presented work, we compare the findings of the HMM with the results of a Support Vector Machine (SVM) model, which fails to take into account the temporal context of data. To achieve this end, we will use synthetic ground truth data from a previously contributed empirically-grounded agent-based model (ABM) of foodborne illness [2]. As Sara M. Knowles shows in her model [2], and reflecting more recent successes in fieldwork by the authors, we further and mainly investigate how use of sentinel reporting of subclinical illnesses via smartphones could improve our inference about the potential outbreaks. To evaluate this, we will simulate two data collection regimes. The first regime focuses complements such traditional data with reports of subclinical illnesses provided by a small sentinel population, constituting just 4% of the total population. While this first data collection regime could be carried out with a number of technologies such as designated social media channels, call-in lines, and web-based mechanisms, we note that such a system has been successfully utilized over many months by the authors in using the Ethica iEpi [3] smartphone-based epidemiological data collection system; this work is currently

being prepared for publication. In the second regime, we will use clinical data only, reflecting presentation by victims of possible foodborne illness to healthcare centers.

2 Overview of the Generative Model

Details of the foodborne illness ABM that serves to generate the synthetic time series that is used to train and test the machine learning models is described in a previous contribution [2]. However, we make some general comments about the model here. This model offers a stylistic depiction of a municipality that includes three types of actors: Persons, Restaurants and Inspectors. In the scenarios examined here, the municipality included a population of 5000 persons, 100 restaurants and one inspector. Restaurants can be either in a non-contaminated or contaminated state, with a transition hazard from the former to the latter such that an average of one restaurant per year becomes contaminated. The inspector can be in one of two modes: Routine inspection and outbreak response. In routine inspection mode, the inspector transitions between restaurants in a round-robin fashion In outbreak response mode, the inspector makes prioritized visits to restaurants according to the number of times that they have been identified (via faulty individual memory or via the geostamped records of sentinels) by those with clinical or (for the sentinel scenario) subclinical illness. Visits to restaurants are remembered by an individual. Independent of their source, foodborne illnesses developed in the model are classified clinical with a small probability (0.005), with the remainder remaining subclinical. Following a fixed period of time (2 days), individuals experiencing either of subclinical and clinical symptoms are treated as recovering, and return to a healthy state.

For analysis, each week, the model reports the incident case counts of clinical and subclinical illness and the count of contaminated restaurants.

3 HMM

HMMs are widely used in classification problems. Given a time horizon, they are used to infer the evolution of the system among a set of latent and non-observable categorical states over that horizon, with each of these states being associated with a specific distribution of observables. In this problem, we focus on discrete time characterization, with each time point representing a single week, and transitioning between two states s_t: a state in which the municipality includes a contaminated restaurant (henceforth termed the "outbreak" state) and $s_t = 1$, and one in which no contaminated restaurant is present and $s_t = 0$. Each such state is associated with a distribution for the observables $y_t(t = 1, ..., n)$: Clinical cases and (for the sentinel scenario) subclinical cases, where n is the n'th week. That is, for a given state s_t, $y_t|s_t \sim f_k(y_t; \theta_k)$, where $k \in \{0, 1\}$, f_k is a pre-specified density (e.g., univariate or multivariate Gaussian or Poisson) and θ_k are parameters to be estimated. The unobserved state space, $s_t(t = 1, ..., n)$ is modelled

by a two-state homogeneous Markov chain of order 1 with stationary transition probabilities $p_{kl} = P(s_{t+1} = l|s_t = k)$, where $k, l \in \{0, 1\}$ denote the two states of s_t (0: non-outbreak; 1: outbreak). Note that in this Markov-dependent mixture model, y_t is conditionally independent of all the remaining variables, given s_t. As we are working with counted data in this experiment (number of reported illnesses), the above mentioned f_k is a Poisson density. So, to make it short, the expected frequency profile for the events of any dataset is definable in the format of a Poisson where all the mentioned conditions are true. An attraction of HMMs is the fact that it is possible to estimate their parameters using a variety of parameter estimation methods including the iterative Expectation Maximization (EM) algorithm. A preinvestigation over the dataset and the histograms corresponding to each of the two datapoint clusters, one can observe: (a) A low level of illness occurrence, where the weekly incident case count can be modeled as a Poisson distribution with parameter λ_1, (b) A high level of illness occurrence, where the weekly incident case count can be modeled as a Poisson distribution with parameter λ_2.

The iterating process of converging the Poisson distributions' lambda parameter is performed by assigning the two above mentioned observed Poisson distribution parameters to specify a starting model for the EM algorithm: $\Omega_0 = (\pi_0, P_0, b_0)$, where π_0 is the initial matrix, P_0 is the (2×2) transition matrix and b_0 is the (1×2) emission matrix containing the first guessed lambda parameters for each of the Poisson distributions. In this study, a package named mhsmm [4] in the R statistical computing framework (R Development Core Team 2010) was used for parameter estimation. For training and cross-validation, the number of contaminated restaurants in successive weeks was rendered into a dichotomous variable serving as ground truth, assuming that any contaminated restaurant number greater than 1 corresponded to the state of an outbreak (whether declared or not). Also, the simulated 10,000-day (almost 27 years) dataset captured from the agent-based model was split up into training dataset (75%) and testing dataset (25%). To get an idea how good HMMs are performing in our problem possessing a temporal context, we utilize a Support Vector Machine (SVM) model with a linear kernel over our dataset. The results for both the models are presented in the next section.

4 Results

– Results of the Hidden Markov Model (Using both subclinical and clinical case counts):
Our HMM model $\Omega = (\pi_0, P_0, b_0)$ was initialized with $\pi_0 = (0.5\ 0.5)$, $P_0 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$ and $b_0 = (1\ 4)$. These parameters are used by the EM algorithm to produce a maximum likelihood estimate Hidden Markov Model to describe the data. We evaluated models in terms of confusion matrix, sensitivity and specificity resulting from a cross-validation procedure over the test data. The best model obtained for the scenario 2 (the case where our observation includes both clinical and subclinical instances) has a set of parameters

as follows: $\pi = (0\ 1)$, $P = \begin{pmatrix} 0.990 & 0.010 \\ 0.055 & 0.945 \end{pmatrix}$ and $b = (7.869088\ 15.860456)$, resulting in a sensitivity of 0.9318182, a specificity of 0.9840764 and a confusion matrix as per Fig. 1.

- Results of the Support Vector Machine Model (Using both subclinical and clinical case counts):

  The predictive performance of the SVM was measured through a cross-validation process over different cost values with 10-fold sampling method and then a model with lowest misclassification error rate with a linear kernel was chosen.

  This model obtained a sensitivity of 0.6590909, a specificity of 0.977707 and a confusion matrix as per Fig. 1 over the testing dataset.
- Results of the HMM and SVM (Using clinical case counts):

  In this scenario where only the clinical incidences were considered, both the HMM and SVM approaches failed in labeling the outbreak state. In this case, the number of reported clinical cases were very rare, and all incidences were labeled as non-outbreak state. Figure 1, shows the confusion matrix for this scenario.

	Total Population	Predicted Condition Positive	Predicted Condition Negative
HMM (Using both subclinical & clinical case counts)	Condition Positive	41	3
	Condition Negative	5	309
SVM (Using both subclinical & clinical case counts)	Condition Positive	29	15
	Condition Negative	7	307
HMM & SVM (sing clinical case counts)	Condition Positive	0	44
	Condition Negative	0	314

Fig. 1. Confusion matrix for different scenarios & ML models

5 HMM-Aided Outbreak Triggering System

To investigate whether the HMM could improve syndromic surveillance monitoring and linked disease outbreak detection systems, the ordinary illness triggering method (which is applied once at least 2 clinical cases happen) as mentioned in detail at Sect. 2, was replaced with the resulted HMM in Sect. 4. To carry out this HMM-based outbreak detection mechanism, the ABM uses the HMM parameters calculated in the HMM results of Sect. 4 to calculate the updated probability of being in an outbreak state in light of the previous value and the reported subclinical and clinical case counts. If the calculated probability at the beginning of any given week is greater than a threshold (which in this experiment was set to 0.6), a message is sent to the inspector to trigger the transition to the outbreak state investigation. The recorded cumulative count of clinical and subclinical illnesses over 10 years for 12 realizations in two different outbreak declaration

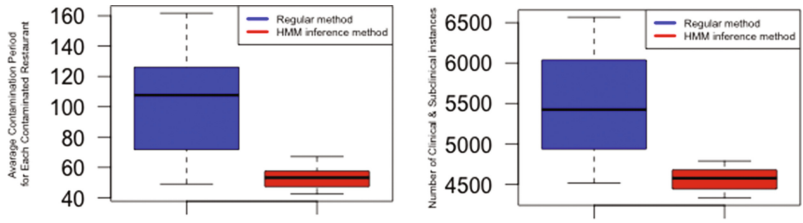


Fig. 2. Regular and HMM-based outbreak declaration comparison over 12 realizations for each: (Left) (Contamination period per contaminated restaurants [day/10-years]) - (Right) (Number of illness incidences [person/10-years]

regime (HMM outbreak triggering method and the ordinary method) is shown in the Fig. 2. Results show a quite significant decrease in the number of illness reports due to the fast detection of contaminated restaurants by applying the HMM outbreak declaration approach. As demonstrated by Fig. 2, this approach reflects a similar decrease in the time period a given contaminated restaurants remains contaminated before being identified and cleared.

6 Conclusion

Performing disease outbreak detection based on reported illness cases is an important function for syndromic surveillance systems. We treated the existence of a foodborne illness outbreak as a latent element of state and developed a Hidden Markov model for syndromic surveillance. We evaluated our disease outbreak detection approach using an empirically grounded previously contributed ABM of foodborne illness, comparing the results from HMM to those secured using an SVM approach. Finally, in light of the highly favourable results from the HMM, we further used the foodborne illness ABM to evaluate the public health gains secured through use of a HMM-based outbreak detection trigger, as compared with a traditional one based on case counts. Despite the highly noisy data present, and overlapping distributions of incident case counts between the outbreak and non-outbreak states, the results reported in this paper suggest a promising future for the use of hidden state variables to model the changing dynamics of observed surveillance time series, and for HMMs in general in outbreak signal detection. Moreover, the results from the first and second scenarios (considering both clinical and subclinical reports vs. considering only clinical reports) reveal that use of smartphones that can record locations and offer channels for reporting mild and moderate foodborne illness reports could improve our inference about the potential outbreaks. Finally, evaluation of HMM-based outbreak triggering mechanisms using ABMs suggest that significant public health gains may be secured when combining new technologies for syndromic surveillance with machine-learning based outbreak signal detection mechanisms. This work suggests promising lines of future work, including in extending our outbreak

detection approach with multiple data streams obtained from mobile applications, such as restaurant-specific traffic and illness counts.

References

1. Morrison, K., Charland, K., Okhmatovskaia, A., Buckeridge, D.: A framework for detecting and classifying outbreaks of gastrointestinal disease. Online J. Pub. Health Inform. **5**(1) (2013)
2. McPhee-Knowles, S.: The complex problem of food safety applying agent-based modeling to the policy process. Digital repository for the College of Graduate and Postdoctoral Studies electronic theses collection, University of Saskatchewan (2014)
3. Ethica Data: Ethicadata.ca, N.p. (2016). Accessed 29 Apr 2016
4. O'Connell, J., Hojsgaard, S.: Hidden semi Markov models for multiple observation sequences: the mhsmm package for R. J. Stat. Softw. **39**(4), 1–22 (2011)

Mitigating the Risks of Financial Exclusion: Predicting Illiteracy with Standard Mobile Phone Logs

Pål Sundsøy(✉)

Telenor Group Research, Big Data Analytics, Snarøyveien 30, 1331 Fornebu, Norway
paal@sundsoy.com

Abstract. The present study provides the first evidence that illiteracy can be predicted from standard mobile phone logs. By deriving a broad set of novel mobile phone indicators reflecting users' financial, social and mobility patterns this study addresses how supervised machine learning can be used to predict individual illiteracy in an Asian developing country, externally validated against a large-scale survey. On average the model performs 10 times better than random guessing with a 70% accuracy. Further it reveals how individual illiteracy can be aggregated and mapped geographically at cell tower resolution. In underdeveloped countries such mappings are often based on out-dated household surveys with low spatial and temporal resolution. One in five people worldwide struggle with illiteracy, and it is estimated that illiteracy costs the global economy more than $1 trillion dollars each year. These results potentially enable cost-effective, questionnaire-free investigation of illiteracy-related questions on an unprecedented scale.

1 Introduction

Illiterates are often trapped in a cycle of poverty with limited opportunities for income generation and financial inclusion. Literacy is also often a hurdle to bring financial services to the unbanked [1]. High-quality literacy statistics is therefore crucial to pinpoint areas where better education is needed: where are the illiterates? Mapping of literacy statistics is currently based on tedious household surveys with a low spatial and temporal frequency [2]. The increasing availability and reliability of new data sources, and the growing demand of comprehensive, up-to-date international literacy data are therefore of high priority. One of the most promising rich Big Data sources are mobile phone logs (CDRs) [3]. CDRs have shown to provide useful proxy indicators for assessing regional poverty levels [4, 5], socioeconomic status [6], unemployment [7, 8], infectious diseases [9] and disasters [10]. This study demonstrates how individual and regional illiteracy can be mapped using a combination of CDRs and financial airtime transactions.

The rest of this paper is organized as follows: Sect. 2 describes the methodological approach, including the features and modelling approach, while Sect. 3 addresses the research results, followed by concluding remarks in Sect. 4.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 309–314, 2017.
DOI: 10.1007/978-3-319-60240-0_37

2 Approach

2.1 Data

Household Survey Data: Data from two nationally representative cross-sectional household surveys of 200,000 individuals in a low-income South Asian country is analyzed. The data is collected at time Q114 and Q214 by an external survey company commissioned by the operator. The survey discriminates between 6 types of educations for the head of household, including being illiterate. The sample includes 6.8% illiterates, 40% primary degree, 26% SSC, 17.6% HSC, 5.6% bachelor, 3.5% master and 0.13% other degrees (incl. Ph.D.). The head of household's is asked for his or her most frequently used phone number. 87% of households in the country has at least one mobile phone.

Mobile Phone Data: Mobile phone logs for 76 000 of the surveyed 200 000 individuals belonging to the leading operator are retrieved from a period of six months and de-identified by the operator. Individual level features are built from the raw mobile phone data and is subsequently coupled with the corresponding de-identified phone numbers from the survey. The social features are subsetted from a graph consisting of in total 113 million subscribers and 2.7 billion social ties. No content of messages or calls are accessible and all individual level data remains with the operator.

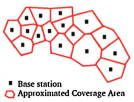
2.2 Features

A structured dataset consisting of 160 novel mobile phone features is built, and categorized into three dimensions: (1) financial (2) mobility and (3) social features, as shown in Table 1. The features are custom made to predict illiteracy, and include various parameters of the corresponding distributions such as weekly or monthly median, mean and variance.

2.3 Model Algorithm

Based on performance of many algorithms, including neural network and SVM, a gradient boosted machines model (GBM) is proposed as the final model [11]. To compensate class imbalance, the minority class in the *training set*, containing illiterates, is up-sampled from 6.8%. The minority class is then randomly sampled, with replacement, to be the same size as the majority class. A 10-fold cross-validation is used as re-sampling technique. In this set-up, each model is trained and tested using a 75/25 split. All results are reported for the test-set.

Table 1. Sample of features from mobile phone metadata used in model

Dimension	Features
Financial 	**Airtime purchases:** Recharge amount per transaction, Spending speed, fraction of lowest/highest recharge amount, coefficient of variation recharge amount etc **Revenue:** Charge of outgoing/incoming SMS, MMS, voice, video, value added sevices, roaming, internet etc. **Handset:** Manufacturer,brand, camera enabled, smart/feature/basic phone etc
Mobility 	Home district/tower, radius of gyration, entropy of places, number of places visited etc.
Social 	**Social Network:** Interaction per contact, degree, entropy of contacts etc. **General phone usage:** Out/In voice duration, SMS count, Internet volume/count, MMS count, video count/duration, value added services duration/count etc.

3 Results

3.1 Individual Illiteracy

Figure 1 shows the final features and their contribution in predicting illiteracy. Concretely, 19 of the features are related to illiteracy and included in the final GBM classifier. The model predicts whether phone users are illiterate with an accuracy of 70.1% (95% CI: 69.6–70.8). The deviation of accuracy from the training set is only 3.8%, which disregard model overfitting. The true positive rate (sensitivity/recall) is 71.6% and true negative rate (specificity) 70%.

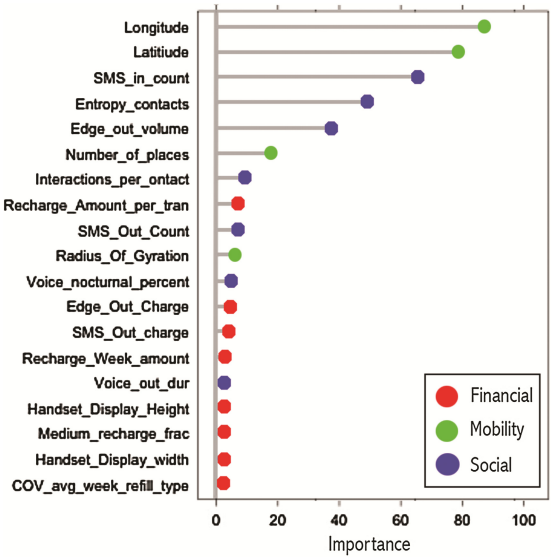


Fig. 1. Top features in the GBM Model colored by their respective feature family

An investigation of the most important predictors, as seen in Fig. 1, reveals some interesting associations. We especially notice that most frequently used longitude and latitude stand out as good predictors: where the people spend most of their time is a good signal of their education level. Another important feature is the number of incoming SMS, which (surprisingly) outperforms outgoing SMS. Moreover, we see that entropy of contacts is important – illiterates tend to concentrate their communication on few people. This is also in line with Eagle's work on geographical level [12], which shows that economic well-being is correlated with social diversity. Further we see that illiterates have limited use of internet (predictor 5), and their mobility pattern is limited to a few base stations (predictor 6).

3.2 Geographical Illiteracy Mapping

A natural next step is to move from individual illiteracy to geographical illiteracy. In big Asian cities there are often thousands of mobile towers that can be used as "sensors" to estimate illiteracy rates in the areas covered by the towers. In the rural areas where towers are less dense, interpolation techniques can be utilized to include information from the neighbour towers. Figure 2a shows the predicted illiteracy rate per tower, in one of the larger cities.

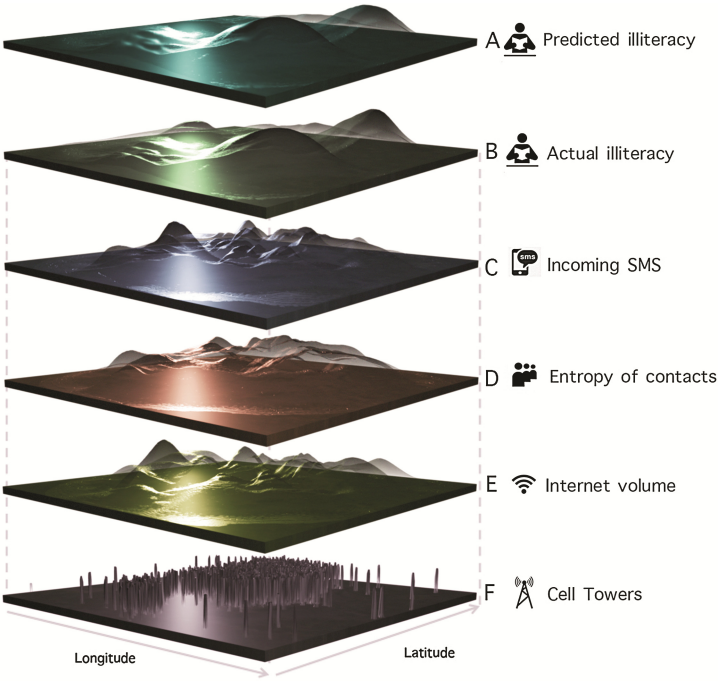


Fig. 2. Geographical mapping of illiteracy, top predictors and the cell tower distribution in one major Asian city. Height (z-axis) is proportional to the tower averages for each given metric.

The individual illiteracy rates are here calculated by using the test set, aggregated and averaged to tower level, and then further spatially interpolated, using an IDW algorithm, to average out the noise of local variations between towers. The *actual* illiterate rates in Fig. 2b is calculated by using the training set as ground truth. We notice three large pockets of larger illiteracy rates in the city. By also including distributions of the top predictors (Fig. 2c-e) it is possible to visually observe spatial correlations. For example, one can observe a large area of high SMS activity (Fig. 2c, left) that can be associated with low illiteracy rates.

4 Conclusion

This study shows how illiteracy can be predicted from mobile phone logs, purely by investigating users' metadata. By deriving economic, social and mobility features for each mobile user we predict individual illiteracy status with 70% accuracy. Further we show how individual illiteracy can be aggregated and mapped geographically with high spatial resolution on cell tower level. Feature investigation indicates that home cell tower and incoming SMS are the superior predictors, followed by diversity of communication

partners and Internet volume. An important policy application of this work is the prediction of regional and individual illiteracy rates in underdeveloped countries where official statistics is limited or non-existing.

References

1. Chibba, M.: Financial inclusion, poverty reduction and the millennium development goals. Eur. J. Dev. Res. **21**(2), 213–230 (2009)
2. IHSN: How (well) is Education Measured in Household Surveys? IHSN working paper 002 (2009)
3. Lokanathan, S., Lucas Gunaratne, R.: Behavioral insights for development from Mobile Network Big Data: enlightening policy makers on the State of the Art. (2014). SSRN 2522814
4. Blumenstock, J., Cadamuro, G., On, R.: Predicting poverty and wealth from mobile phone metadata. Science **350**(6264), 1073–1076 (2015)
5. Steele, J.E., Sundsøy, P., Pezzulo, C., Alegana, V., Bird, T., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.A., Iqbal, A., Hadiuzzaman, K., Lu, X., Wetter, E., Tatem, A., Bengtsson, L.: Mapping poverty using mobile phone and satellite data. J. R. Soc. Interface **14**(127), 20160690 (2017)
6. Sundsøy, P., Bjelland, J., Reme, B.A., Iqbal, A., Jahani, E.: Deep learning applied to mobile phone data for Individual income classification. In: ICAITA (2016)
7. Toole, J.L., Lin, Y.R., Muehlegger, E., Shoag, D., González, M.C., Lazer, D.: Tracking employment shocks using mobile phone data. J. R. Soc. Interface **12**(107), 20150185 (2015)
8. Sundsøy, P., Bjelland, J., Reme, B.A., Jahani, E., Wetter, E., Bengtsson, L.: Estimating individual employment status using mobile phone network data. arXiv preprint arXiv: 1612.03870 (2016)
9. Wesolowski, A., Qureshi, T., Boni, M.F., Sundsøy, P.R., Johansson, M.A., Rasheed, S.B., Engø-Monsen, K., Buckee, C.O.: Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proc. Natl. Acad. Sci. **112**(38), 11887–11892 (2015)
10. Lu, X., Wrathall, D.J., Sundsøy, P.R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Canright, G.S., Engø-Monsen, K., Bengtsson, L.: Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication, mobility and consumption patterns during cyclone Mahasen. Clim. Change **138**(3–4), 505–519 (2016)
11. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat., 1189–1232 (2001)
12. Eagle, N., Macy, M., Claxton, R.: Network diversity and economic development. Science **328**(5981), 1029–1031 (2010)

Multi-layer Network Composition Under a Unified Dynamical Process

Xiaoran Yan[1(✉)], Shang-Hua Teng[2], and Kristina Lerman[3]

[1] Indiana University Network Science Institute,
Indiana University, Bloomington, USA
everyxt@gmail.com
[2] Computer Science Department,
University of Southern California, Los Angeles, USA
[3] Information Sciences Institute, University of Southern California, Los Angeles, USA

Abstract. In this paper, we take a step towards a principled method of network composition from multi-layer data. We argue that inter-layer dynamics is a essential component of understanding the structure as a whole. Mathematically, we consider the following abstract problem: given multiple layers of network data over a shared vertex set, and additional parameters for inter-layer transitions, construct a (single) weighted network that best integrates the multi-layer dynamics. In this context, we will also study an empirical use case of the composition framework.

1 Introduction

Studies of network structures have lead to fundamental insights into the organization and function of social, biological and technological systems [13]. On top of these network structures, different dynamical processes unfold [4,8], leading to applications ranging from ranking web pages to maximizing social influence and controlling epidemics [9,14]. Traditionally, most research has focused on the simple graph representation where all verticies and edges are of a single type. More recently, there has been a great interest in network models that are capable of capturing multiple types of connections [1,15]. In this paper, we adopt the general notion of *multi-layer networks*, with *multiplex network* being a special case when inter-layer structures are absent [10].

Structure and dynamics of multi-layer networks have been explored in both theoretical graphs and real world data [2,5,12]. However, it remains an open question as how to build a multi-layer network in the first place. They are often constructed simply by stacking or projecting layers into a single network. When inter-layer edges are explicitly modeled, they are usually captured by a simple parameter called the *coupling strength*. One challenge for modeling inter-layer structures is that they are empirically difficult to measure in most cases [7].

In this paper, we propose a two-stage framework for multi-layer network composition based on a unified dynamical process. In Sect. 2, we will first briefly discuss how to transform the layers into homogeneous Markov processes, followed by the main theorem which infer inter-layer edge weights based on inter-layer dynamics. A real world example will be investigated in Sect. 3.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 315–321, 2017.
DOI: 10.1007/978-3-319-60240-0_38

2 The Multi-layer Composition Framework

We first introduce some basic notations.

Single-layer data: A standard network is represented by weighted directed graph $G = (V, E, \boldsymbol{A})$, where $V = \{1, ..., n\}$ and for $u, v \in V$, $a_{uv} \geq 0$ assigns a weight to edge $(u, v) \in E$. We follow the convention that $a_{uv} = 0$ if and only if $(u, v) \notin E$. G may have self-loops. For $u \in V$, let $d_u^{out} = \sum_{v=1}^{n} a_{u,v}$ denote the *out-degree* of vertex u. In this paper, we use $\boldsymbol{D_A}$ (or $\boldsymbol{D}$ when the context is clear) to denote the diagonal matrix whose entries are out-degrees.

Multi-layer data: We consider *vertex-aligned* multi-layer networks [10]. We use l to denote number of layers, and use $G^i = (V, E^i, \boldsymbol{A}^i)$ to denote the network at i^{th} layer. For clarity, we will use superscripts i, j, r for the layers and subscripts u, v, w for vertices. Note that the vertex set V is the same across the layers.

The simplest dynamical process on graphs G is the discrete time *unbiased random walk* (URW), represented by the transition matrix $\boldsymbol{M}$.

Lemma 1. *For every directed network $G = (V, E, \boldsymbol{A})$, there is a unique transition matrix, $\boldsymbol{M_A} = \boldsymbol{A}\boldsymbol{D_A}^{-1}$, that captures the URW Markov process on G. Conversely, given a transition matrix $\boldsymbol{M}$, there is in fact multi-layer an infinite family of adjacency matrices whose random walk Markov process is consistent with $\boldsymbol{M}$: $\boldsymbol{\mathcal{A}_M} = \{\boldsymbol{M\Gamma} : \boldsymbol{\Gamma}$ is a positive diagonal matrix.$\}$*

In other words, every directed network uniquely defines a random walk process. However, given a transition matrix $\boldsymbol{M}$, there remains n degrees of freedom to specify the underlying network. They are vertex scaling factors. For undirected graphs, there is only one global scaling factor.

In [8], we argued that perceived network structure is a result of the interplay between the network topology and the dynamical process on top of it. We believe this interplay is even more pronounced in multi-layer networks, with each layer represents a different type of connection. It is thus essential to account for the different intra-layer dynamics before we put them together.

For this purpose, we reintroduce the parametrized Laplacian [8], $\boldsymbol{\mathcal{L}} = (\boldsymbol{D'} - \boldsymbol{BA})(\boldsymbol{D'T})^{-1}$, where $\boldsymbol{A}$ is the adjacency and $\boldsymbol{D'}$ is the reweighed degree matrix now defined as: $d'_u = \sum_v [\boldsymbol{BA}]_{uv}$. The diagonal matrix $\boldsymbol{T}$ controls the time delay factors at each vertex. The bias factors form the other diagonal matrix $\boldsymbol{B}$. Under the framework, we can transform each input layer to equivalent graphs underlying continuous time URWs as the unifying dynamical process (Please refer to for the details and proofs).

Theorem 1. *For a directed network $G = (V, E, \boldsymbol{A})$, the dynamics $\boldsymbol{\mathcal{L}} = (\boldsymbol{D'} - \boldsymbol{BA})(\boldsymbol{D'T})^{-1}$ is equivalent to a continuous time URW with uniform delay factors on another transformed graph.*

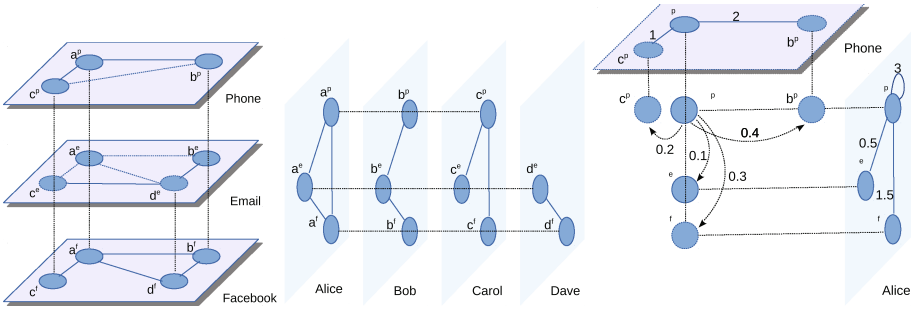
Corollary 1. *For an undirected network $G = (V, E, \boldsymbol{A})$, and the dynamical parameters $B, \boldsymbol{T}$, the interaction matrix $\boldsymbol{W} = \alpha(B\boldsymbol{A}B + (\boldsymbol{T} - \boldsymbol{I})\boldsymbol{D'})$ is unique up to a global scaling factor.*

We are now ready to discuss the second stage of the framework. For multiplex composition, simple matrix addition does the trick. However, we need a more general framework when inter-layer structures matter. Consider the following:

Formulation 1. *Given l transformed layers $G^1 = (V, E^1, \boldsymbol{W}^1), ..., G^l = (V, E^l, \boldsymbol{W}^l)$, and egocentric inter-layer dynamics $(\boldsymbol{M}_v : v \in V)$, compose a $(ln \times ln)$ weighted super-adjacency matrix, $\mathbb{W} = \begin{bmatrix} \boldsymbol{W}^1 & \boldsymbol{W}^{12} & ... & \boldsymbol{W}^{1l} \\ ... & & & \\ \boldsymbol{W}^{l1} & \boldsymbol{W}^{l2} & ... & \boldsymbol{W}^{l} \end{bmatrix}$ to integrate the multi-layer network data. In addition, $\mathbb{W}$ represent a diagonal multi-layer networks, as defined in [10], which means that all inter-layer edges are between the same vertex at different layers. Here, in $\mathbb{W}$, the l diagonal $(n \times n)$-blocks are directly fed from the first stage $\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^l$.*

We have used the model of egocentric inter-layer dynamics for each vertex $(\boldsymbol{M}_v : v \in V)$, with $\boldsymbol{M}_v$ being the stochastic transition matrix for the inter-layer instances of the same vertex v. Such egocentric models are considered to be fundamental in the formation of social structures [6].

Figure 1a, b is a toy example with three horizontal layers, consisting of (hypothetical) phone contacts, email exchanges and Facebook friendships. At the same time, egocentric inter-layer dynamics form a vertical perspective of the same system. Figure 1c represents the Markov transitions of this joint system when Alice receives a phone call. She might pass on the message directly by calling others with probability $0.6 = 0.4 + 0.2$, or relay the message through emails with probability 0.3, or post it on a Facebook wall with probability 0.1.



(a) Traditional view (b) The vertical perspective (c) Markov model around a^p

Fig. 1. A hypothetical toy example

By Lemma 1 each layer $\boldsymbol{W}^i$ uniquely defines a Markov model, $\boldsymbol{M}_{\boldsymbol{W}^i}$. Our task is to combine them with the n egocentric Markov models $(\boldsymbol{M}_v : v \in V)$. Therefore, we aim to identify a weighted $(ln \times ln)$-adjacency matrix $\mathbb{W}$, whose random-walk Markov model, $\boldsymbol{M}_{\mathbb{W}}$, satisfies the following two basic conditions:

1. **Layer Consistency**: The random-walk Markov model of each layer, $\boldsymbol{M}_{\boldsymbol{A}^i}$, $i \in [1, 2, .., l]$, is the *projection* of $\boldsymbol{M}_{\mathbb{W}}$ to that layer, and
2. **Ego Consistency**: The egocentric inter-layer dynamics, $\boldsymbol{M}_v$, of vertex $v \in V$, is the *layer marginals* of $\boldsymbol{M}_{\mathbb{W}}$ at vertex v.

The projection of $\boldsymbol{M}$ onto a subset is simply the stochastic normalization of corresponding principal submatrix of $\boldsymbol{M}$. Thus, Condition 1 is automatically achieved by setting diagonal blocks of $\mathbb{W}$ as $\boldsymbol{W}^1, ..., \boldsymbol{W}^l$ in Formulation 1.

For Condition 2, notice that $\mathbb{W}$ also defines an $l \times l$ interlayer adjacency matrix $\boldsymbol{W}_v$ at each $v \in V$. The corresponding random-walk process, $\boldsymbol{M}_{\boldsymbol{W}_v}$, is the projection of $\boldsymbol{M}_{\mathbb{W}}$ to these vertical slices. Let $q_{v,i}$ denote the transition probability of going from vertex v in the i^{th} layer to some u in the same layer, according to $\boldsymbol{M}_{\mathbb{W}}$. Let $\boldsymbol{Q}_v$ be the $l \times l$ diagonal matrix of $[q_{v,i} : i \in [l]]$. Then, $\boldsymbol{Q}_v + \boldsymbol{M}_{\boldsymbol{W}_v} \cdot (\boldsymbol{I} - \boldsymbol{Q}_v)$ denote the layer marginals of the joint Markov model $\boldsymbol{M}_{\mathbb{W}}$ at vertex v. Consequently, Ego Consistency requires that $\boldsymbol{M}_v = \boldsymbol{Q}_v + \boldsymbol{M}_{\boldsymbol{W}_v} \cdot (\boldsymbol{I} - \boldsymbol{Q}_v)$. Intuitively, $\boldsymbol{M}_v$ bridges between the orthogonal projections by including both $\boldsymbol{Q}_v$ and $\boldsymbol{M}_{\boldsymbol{W}_v}$. Now we present the main theorem of this paper:

Theorem 2. *For any multi-layer data* $(\boldsymbol{A}_i : i \in [l], \boldsymbol{M}_v : v \in V)$*, there exists a unique and feasible super-adjacency* $\mathbb{W}$ *that satisfies both Layer Consistency and Ego Consistency.*

Proof. Because Formulation 1 requires that all off-diagonal blocks of $\mathbb{W}$ are diagonal matrices, we have $(l^2 - l)n$ degrees of freedom after meeting Layer Consistency. Notice that $(\boldsymbol{M}_v : v \in V)$ are n stochastic $l \times l$ matrices. Thus, Ego Consistency represents $(l^2 - l)n$ dimensional constraints, which matches perfectly with the remaining degrees of freedom. Uniqueness proven.

To prove the feasibility of the unique solution, we introduce Algorithm 1,

Algorithm 1. Multilayer network composition

Input: weighted network layers: $G^1 = (V, E^1, \boldsymbol{A}^1), G^2 = (V, E^2, \boldsymbol{A}^2), ..., G^l = (V, E^l, \boldsymbol{A}^l)$, parameters of the dynamics: $\boldsymbol{T}^1, B^1, \boldsymbol{T}^2, B^2, ..., \boldsymbol{T}^l, B^l$, and n $l \times l$ egocentric inter-layer dynamics M_u for each vertex $u \in V$

Algorithm: For each layer i,

- Apply the bias transformation $\boldsymbol{A}'^i = B^i \boldsymbol{A}^i$ ($\boldsymbol{A}'^i = B^i \boldsymbol{A}^i B^i$ for undirected graphs)
- Apply the delay transformation with a global scaling $\boldsymbol{W}^i = \alpha^i(\boldsymbol{A}'^i + (\boldsymbol{T}^i - \boldsymbol{I})\boldsymbol{D}'^i)$
- Create a $ln \times ln$ empty matrix $\mathbb{W}$
- Fill the l diagonal blocks (each of size $n \times n$) with $\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^l$
- Construct the off diagonal blocks $\boldsymbol{W}^{ij}$ (each of size $n \times n$) for all layer pairs i and j based on Algorithm 2 with $\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^l$ as inputs

Output The super adjacency matrix $\mathbb{W}$

We need a subroutine Algorithm 2 to satisfy the Ego Consistency. Rearrange $\mathbb{W}$ so that the counterparts of the same vertex are grouped together.

$\bar{\mathbb{W}} = \begin{bmatrix} \boldsymbol{W}_1 & \boldsymbol{W}_{12} \ldots \boldsymbol{W}_{1n} \\ \ldots & \\ \boldsymbol{W}_{n1} & \boldsymbol{W}_{n2} \ldots \boldsymbol{W}_{n}, \end{bmatrix}$ where $\boldsymbol{W}_{u,v}$ are $l \times l$ matrices that have already been fixed by Layer Consistency. The diagonal blocks $\boldsymbol{W}_v : v \in V$ contains all entries set by Ego Consistency. The diagonal entries of $\boldsymbol{W}_v, v \in V$, are also set by Layer Consistency, because they are unaffected by the rearrangement of $\mathbb{W}$. The rest $n(l^2 - l)$ entries lead to the same degrees of freedom we discussed earlier.

The reordered $\boldsymbol{W}_u$ blocks are closely related to the egocentric adjacencies $\boldsymbol{X}_u$ underlying the egocentric inter-layer dynamics $\boldsymbol{M}_u$. The vertical slice in Fig. 1c demonstrates such a $\boldsymbol{X}_a$, where intra-layer transitions are captured using self-loops. Subroutine Algorithm 2 can now be specified as

Algorithm 2. Building inter-layer blocks

Input: transformed layers: $G_1 = (V, E^1, \boldsymbol{W}^1), G_2 = (V, E^2, \boldsymbol{W}^2), ..., G_l = (V, E^l, \boldsymbol{W}^l)$, and a $l \times l$ egocentric inter-layer transition matrix $\boldsymbol{M}_u$ for vertex $u \in V$.
Algorithm:

- Create a $l \times l$ empty matrix $\boldsymbol{X}_u$
- Fill the diagonal elements with $\boldsymbol{X}_u^{ii} = d_u^i(out)$
- Construct the off diagonal elements $\boldsymbol{X}_u^{ij} = \frac{\boldsymbol{M}_u^{ij}}{\boldsymbol{M}_u^{ii}} d_u^i(out)$

Output Block $\boldsymbol{X}_u$ and repeat for each $u \in V$

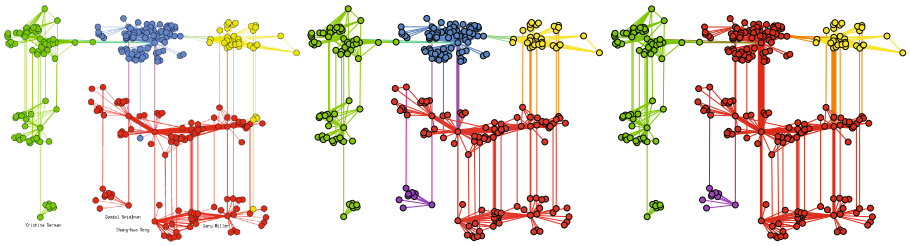
Using Lemma 1, we can rewrite the steps in Algorithm 2 as $\boldsymbol{X}_u = \boldsymbol{M}_u \boldsymbol{\Gamma}$, by setting the i^{th} entry of $\boldsymbol{\Gamma}$ uniquely as $d_u^i(out)/\boldsymbol{M}_u^{ii}$, where $d_u^i(out)$ is the total out edge weights of vertex u in layer i. Intuitively, we are simply using the intra-layer dynamics to determine the vertex scaling factor.

From Fig. 1c, it is clear that the off-diagonal parts of $\boldsymbol{X}_u$ is exactly what we are looking for in $\boldsymbol{W}_u$ blocks. Or $\boldsymbol{W}_u = \boldsymbol{X}_u - D_u(out)$, where the diagonal matrix $D_u(out)$ is composed of $d_u^i(out)$ entries. With the uniquely solvable $\boldsymbol{X}_u$ blocks, we can now complete the output $\mathbb{W}$ by filling its off diagonal blocks $\boldsymbol{W}^{ij}$ with reordered $\boldsymbol{W}_u$ blocks. On top of that, Algorithm 2 will always lead to feasible solutions with the constrains $\boldsymbol{X}_u^{ij} \geq 0$, provided that $\boldsymbol{M}_u$ entries are well defined. Uniqueness and feasibility proven.

3 A Real World Example

Figure 2 presents collaboration networks (undirected) centered around four authors: Shang-hua Teng, Daniel Spielman, Gary Miller and Kristina Lerman, as well as their coauthors on papers appearing in the ACM Digital Library. Each layer represents a separate time period: from bottom to top, 1985–1994, 1995–2004, and 2005–2014. The weight of an intra-layer edge represents the number of times two authors collaborated during that decade. The traditional approach, as

shown by Fig. 2(a), simply connect the same author between neighboring decades with a constant coupling strength of 2. To use our framework, we assume that an author has a 10%/20% transition probability to connect with former self in the previous decade. Specifying inter-layer edges using Algorithms 1 and 2, first between the top tow layers then the bottom two, we have Fig. 2(b), (c).



(a) Constant couplings (b) 10% transition (c) 20% transition

Fig. 2. Community structures of coauthor networks using different compositions (Color figure online)

For comparison, we have visualized the community structures with different multi-layer compositions in Fig. 2, using the Louvain algorithm [3], with the resolution parameter set to 5 [11]. The traditional approach produced some counter-intuitive cross-layer communities, because the constant coupling strength is too strong for peripheral vertices. Our framework, on the other hand, leads to much more sensible results with different inter-layer strength for vertices with different degrees. With a 10% transition probability, we can see that authors surrounding Teng and Spielman later separated from the red community (which became yellow) of theoretical computer scientists like Miller. As Teng started to collaborate with the social network community surrounding Lerman (green), the newly formed blue community now represents the filed of graph theory.

References

1. Acar, E., Yener, B.: Unsupervised multiway data analysis: a literature survey. IEEE Trans. Knowl. Data Eng. **21**(1), 6–20 (2009)
2. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., Vespignani, A.: Multiscale mobility networks and the spatial spreading of infectious diseases. Proc. Natl. Acad. Sci. **106**(51), 21484–21489 (2009)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theor. Exp. **10**, 10008 (2008)
4. Borgatti, S.: Centrality and network flow. Soc. Netw. **27**(1), 55–71 (2005)
5. De Domenico, M., Sole, A., Gomez, S., Arenas, A.: Random walks on multiplex networks. ArXiv e-prints, June 2013
6. Dunning, D., Cohen, G.L.: Egocentric definitions of traits and abilities in social judgment. J. Personal. Soc. Psychol. **63**(3), 341–355 (1992)

7. Gallotti, R., Barthelemy, M.: The multilayer temporal network of public transport in Great Britain. Sci. Data **2**, 140056 (2015)
8. Ghosh, R., Teng, S.H., Lerman, K., Yan, X.: The interplay between dynamics and networks: centrality, communities, and cheeger inequality. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1406–1415. KDD 2014, ACM, New York (2014)
9. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: KDD 2003, pp. 137–146. ACM (2003)
10. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J.P., Moreno, Y., Porter, M.A.: Multilayer networks. ArXiv e-prints, September 2013
11. Lambiotte, R., Delvenne, J.C., Barahona, M.: Laplacian dynamics and multiscale modular structure in networks. arXiv preprint arXiv:0812.1770 (2008)
12. Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**, 876 (2010)
13. Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)
14. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web (1999)
15. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications, vol. 8. Cambridge University Press, Cambridge (1994)

Extracting Information from Negative Interactions in Multiplex Networks Using Mutual Information

Alireza Hajibagheri[1], Gita Sukthankar[1(✉)], and Kiran Lakkaraju[2]

[1] University of Central Florida, Orlando, FL, USA
{alireza,gitars}@eecs.ucf.edu
[2] Sandia National Labs, Albuquerque, NM, USA
klakkara@sandia.gov

Abstract. Many interesting real-world systems are represented as complex networks with multiple types of interactions and complicated dependency structures between layers. These interactions can be encoded as having a valence with positive links marking interactions such as trust and friendship and negative links denoting distrust or hostility. Extracting information from these negative interactions is challenging since standard topological metrics are often poor predictors of negative link formation, particularly across network layers. In this paper, we introduce a method based on mutual information which enables us to predict both negative and positive relationships. Our experiments show that SMLP (Signed Multiplex Link Prediction) can leverage negative relationship layers in multiplex networks to improve link prediction performance.

Keywords: Multiplex link prediction · Complex networks · Mutual information

1 Introduction

While both positive and negative relationships clearly exist in many social network settings, the vast majority of research has only considered positive relationships. On social media platforms, people form links to indicate friendship, trust, or approval, but they also link to signify disapproval of opinions or products. In this paper, we address the problem of predicting future user interactions from other layers of the network where a layer could represent either negative or positive user interactions. To do this, it is crucial to determine the correlation between layers. Two network layers of opposite valence are likely to have negatively correlated link formation processes. To capture the interdependencies between different layers, we use mutual information to determine the sign of the correlation. Mutual information expresses the reduction in uncertainty due to another variable; we demonstrate that the average value of a layer's mutual information can be used to calculate the correlation of link formation processes across network layers with unknown valences.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 322–328, 2017.
DOI: 10.1007/978-3-319-60240-0_39

2 Proposed Method

In previous work, we introduced MLP [1], a hybrid architecture that utilizes multiple components to address different aspects of the link prediction task. While MLP utilizes information from all layers of a network to improve link prediction in a target layer, it is not able to capture negative correlations among different layers of a network. For example, let us consider α and β as layers of a network N where α and β represent positive (trades) and negative (raids) relationships between users in a game respectively. Now, if our goal is to predict future links of layer α using information from β, MLP fails to capture the negative effect of a link between two nodes on β. In order to model such relationships, we propose a model based on *mutual information* that can capture the correlation sign between different layers in order to modify the MLP weighting procedure. In this section, we first provide a short description on the concept of mutual information and describe how it has been previously used for the link prediction task in single layer networks. Finally, we introduce our new signed multiplex link prediction model.

2.1 Using Mutual Information for Link Prediction

Considering a random variable X associated with outcome x_k with probability $p(x_k)$, its self-information $I(x_k)$ can be denoted as $I(x_k) = \log \frac{1}{p(x_k)} = -\log p(x_k)$ [2]. The higher the self-information is, the less likely the outcome x_k occurs. On the other hand, the mutual information of two random variables can be denoted as:

$$I(x_k; y_j) = \log \frac{p(x|y)}{p(x)} = -\log p(x_k) - (-\log p(x_k|y_j)) = I(x_k) - I(x_k|y_j) \quad (1)$$

The mutual information is the reduction in uncertainty due to another variable. Thus, it is a measure of the dependence between two variables. It is equal to zero if and only if two variables are independent. Tan et al. [3] proposed the following link prediction model based on mutual information. Let $\Gamma(x)$ represent node x's neighbors, then for the node pair (x, y), the set of their common neighbors is denoted as $O_{xy} = \Gamma(x) \cap \Gamma(y)$. Given a disconnected node pair (x, y), if the set of their common neighbors O_{xy} is available, the likelihood score of node pair (x, y) is defined as:

$$s_{xy}^{MI} = -I(L_{xy}^1|O_{xy}) \quad (2)$$

where $I(L_{xy}^1|O_{xy})$ is the conditional self-information of the existence of a link between node pair (x, y) when their common neighbors are known (refer to [3] for more details).

2.2 Signed Multiplex Link Prediction

Various experiments on multilayer link prediction have indicated that using neighborhood information from different layers of a network in a multiplex environment can improve the performance of link prediction. Hristova et al. [4] proposed the concept of a multilayer neighborhood where a link that exists on more than one layer in a multiplex network is called a *multiplex link*. Following the definition of a multilayer network, the ego network of a node can be redefined as the multilayer neighborhood. While the simple node neighborhood is the collection of nodes one hop away from it, the multilayer global neighborhood (denoted by GN) of a node i can be derived by the total number of unique neighbors across layers:

$$\Gamma_{GNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cup \beta}\} \tag{3}$$

Similarly, the core neighborhood (denoted by CN) of a node i across layers of the multilayer network is defined as:

$$\Gamma_{CNi} = \{j \in V^{\mathcal{M}} : e_{i,j} \in E^{\alpha \cap \beta}\} \tag{4}$$

The goal is to determine the type of correlation (negative or positive) between two layers of a multiplex network. To this end, we calculate the average mutual information value for a target layer based on predictor layers. There are two assumptions: (1) Using the core neighborhood, if there is a negative correlation between two layers, the value of average mutual information would decrease substantially but would not change significantly if there is a positive correlation between the two layers. (2) If the global neighborhood is used to calculate the value of mutual information for a node pair, this would increase the value of average mutual information if there is a positive correlation between two layers and would not change significantly otherwise. After the sign of the relationship between the target layer and other predictor layers has been determined, MLP is used for predicting future links with the minor addition that layer weights can have both negative and positive values. Given a disconnected node pair (x, y), the mutual information of link existence (assuming that the set of their common neighbors O_{xy} is available) can be derived as:

$$I(L^1_{xy}; O_{xy}) = I(L^1_{xy}) - I(L^1_{xy}|O_{xy}) \tag{5}$$

For a multiplex network, O_{xy} can be redefined to use the global and core neighborhood of the two nodes. As a result, the average mutual information of a target layer would be defined as:

$$MI^{\alpha} = \frac{1}{|E^{\alpha}|} \sum_{x,y \in E^{\alpha} \& x \neq y} I(L^1_{xy}; O_{xy}) \tag{6}$$

The value of MI^{α} is calculated using information from a predictor layer β. It can be used to indicate the sign (negative or positive) of the correlation between the link formation in two layers. This sign is then used within the second phase

of link prediction task (MLP) to assign weights to all predictor layers and hence improve the performance of link prediction task for the target layer α.

3 Experimental Study

This paper evaluates the SMLP framework on networks extracted from two real-world datasets, Travian and Cannes2013. Not only do we compare our results with two other approaches for fusing cross-layer information, but we also consider scores generated by mutual information paired with core and global neighborhood as distinct methods.

3.1 Datasets

We use two real-world dynamic multiplex networks to demonstrate the performance of our proposed algorithm. These networks are considerably disparate in structure and were selected from different domains. Negative links (raids) exist between users of our MMOG dataset to evaluate the performance of our method in predicting negative links as well as examining these layers on positive ones.

- **Travian MMOG** [5] Travian is a browser-based, real-time strategy game. Trades, messages, and raids networks were extracted from Travian for this research.
- **Twitter Interactions** [6] This dataset consists of Twitter activity before, during, and after an "exceptional" event as characterized by the volume of communications. The Cannes2013 dataset was created from tweets about the Cannes film festival that occurred between May 6, 2013 to June 3, 2013.

3.2 Evaluation Metrics

For the evaluation, we measure receiver operating characteristic (ROC) curves for the different approaches. The ROC curve is a plot of the *true positive rate (tpr)* against the *false positive rate (fpr)*. We report area under the ROC curve (AUROC), the scalar measure of the performance over all thresholds.

3.3 Analysis of Multilayer Neighborhood

As mentioned before, our assumption is that both core and global neighborhood definitions would enable us to study correlations between different layers of a network. Figure 1 shows average mutual information values for Travian network for each day of the 30 day period. There are negative and positive correlations between trades (or messages)-raids and trades-messages respectively. As shown in the figure, at every timestep of the network, the average value decreases when the core neighborhood is used in the case of negative correlation (Fig. 1(a) and (b)) and does not change significantly when global neighborhood is used to calculate the average value. On the other hand, as shown in Fig. 1(c), when dealing

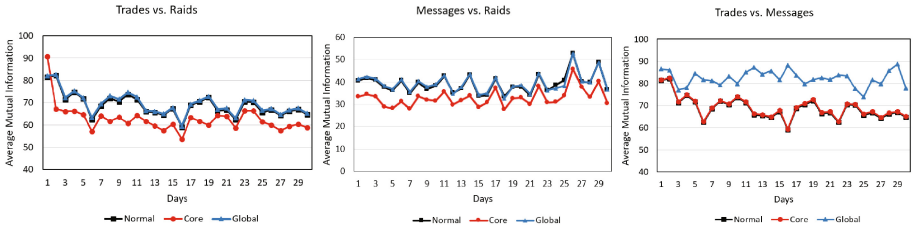


Fig. 1. Average mutual information values over time calculated using normal and core neighborhood definitions.

with positive correlations (trades-messages), the value of average mutual information does not change significantly using the core neighborhood but increases drastically using the global neighborhood definition (the average difference is less than one). Hence, this justifies our assumption that the average mutual information value of a certain layer can be used to determine the sign of its role in target layer link prediction.

3.4 Performance of Signed Multiplex Link Prediction (SMLP)

Table 1 shows the results of different algorithms on the Travian and Cannes2013 datasets. With 30 days of data from Travian and 27 days for Cannes2013, we were able to extensively compare the performance of the proposed methods and the impact of using different elements. AUROC performances for a target layer averaged over all snapshots are calculated, and our proposed framework is shown at the top of the table, followed by variants of mutual information (MI) based link prediction models using different definitions of neighborhood (N which stands for Normal proposed by Tan et al. [3], CN stands for Core Neighborhood and GN stands for Global Neighborhood). The algorithms shown in the bottom half of the table (Average Aggregation (AA) and Entropy Aggregation (EA) [7]) are techniques for multiplex networks proposed by other research groups. The settings given in [1] were used for MLP.

Table 1. AUROC performances for a target layer averaged over all snapshots and ten runs.

	Trade	Message	Raids	Retweet	Mention	Reply
SMLP	$\mathbf{0.871 \pm 0.013}$	$\mathbf{0.843 \pm 0.031}$	$\mathbf{0.793 \pm 0.007}$	$\mathbf{0.812 \pm 0.002}$	$\mathbf{0.834 \pm 0.003}$	$\mathbf{0.839 \pm 0.002}$
MLP	0.821 ± 0.001	0.803 ± 0.002	0.758 ± 0.001	0.812 ± 0.002	0.834 ± 0.003	0.839 ± 0.002
MI (CN)	0.740 ± 0.016	0.753 ± 0.011	0.744 ± 0.013	0.774 ± 0.009	0.759 ± 0.012	0.782 ± 0.005
MI (GN)	0.737 ± 0.010	0.746 ± 0.011	0.747 ± 0.009	0.771 ± 0.011	0.767 ± 0.012	0.773 ± 0.009
MI (N)	0.716 ± 0.012	0.727 ± 0.006	0.703 ± 0.012	0.731 ± 0.007	0.725 ± 0.013	0.742 ± 0.014
AA	0.744 ± 0.030	0.752 ± 0.020	0.658 ± 0.017	0.740 ± 0.003	0.737 ± 0.011	0.761 ± 0.003
EA	0.731 ± 0.004	0.763 ± 0.020	0.661 ± 0.009	0.749 ± 0.003	0.758 ± 0.031	0.744 ± 0.002

Bold numbers indicate the best results on each target layer considered. As expected, SMLP is the best performing algorithm in all cases since not only it utilizes both historical and cross-layer information, but also mutual information enables SMLP to capture negative correlations between different layers. As a result, node pairs that are connected by raids in Travian, are penalized for this connection and eventually receive a lower score compared to node pairs that are only connected on messages and trades. This holds true when raids is the target layer and two nodes are connected on either messages or trades layers which are negatively correlated with raids. On the other hand, it is evident that methods specifically designed for multiplex link prediction outperform Mutual Information (N) which is unable to leverage cross-layer information. Also, Average Aggregation and Entropy Aggregation are able to achieve higher AUROC scores compared with Mutual Information based methods since they collect more information from the network using different similarity metrics such as common neighbors, Adamic/Adar, etc. Finally, for Twitter layers, SMLP and MLP achieve similar results since there are no negative layers to modify the sign of the weights associated with different layers of the network.

4 Conclusion and Future Work

In this paper, we introduce a new link prediction framework, SMLP (Signed Multiplex Link Prediction), that employs a holistic approach to accurately predict links in dynamic multiplex networks by incorporating negative relationships between users. Our analysis on real-world networks created by a variety of social processes suggests that SMLP effectively models multiplex network coevolution in many domains and is also able to capture negative correlation between layers in order to improve link prediction performance.

In future work, we are planning to extend our results to other multilayer networks that contain negative interactions; we are particularly interested in studying datasets that contain multiple negative interaction layers to see if they result in positive correlations. Another promising line of inquiry is studying the usage of multiple features (beyond common neighbors) to calculate mutual information, since different structural features highlight different aspects of network formation.

Acknowledgments. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. The Travian dataset was provided by Drs. Rolf T. Wigand and Nitin Agarwal (University of Arkansas at Little Rock, Department of Information Science).

References

1. Hajibagheri, A., Sukthankar, G., Lakkaraju, K.: A holistic approach for link prediction in multiplex networks. In: Spiro, E., Ahn, Y.-Y. (eds.) SocInfo 2016. LNCS, vol. 10047, pp. 55–70. Springer, Cham (2016). doi:10.1007/978-3-319-47874-6_5
2. Shannon, C.E.: A mathematical theory of communication. ACM SIGMOBILE Mobile Comput. Commun. Rev. **5**(1), 3–55 (2001)
3. Tan, F., Xia, Y., Zhu, B.: Link prediction in complex networks: a mutual information perspective. PloS One **9**(9), e107056 (2014)
4. Hristova, D., Noulas, A., Brown, C., Musolesi, M., Mascolo, C.: A multilayer approach to multiplexity and link prediction in online geo-social networks. arXiv preprint (2015). arXiv:1508.07876
5. Hajibagheri, A., Lakkaraju, K., Sukthankar, G., Wigand, R.T., Agarwal, N.: Conflict and communication in massively-multiplayer online games. In: Agarwal, N., Xu, K., Osgood, N. (eds.) SBP 2015. LNCS, vol. 9021, pp. 65–74. Springer, Cham (2015). doi:10.1007/978-3-319-16268-3_7
6. Omodei, E., De Domenico, M., Arenas, A.: Characterizing interactions in online social networks during exceptional events. arXiv preprint (2015). arXiv:1506.09115
7. Pujari, M., Kanawati, R.: Link prediction in multiplex networks. Netw. Heterogeneous Media **10**(1), 17–35 (2015)

A Blockchain-Enabled Participatory Decision Support Framework

Marek Laskowski[1,2,3(✉)]

[1] Mathematics and Statistics, York University, Toronto, Canada
marek.laskowski@gmail.com
[2] Population Medicine, University of Guelph, Guelph, Canada
[3] Information and Computing Technologies, Seneca College, Toronto, Canada

Abstract. In this "post truth" age of "fake news" and "alternative facts" uncertainty in the provenance of policy makes transparency in decision making increasingly important for evidence-based policymakers. This paper demonstrates how the convergence of Agent Based Modelling, Smart Contracts, Blockchain, and Virtual Reality (VR) technologies make possible the development of participatory decision support frameworks. The concept is demonstrated in a public health context by implementing an Agent Based Model of disease spread within a simulated population (SIR model) as a so-called Smart Contract deployed on a public Blockchain network. We demonstrate that in order to "close the loop" the simulation outcomes can be visualized using commodity VR hardware, and future extensions towards fully interactive simulations are proposed.

Keywords: Blockchain · Participatory decision support · Agent based model · Data analysis · Ethereum · Smart contract · Virtual reality

1 Introduction

When decisions are made behind closed doors by unknown officials using unseen evidence, there can be a breakdown in communication leading to lowered uptake or acceptance of public policy interventions [1]. As a result, in many developed and developing countries individuals disregard policy (e.g. public health) messages that they do not understand or trust. Furthermore, in recent years there have been increasing calls for transparency and openness in all aspects of public policy decision making processes [2] including the ethical management of individual's data [3].

For applications in public health in particular, the effectiveness of evidence-based decision making relies on not only the compliance of the public, but the availability of highly detailed contact network data and epidemiological surveillance [4] requiring increased trust and buy-in from all stakeholders. We argue that for there to be increased trust between public policy makers and the public, all parts of the decision support phases described by Fig. 1 must take into account: fair governance of data and resources, availability of source data and analytics code for all stakeholders, as well as provenance of data and resulting analytic evidence.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 329–334, 2017.
DOI: 10.1007/978-3-319-60240-0_40

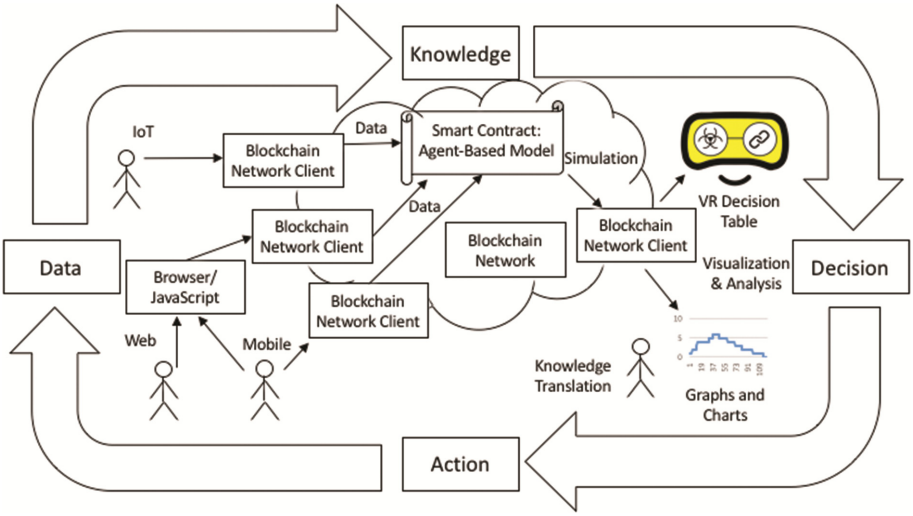


Fig. 1. An evidence-based decision support process alongside proposed framework components. Data from a variety of sources including web, mobile, and IoT is collected and stored on a Blockchain network. The data is combined in a Smart Contract with expert knowledge to create an Agent Based Model which can be used to carry out simulations on the Blockchain network. During the decision phase, simulation results can be analyzed using traditional plots or by using a Virtual Reality "decision support table" to assist with knowledge translation leading into the action phase.

The key emerging technologies that enable participatory decision support include:

(1) Blockchain - A Blockchain can be said to implement a highly-secure, distributed ledger or database [5–7]. Unlike a centralized database controlled by one organization, instead a peer-to-peer network shares responsibility for maintaining the state or integrity of the data contained within. Data stored within a Blockchain is considered immutable, although, in effect, newer records can be created to serve as updates for older records. Value represented as tokens can be transferred using a Blockchain, as demonstrated by networks such as Bitcoin [5] and Ethereum [8].

(2) Smart Contracts - The term "Smart Contract" has come to mean a computer program which has its execution carried out on a Blockchain, and furthermore its state and outputs are secured by said Blockchain. A "Smart Contract" can encode complex business logic and allows for highly-structured computations to be carried out without direct human intervention [8, 9].

(3) Agent Based Models (ABMs), incorporate rules from expert stakeholders, as well as a numerous data streams to parameterize the numerical and statistical parameters that underlie stochastic phenomena in the model [10]. In the resulting model that is both heuristic and data driven in nature, complex behavior emerges from the relatively simple set of rules relating a population of agents to the environment and to each other. Agent Based Modelling is increasingly used for modelling complex social phenomena [10] including the spread of infectious disease [4].

(4) Virtual Reality (VR), is a technology that presents realistic 3D highly immersive simulations to users, in which the users can move their bodies to interact with the simulation to varying degrees [11]. In contemporary designs, headgear is used to mount increasingly capable consumer devices (e.g. smartphones) directly in front of the user's eyes showing a different stereoscopic image to each eye.

The remainder of this paper will discuss how each key technology is used in the context of a prototype participatory decision support framework, and evaluate the prototype against the aforementioned principles.

2 Prototype Implementation

An evidence-based decision making process encompassing "from data to knowledge to decision to action" along with the corresponding proposed framework components is shown in Fig. 1. In the data collection phase, input data can come from individuals using a Web Browser interface as well as mobile devices including smartphones and Internet of Things (IoT) devices. A Blockchain network client (i.e. geth) is used directly or through an intermediate Javascript/Web layer to deliver data onto the Blockchain.

In the context of this proof-of-concept, anonymous data is aggregated and stored using a relatively simple mechanism within a Smart Contract on the Blockchain network, implemented using the Solidity programming language. The data is combined with expert knowledge encoded within the Smart Contract to implement the Agent Based Model used to model the spread of infectious disease. The disease model here is based on a previously validated model [4]. After simulations are carried out using the computational capacity of the Blockchain network, the relevant simulation logs are extracted using the Blockchain network client. In the "decision" phase, simulation results can be analyzed using traditional aggregate graphs, charts, and statistics, such as Fig. 2. Results can also be visualized using a Virtual Reality "Decision Support Table" implemented using the Google Daydream VR kit[1] and Unity3D[2]. VR represents an intuitive and explicit means of visualizing data and simulation outcomes during the decision phase, and can be used for knowledge translation to convey the nature of simulation based modelling to stakeholders and the public during knowledge dissemination in the action phase.

Within a typical Agent Based Modelling workflow, model refinements are often made by working with the model code, redeploying the code, and re-running simulations. Therefore, at this stage of development some phases of the decision making process understandably involve human intervention, such as deploying or upgrading the Smart Contract, data entry, as well as importing the simulation log into the VR application.

[1] https://vr.google.com/daydream/.
[2] https://unity3d.com/.

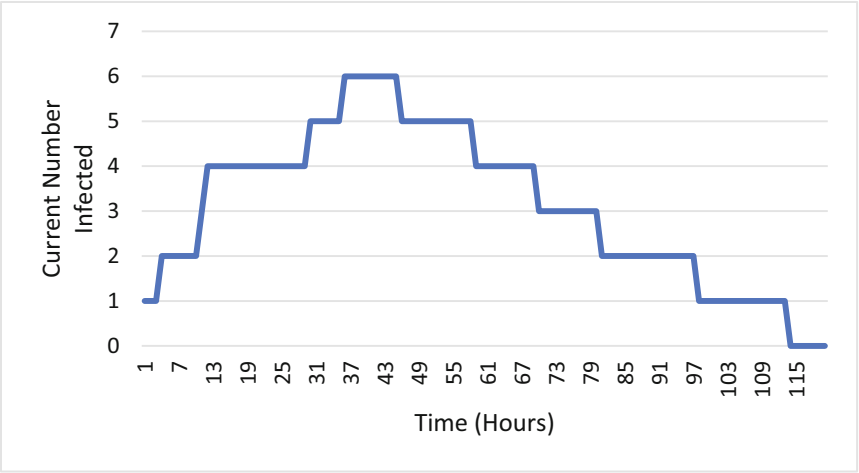


Fig. 2. Number of currently infected individuals on each hour of the simulation.

For further details, please see the project's Github page: https://github.com/professormarek/Intellichain or [9] for more background on technical implementation. For brevity, screenshots, downloads, and other details concerning the VR decision table demo application are available at: http://www.intellicha.in.

3 Results

A scenario including seven locations, and ten agents, one initially infected was used to demonstrate and test the functionality of the prototype framework; the results of this test run are shown in Fig. 2. It was decided to follow the Ethereum community's best practices and carry out tests of the smart contract with the use of a test network. Once preliminary tests were completed, a contract was deployed to the Main Ethereum network at address 0x203028e846f512ef3320c10f0d39739906e65797 - in order to demonstrate the open and participatory features of the framework, anyone can interact with the contract directly using an Ethereum client or using the Application Binary Interface (ABI) published on the project's aforementioned GitHub page.

The simulation results can be interpreted using standard analytical tools such as graphs and charts. One such graph is Fig. 2 which shows the number of currently infected individuals as a function of time during a simulation with particular initial conditions carried out on the test network. The published VR app demo permits users to experience the simulation carried out earlier on an Ethereum test network, presented as a virtual "decision table". Both Fig. 2 and the simulation that plays out in the VR app communicate the important epidemiological principle of a super-spreader infecting several secondary cases early on in the outbreak who in turn infect others later on.

As the purpose of this article is to expound the concept and to demonstrate its viability, a more rigorous study of performance characteristics and analysis of scalability will be the subject of a future publication.

4 Discussion

Regarding the fair use of data and resources, the current iteration of the prototype is focused on the use of open data, as well as anonymized data volunteered by informed participants. However, in general, fairness may imply compensation of participants for the use of their data. Fortunately, the Ethereum Blockchain network has secure value token exchange as a first-class feature of the protocol, making it very easy to compensate participants. Those individuals running nodes that perform calculations and maintain the ledger (i.e. miners) are already rewarded using this concept.

Regarding the participatory and open nature of the framework towards all stakeholders, we argue that this framework is successful in this regard, because all stakeholders including the public are able to access to the same data, results, and tools as experts.

Regarding the clear governance and provenance of data, beyond the aforementioned compensation of participants, the governance of data through the use of smart contracts on a Blockchain, has already been demonstrated elsewhere [12]. Metadata, if available, concerning the provenance of data can be entered into the Blockchain alongside said data. When computation and analytics are carried out using an open Blockchain network, the simulations are verifiable and the provenance of all results are clear owing to the immutability and auditability of the Blockchain [9].

4.1 Future Work

Scientific validation of the model presented here is a priority, however it is a time-consuming process that involves expert stakeholder involvement. Another pressing goal, enabled by advances in mobile Ethereum clients is to make the VR application fully interactive with "live" simulations being executed on the Blockchain from within the VR app. The VR app is also being upgraded to allow multiple users to interact with the simulation and each other in the same shared simulated experience. Now that an initial proof-of-concept has been presented, streamlining of the implemented phases can take place, and increasing the overall automation in the process. As mentioned, extensive performance analysis of the simulation framework will be carried out to better understand scalability and performance limitations.

5 Concluding Remarks

In summary, the framework presented here represents a step forward in participatory decision support, owing to its open design principles as well as advanced analysis and visualization capabilities. The code, data and outcomes are open and the data and simulations are governed by smart contracts on a Blockchain. Experts and the public they serve have access to the same platform where ideas and outcomes can flow freely.

Evidence-based decision support efforts relying on simulation should take into account potentially valuable tradeoffs in being able to substantiate the provenance and

quality of the decision support despite any additional overhead incurred. In the near future this may be part of a cost-benefit analysis when deciding on a modelling strategy.

The utility of public Blockchain networks for creating an open, participatory, decision support framework is demonstrated in an epidemiological, public health, context – however, extensions to other domains remain possible. Perhaps one day soon, individuals will run participatory decision support software at home much as they do seti@home or folding@home in the present day.

References

1. Makri, A.: Give the public the tools to trust scientists. Nature **541**(7637), 261 (2017). doi:10.1038/541261a
2. Helbing, D., Bishop, S., Conte, R., Lukowicz, P., McCarthy, J.B.: FuturICT: participatory computing to understand and manage our complex world in a more sustainable and resilient way. Eur. Phys. J. Spec. Top. **214**(1), 11–39 (2012)
3. Pentland, A.: Personal data: the emergence of a new asset class. Report, World Economic Forum (2011)
4. Najafi, M., Laskowski, M., de Boer, P.T., Williams, E., Chit, A., Moghadas, S.M.: The Effect of Individual Movements and Interventions on the Spread of Influenza in Long-term Care Facilities. Medical Decision Making (2017, in press)
5. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2008). http://www.bitcoin.org/bitcoin.pdf
6. Swan, M.: Blockchain: Blueprint for a New Economy. O'Reilly Media Inc, Blockchain (2015)
7. Tapscott, D., Tapscott, A.: Blockchain Revolution: How the Technology Behind Bitcoin is Changing Money, Business, and the World. Penguin, New York (2016)
8. Buterin V.: Ethereum white paper (2013)
9. Kim, H.M., Laskowski, M.: Towards an ontology-driven Blockchain design for supply chain provenance. In: 26th Workshop on Information Technologies and Systems (2016)
10. Bonabeau, E.: Agent-based modeling: methods and techniques for simulating human systems. Proc. Natl. Acad. Sci. **99**, 7280–7287 (2002)
11. Steuer, J.: Defining virtual reality: dimensions determining telepresence. J. Commun. **42**(4), 73–93 (1992)
12. Azaria, A., Ekblaw, A., Vieira, T., Lippman, A.: MedRec: using blockchain for medical data access and permission management. In: International Conference on Open and Big Data (OBD), pp. 25–30. IEEE, 22 August 2016

Hyperparameter Optimization for Predicting the Tolerance Level of Religious Discourse

Donald E. Brown[1(✉)], Hope McIntyre[1], Peter J. Grazaitis[2], Riannon M. Hazell[2], and Nicholas Venuti[1]

[1] Data Science Institute, University of Virginia, Charlottesville, VA, USA
brown@virginia.edu
[2] Human Research and Engineering Directorate, U.S. Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA
https://dsi.virginia.edu/

Abstract. To address the rising tide of religious violence as it affects U.S. Military deployments, the Army requires analytic methods that can be generalized to predict religious group violence around the globe and scalable to the number of potential groups in an area of operation. Current computational methods based on semantics and topics are lacking in predictive performance for the generalized problem and topic modeling performs poorly in predicting the tolerance level of new groups towards U.S. Military presence. The research in this paper aims to discover the association between religious speech and behaviors and provide a foundation for proactive engagement with these groups. The approach builds on the work from ethnolinquistics to model how things are said (performative analysis) rather than word meanings (semantic analysis). Recent research has developed computational approaches to streamline the manually intensive performative analysis of religious text. While producing promising results, these computational methods lack systematic optimization of the hyperparamters in the learning algorithms. Hence, we do not know the sensitivity of the results to parameter settings. This paper reports on results for predicting religious tolerance by optimizing the parameters in the signal processing algorithms and shows that the predictive power of performative approach is robust to parameter settings.

Keywords: Behavior analysis · Military · Computational linguistics

1 Introduction

The U.S. Army's role continues to expand into operations that are "characterized by ambiguity in the nature of the conflict, the parties involved, or the relevant policy and legal frameworks" [15]. These operations frequently require interactions with the indigenous populations with various communities, groups, state and non-state actors for effective mission execution and success. It is well recognized by the U.S. Military that ideological based groups including religious

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 335–341, 2017.
DOI: 10.1007/978-3-319-60240-0_41

groups pose threats to mission success [14]. This paper describes research that focuses on the use of the language of religious non-state actors to predict their behaviors.

Many approaches to predicting religious attitudes use the literal or emotive attributes of religious speech as encoded in responses to questions and scenarios [1]. In contrast, some scholars of religions have argued for an approach to understanding religious discourse using the capacity for religious language to encapsulate an action or identity and the flexibility of this language for use in inter-group communications [12]. This approach is *performative* analysis.

However, application of these ideas to actual dialog and group interactions is difficult because of the time and effort required to manually process religious text from a variety of groups. To develop the classification for the documents used in this paper, we have had teams of students and faculty members review documents. This process takes approximately one hour for a 500–750 word document.

Recent work has applied machine learning with signal processing to help automate performative analysis [16]. While showing promising results in labeling the tolerance levels of religious groups, the signal processing used in [16] was not optimized. Hence, we have no idea how sensitive those results were to parameters in the models. This paper provides that sensitivity analysis and gives results for computational performative analysis of religious documents with optimized hyperparameters.

2 Related Work

Linguistic approaches to religious text have focused primarily on semantics to classify the goals and intentions of groups (e.g., see [3]. The resulting correlations with behaviors are often difficult to understand and interpret. For example, Hassner, et al., used Latent Semantic Analysis (LSA) to track temporal shifts in language usage for Iranian leaders. While topical shifts were identified in this approach, few predictive capabilities were developed through this methodology [9]. When researchers have employed quantitative methods for semantic analysis they have narrowly focused on specific incidences of violence that lack generality [17].

In contrast to semantic analysis, performative analysis considers word usage as signaling intentions. Barsalou [2] showed that shifts in definitions correlate with the representational flexibility of a concept. Additionally, Sagi, et al., [13] found that a diversity of contexts directly correlates with the performative characteristics of words.

Results by [5] showed that neither word stop-lists nor word stemming significantly improved performance on word-word co-occurrence statistics. Minimal context window size gave the best performance, while dimensionality reduction through singular value decomposition (SVD) also improved performance. Boussidan and Ploux [4] used a graph built from subsetted co-occurrence tables to create a map of lexical usages of words. While they got good results, the complexity of computing cliques limits the scalability of their method. [6] used a similar approach to detect 'amelioration' (a word losing a negative meaning) and 'pejoration.'

3 Data

The data for the analysis comes from online repositories for the groups and religious leaders shown in Table 1. Students and faculty members in the Global Covenant of Religion (GCR) [11] rated the documents produced by each group with a language flexibility score from 1–9 where 1 means the least flexible use of language and 9 means very encompassing use of language. The GCR has a systematic approach for obtaining this score and their members can provide details [11]. This flexibility score serves as the response variable for supervised learning. Table 1 shows the scores, affiliation and number of documents for each group or religious leader.

Table 1. Data sources

Group	Score	Affiliation	Number
Westboro Baptist	1	Baptist	419
Faithful Word Baptist	2	Baptist	228
Nouman Ali Khan	3	Sunni Muslim	88
Dorothy Day	4	Catholic	774
John Piper	4	Baptist	579
Steve Shepherd	4	Christian	728
Rabbinic Texts	6	Jewish	166
Unitarian Texts	7	Unitarian	276
Meher Baba	8	Spiritualist	265

The documents were randomly put in bins of 5 or more documents. We cleaned the text by removing punctuation, converting to lowercase, and converting all numbers to a single symbol. The tokens were stemmed, but stopwords (such as, "the", "an") were not removed. For Part-of-Speech (POS) labeling we applied the Maxent POS Tagger from the python Natural Language Toolkit (NLTK) package to the corpus [10]. Word counts were generated for each unique word/POS tag combination for each bin. The top 10 most frequent adjectives and adverbs were selected for each bin and used as the keywords. For each group we put 70% of the bins in a training set and the remaining 30% in a testing set.

4 Hyperparameter Optimization

There are two major components we used for the computational performative analysis of religious text: text signal processing and machine learning. The work in this paper focused on optimizing hyperparameters for signal processing or feature engineering. The details of the different signals acquired from religious text and the subsequent machine learning methods are given in [16]. Here we

describe the optimization of the three major signals: co-occurrence window; context window; and network adjacency angle.

To capture the variations in linguistic flexibility of keywords within religious discourse, we implemented a context vector semantic density algorithm in python based on research by [13]. Using the pre-processed tokens as described in the data section above, a co-occurrence matrix, $X = [x_{ij}]$ was constructed with $i, j \in \{1, 2, \ldots, v\}$, where v is the size of the vocabulary. The x_{ij} elements of X capture how often each word appeared with other words in the vocabulary. This was done by iterating over each word in the bin and counting the words within a $\pm k$ sized co-occurrence window.

This co-occurrence window directly affects the representation of the distribution of the language within the corpus. A larger window includes more information about the words occurring around other words. This may better capture the linguistic signals but it may also add noise. Our optimization explored windows of size 2 to 6.

Once the co-occurrence matrix, X, was constructed, a distributional semantic matrix, D, was developed to reduce the computational load. D was obtained from X using truncated SVD. This reduced the column space of D to 50 components.

Next, context vectors were created from D. The context vectors for each keyword in the bin were developed by extracting the words within an r-sized context window surrounding the target and summing the rows of D for the words within the window. In contrast to the co-occurrence window which measures general word usage, the context window size impacts the specific measure of a word's usage. Without knowledge of how the proximity of a word affects the analysis of the variability of its usage, we varied the window size from 2 to 6.

Utilizing the distributional semantic matrix, D, we created a graph to estimate semantic density using the igraph package in python [7]. This was done by first creating a v-v adjacency matrix by computing the cosine similarity of each row h in D. Values in this matrix were converted to 0 or 1 if they were above or below the determined network adjacency angle threshold, respectively. This matrix was then used to create a graph where a node represents a word and an edge is assigned between two nodes if the associated value in the adjacency matrix is 1.

The network adjacency angle can significantly impact the accuracy of estimates of word usage variability. This is due to the fact that an inaccurate value of network adjacency angle can over or under estimate the relationship of words in the graph. To optimize network adjacency angle its value was varied by 15° increments over the range 15–75.

5 Results and Conclusions

To judge the relationship of the hyperparameters to linguistic flexibility we used two measures for linguistic flexibility: semantic density and eigenvalue centrality. Eigenvalue centrality is defined in [8]. The semantic density is computed for each target bin word. Let C_i be the set of context vectors for target bin word i. We

then estimated the average cosine similarity of all the context vectors for that word by randomly sampling 2 context vectors from the set, C_i and calculated the cosine similarity between them. This was performed over $n = 1000$ iterations. The resulting values were averaged to produce the semantic density, SD, for that target word as shown below.

$$SD(w_i) = \frac{1}{n} \sum_{k=1}^{n} \frac{c_{ia_k} \cdot c_{ib_k}}{||c_{ia_k}|| * ||c_{ib_k}||}$$

where a_k, b_k are randomly chosen from C_i in iteration k.

Optimization results show that up to a point an increase of the co-occurrence and context vector window sizes results in an increase in the average semantic density of the keywords. For example, for random forests as the machine learning method the optimal value occurs when both co-occurrence and context vector window are set to 5. These values allow for important signals to be held in the context vectors, while avoiding excess noise.

The response surface for average eigenvector centrality was multi-modal. Nonetheless, the average eigenvector centrality trends upward as the co-occurrence window size increases and the network adjacency angle decreases. An increase in the co-occurrence window size results in a greater connection between the words within the discourse, and the lower adjacency angle produces more edges between the nodes within the graphs. As eigenvector centrality is a bounded variable between 0 and 1, the limit of 1 occurs under the extreme conditions of 30° and a co-occurrence window size of 6.

For model performance, we used a robust measure that allows an error margin of 1, as shown in Eqs. 1 and 2. Let m be the machine learning method, i.e., random forests or support vector machines (SVM). Let $\hat{y}$ be the predicted flexibility score and y be the actual score. Finally, let B be the set of document bins. So, for $b \in B$ the accuracy, acc, is given by

$$acc(b, m) = \begin{cases} 1, & \text{if } |\hat{y} - y| \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$acc(m) = \frac{1}{|B|} \sum_{b \in B} acc(b, m). \tag{2}$$

Finally, Table 2 shows the overall increase in accuracy from parameter optimization. The table shows the machine learning method (RF: random forest and SVM: support vector machines), the previous results reported in [16], and the results from optimized hyperparameters. These results come from three-fold cross-validation.

The results show performative analysis can provide promising predictions of the flexibility evident in religious documents from a wide range of groups and the methods are not highly sensitive to parameter settings. The results in Table 2 indicate modest loss of accuracy from using default or arbitrarily chosen settings. This suggests that carefully chosen settings will put the performance in ranges

Table 2. Accuracy comparison for signal processing

Method	Previous	Optimized
SVM	84%	86%
RF	86%	92%

not much different from those found by hyperparameter optimization. Clearly the robustness this approach suggests that the signals, i.e., the feature engineering when combined with the classification methods is doing the heavy lifting in predicting the flexibility of religious speech. Notice that the difference between the performance of optimized random forests and support vector machines is the same as the difference previous random forest results and hyperparameter optimized results.

These results suggest that future work should focus on finding new signals or relationships between signals. Clear areas for this new work are in the application of word embedding techniques which may produce signals not easily engineered or identified by feature engineering.

References

1. Abu-Nimer, M.: Conflict resolution, culture, and religion: toward a training model of interreligious peacebuilding. J. Peace Res. **38**(6), 685–704 (2001)
2. Barsalou, L.: Flexibility, structure, and linguistic vagary in concepts: manifestations of a compositional system of perceptual symbols. Theor. Mem. **1**, 29–31 (1993)
3. Blatter, B., Patel, V.: Exploring dangerous neighborhoods: latent semantic analysis and computing beyond the bounds of the familiar. In: AMIA 2005 Symposium Proceedings, pp. 151–155 (2005)
4. Boussidan, A., Ploux, S.: Using topic salience and connotational drifts to detect candidates to semantic change. In: Proceedings of the Ninth International Conference on Computational Semantics, pp. 315–319. Association for Computational Linguistics (2011)
5. Bullinaria, J.A., Levy, J.: Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. Behav. Res. Methods **44**(3), 890–907 (2012)
6. Cook, P., Stevenson, S.: Automatically identifying changes in the semantic orientation of words. In: LREC (2010)
7. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJ. Complex Syst. **1695**(5), 1–9 (2006)
8. Estrada, E., Rodriguez-Velazquez, J.A.: Subgraph centrality in complex networks. Phys. Rev. E **71**(5), 056103 (2005)
9. Hassner, R.E.: War on Sacred Grounds. Cornell University Press, Ithica (2009)
10. Loper, E., Bird, S.: NLTK: the natural language toolkit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, vol. 1, pp. 63–70. Association for Computational Linguistics (2002)

11. Marcus, B.: Global covenant of religion. (2015). https://sites.google.com/a/globalcovenant.org/global-covenant/team
12. Ochs, P.: The possibilities and limits of inter-religious dialogue. In: Omer, A., Appleby, R.S., Little, D. (eds.) Religion, Conflict, and Peacebuilding, pp. 488–534. Oxford University Press, New York (2015)
13. Sagi, E., Kaufmann, S., Clark, B.: Semantic density analysis: Comparing word meaning across time and phonetic space. In: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, pp. 104–111. Association for Computational Linguistics (2009)
14. Tompkins, P.J.: Human factors considerations of undergrounds in insurgencies. U.S. Army Special Operations Command, Ft. Bragg, NC, January 2013
15. United States Special Operations Command: Gray zone, September 2015. https://info.publicintelligence.net/USSOCOM-GrayZones.pdf
16. Venuti, N., Sachtjen, B., McIntyre, H., Mishra, C., Hays, M., Brown, D.E.: Predicting the tolerance level of religious discourse through computational linguistics. In: 2016 IEEE Systems and Information Engineering Design Symposium (SIEDS), pp. 309–314. IEEE Press (2016)
17. Yang, M., Wong, S., Coid, J.: The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. Psychol. Bull. **136**(5), 740 (2010)

APART: Automatic Political Actor Recommendation in Real-time

Mohiuddin Solaimani[1], Sayeed Salam[1], Latifur Khan[1], Patrick T. Brandt[2], and Vito D'Orazio[2(✉)]

[1] Department of CS, The University of Texas at Dallas, Richardson, USA
{mxs121731,sxs149331,lkhan}@utdallas.edu
[2] School of Economic, Political, and Policy Sciences, The University of Texas at Dallas, Richardson, USA
{pbrandt,dorazio}@utdallas.edu

Abstract. Extracting actor data from news reports is important when generating event data. Hand-coded dictionaries are used to code actors and actions. Manually updating dictionaries for new actors and roles is costly and there is no automated method. We propose a dynamic frequency-based actor ranking algorithm with partial string matching for new actor-role detection, based on similar actors in the CAMEO dictionary. This is compared to a graph-based weighted label propagation baseline method. Results show our method outperforms the alternatives.

1 Introduction

Political event data [6,17] are coded from news reports and take the form of "who-did/said-what to whom." Automated event coders (e.g., PETRARCH [2] or BBN Accent [8]) use manually-entered dictionaries to identify actions and actors, and assign roles (e.g., government employee, media, etc.). The set of actions or verbs is rather finite and matched to CAMEO [16], but the set of nouns for actors and roles is large and constantly changing. Accurate and timely event data coding thus needs a real-time system to detect new actors and roles.

Designing a dictionary update system poses several challenges. First, actors may have multiple aliases: 'Barack H. Obama', or 'President Obama'. Second, the roles of actors change over time: Shimon Peres served multiple Israeli political roles. Finally, processing a large volume of international news articles demands scalable, distributed computing to detect this. We develop a real-time, distributed recommendation framework to identify actors and roles. We gather international news articles and pre-process them via Stanford CoreNLP and PETRARCH to extract actor data. Next, an unsupervised ranking algorithm recommends new actors and roles. A graph-based, weighted label propagation [11] actor-role recommendation method is implemented as a baseline.

We make three contributions: (1) is a novel time frequency- and window-based unsupervised new actor and role recommendation technique with alias actor grouping; (2) is a scalable real-time framework for coding actors; (3) is an improvement over a graphical propagation based actor-role recommendation.

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 342–348, 2017.
DOI: 10.1007/978-3-319-60240-0_42

2 Background

The first machine coder for event data was introduced by [15] anad was then developed into TABARI [14]. The DARPA-funded Integrated Crisis Early Warning System builds on this [12] and now provides global event data from 1995 [7]. Schrodt and Van Brackle [17] illustrate generating events from news texts. This earlier work focuses on event data generation and analysis, not incorporating dynamic dictionaries. Beieler et al. [6] note that an event data challenge is manually developing CAMEO dictionaries with new entries. Relatedly, Saraf et al. [13] show a recommendation model to detect reports of civil unrest.

Dynamically building actor dictionaries uses several tools: **Stanford CoreNLP** is a tool to annotate text with part-of-speech (POS) taggers and named entity recognition (NER), etc. [3]. **CAMEO** (Conflict and Mediation Event Observations) codes events, including actor specifications, to record political events [16]. **PETRARCH** (A Python Engine for Text Resolution And Related Coding Hierarchy) is a program that takes text in Penn Tree format [4] from CoreNLP and generates CAMEO-coded events [2]. **Apache Spark** is an open-source, distributed framework for data analytics that avoids the I/O bottleneck of the conventional two-stage MapReduce programs [1].

3 Framework

Figure 1 shows the actor recommendation framework, an extension of [18]. A web-scraper [5] collects news periodically from about 400 RSS Feeds and extracts the main content which is then shipped through Apache Kafka to Apache Spark-based processing modules [18]. In the data processing unit, CoreNLP parses the reports and extracts meta-data, a parse tree, and NER. PETRARCH codes events from the parse tree, while New Actor uses NER and PETRARCH output. It captures PETRARCH generated events where political actors are unknown, and crosschecks it with NER to find new actors. Using our New Actor algorithm, new actors and roles are recommended. We provide a GUI and dashboard so that users can validate recommended actors and their roles for dictionary updates.

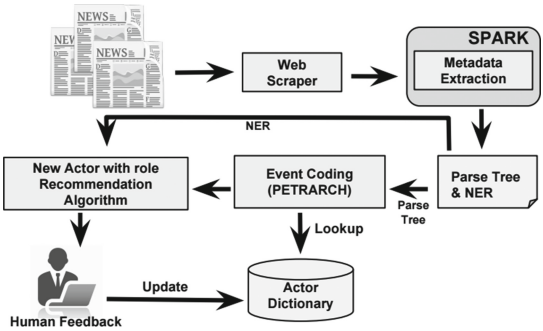


Fig. 1. Framework for real-time new political actor recommendation

4 Recommending a New Actor and Role

Actors: The initial list of potential new actors and roles are those where PETRARCH suggests an event, but the actor is not present in CAMEO. From this list, we group the variations of an actor's name under a single actor identity (e.g., *Barack H. Obama* for *President Obama*, *Barack Obama*, etc.). Several similarity measures are used, such as Levenshtein Distance [10] and MinHash [9] to group the name variations. Each of these methods requires a similarity threshold sim_{th}. A score is generated for each actor via the following equation: $rank(a) = \sum_{d \in D} tf(a,d) \times df(a,D)$. The term frequency, $tf(a,d)$, shows the frequency of an Actor a in Document d. We use document frequency $df(a,D)$ to show the time window of document set D, where Actor a appears at least once. A buffered time window W of length L is maintained to find N actors, which are merged with the previous window's list. The rank statistics are updated after the merge. After L windows, new actors and their roles are recommended if their occurrences in the L_{th} buffered window exceed the threshold TH.

Roles: Role recommendations are based the similarity of new actors to those with whom they interact. Actors from one country or government are more likely to interact with ones from the same country or government. When a new potential actor is identified, the top M most frequent roles are recommended from among the co-occurring roles of any co-occurring actors. Co-occurring actors are those actors that appear in the same document. When an actor appears in two documents, the roles of all co-occurring actors in both articles are included. When a co-occurring actor has multiple roles, we include each.

Recommendations: In each time window all articles are scanned for the potential actor list with rankings and role recommendations. If an actor comes in the top N rankings in multiple time windows, s/he has high probability of being a new political actor. For the threshold TH, new actors are those that appear in 5 or more windows, but those who appear less than 3 are discarded.

5 Experiments

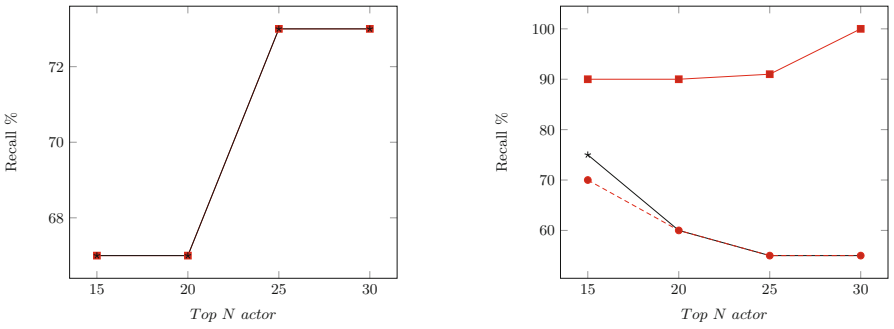
To evaluate the framework, in $L = 9$ time windows at 2 h intervals, newly published news articles from about 400 RSS feeds were scraped. For this empirical study, we estimate thresholds for edit distance and min-hash based methods to be $sim_{th} = .75$ and $sim_{th} = .45$, respectively, based on minimizing false positives using known alias groupings provided by the CAMEO actor dictionary.

A graph-based role detection technique is the baseline comparison. This models the interactions between actors using a Graph, $G = (V, E)$, where V is the set of existing and recommended actors and E contains their edge interactions to infer roles using the weighted label propagation technique [11]. The co-occurrence of two actors in the same document is an interaction and the frequency of co-occurring actors is their weight. After formulating the graph, we begin the weighted label propagation algorithm. The existing actors, those that

are in CAMEO, have their roles as the labels. The recommended actors, those not in CAMEO, begin with an *empty* label.

We next eliminate some well known actors from the CAMEO actor dictionary and try to recover them. This experiment removed 5, 10, and 15 actors, but due to space we only present the results from removing 15. The results are consistent when removing 5 and 10. Since PETRARCH will not code events for the deleted actors, they are recommended as if they are newly discovered. The recall is the percentage of the removed actors then recommended by our algorithm. Precision is not computed because all the recommended actors are political.

Performance Evaluation. Figure 2(a) plots recall when grouping actor aliases using edit distance and MinHash. As the top N recommended actors increase, the recall for retrieving a deleted actor increases. Of the 15 actors removed from CAMEO, 12 are included in the top 25 recommendations. Although not apparent in Fig. 2(a), for closely matched actor aliases like 'Donald Trump' and 'Donald J. Trump', edit distance performs slightly better than MinHash. In Fig. 2(b), examines recall with 5 suggested roles. The suggestion is a success if one of the recommended roles is the true role in CAMEO. For Exact match, the recommended role must be exactly identical to the true role. For edit distance and MinHash, partial matching is allowed using substrings (e.g., suggesting USA for USAGOV is a success) and thresholds, to allow for roles that are near the true role but not identical. While edit distance is the better method, MinHash and Exact show good recall.



(a) Deleted actor = 15 (b) Deleted actor = 15

Fig. 2. Performance for (a) Actor recommendation and (b) Role recommendation. Recall: Edit distance —■—, MinHash —⋆—, Exact match - -●- - (b only)

Word2Vec. We have experimented with using Word2Vec for role recommendations by training a model to predict co-occurring actors. Specifically, for each document in each of the L windows, we extract all actors recognized by NER and input that actor list as an observation for training the Word2Vec model. When the New Actor algorithm proposes a new actor, the Word2Vec model predicts

co-occurring actors, extracts their roles, and proceeds with role recommendation. In this way, it is possible that an actor that does not co-occur, in the strict sense of appearing in the same article, may be among the list of predicted co-occurring actors whose roles are extracted. After applying this method, the suggested roles had less than 5% recall due to poor performance by the co-occurring actor predictions. We suspect this was due to the lack of alias groupings as a pre-processing step, and leave this for future research.

Graph-based Comparison. The frequency-based approach for role recommendation outperforms the graph-based approach by a considerable margin. We fix the values of the number of recommended actors per window at $N = 20$, and again delete 15 actors from CAMEO. The frequency-based approach outperforms the graph-based one for each similarity measure: 90 to 27 for edit distance, 60 to 20 for MinHash, and 61 to 20 for Exact. The frequency-based approach considers roles from existing actors in the CAMEO dictionary, while the graph-based approach considers roles from neighbors who are either existing or new actors. Thus the error with one role assignment can propagate to others.

Validation. As validation of the system, Table 1 shows the top recommended actors across all windows for the two string similarity measures, using the same thresholds and time-windows as in the experiments. Both methods detect similar roles for identical actors, but suggest different actor lists. Given that we are building a recommendation system for expanding a political actor dictionary, we can see that the new actors are quite appropriate—Donald Trump, Amir Sheik Sabah, Rodrigo Duterte, and others are prominent government officials.

Table 1. List of recommended actors with their roles

Edit distance		MinHash	
Actor	Top 3 roles	Actor	Top 3 roles
Donald Trump	USA, USAELI, LEG	Sediq Sediqqi	AFG, PPL, UAF
Sediq Sediqqi	UAF, AFG, PPL	Donald Trump	USA, USAELI, LEG
Amir Sheik Sabah Al-Ahmad Al-Jaber Al-Sabah	SAUMED, SAUGOV, KWTMEDGOV	Amir Sheik Sabah Al-Ahmad Al-Jaber Al-Sabah	SAUMED, SAUMEDGOV, KWTMEDGOV
Lynne O' Donnel	AFG, AFGUAF, PPL	Rodrigo Duterte	PHLGOV, GOV, PHL
Rodrigo Duterte	PHLGOV, GOV, PHL	Antio Carpio	JUD, PHL, CRM

6 Conclusion and Future Work

Political actor dictionaries are integral to political event data coding. We address the problem of detecting and recommending new actors and their roles in real-time. A Spark-based framework with unsupervised rankings of new actor aliases on a periodic basis is proposed. Currently, this is only to find new actors, but it can be extended to recommend new events in the CAMEO verb dictionary.

Acknowledgments. Support from the National Science Foundation (NSF) SBE-SMA-1539302, CNS-1229652, and SBE-SES-1528624; and the Air Force Office of Scientific Research (AFOSR): FA-9550-12-1-0077. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the NSF or the AFOSR.

References

1. Apache Spark. http://spark.apache.org/
2. Petrarch. http://petrarch.readthedocs.org/en/latest/
3. Stanford CoreNLP. http://nlp.stanford.edu/software/corenlp.shtml
4. The Penn Treebank Project. https://www.cis.upenn.edu/~treebank/
5. Web Scraper. http://oeda-scraper.readthedocs.io/en/latest
6. Beieler, J., Brandt, P.T., Halterman, A., Schrodt, P.A., Simpson, E.M.: Generating political event data in near real time: opportunities and challenges. In: Michael Alvarez, R. (ed.) Computational Social Science: Discovery and Prediction, pp. 98–120. Cambridge University Press, Cambridge (2016)
7. Boschee, E., Lautenschlager, J., O'Brien, S., Shellman, S., Starz, J., Ward, M.: ICEWS Coded Event Data (2016)
8. Boschee, E., Natarajan, P., Weischedel, R.: Automatic extraction of events from open source text for predictive forecasting. In: Subrahmanian, V.S. (ed.) Handbook of Computational Approaches to Counterterrorism, pp. 51–67. Springer, New York (2013)
9. Broder, A.Z.: On the resemblance and containment of documents. In: Compression and Complexity of Sequences 1997, Proceedings, pp. 21–29. IEEE (1997)
10. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady **10**, 707 (1966)
11. Lou, H., Li, S., Zhao, Y.: Detecting community structure using label propagation with weighted coherent neighborhood propinquity. Phys. A Stat. Mech. Appl. **392**(14), 3095–3105 (2013)
12. O'Brien, S.: Crisis early warning and decision support: contemporary approaches and thoughts on future research. Int. Stud. Rev. **12**(1), 87–104 (2010)
13. Saraf, P., Ramakrishnan, N.: EMBERS autogsr: automated coding of civil unrest events. In: ACM SIGKDD, San Francisco, CA, USA, 13–17 August 2016, pp. 599–608 (2016)
14. Schrodt,P.A.: TABARI: Textual Analysis by Augmented Replacement Instructions (2009). http://eventdata.psu.edu/tabari.html
15. Schrodt, P.A., Davis, S.G., Weddle, J.L.: Political science: KEDS-a program for the machine coding of event data. Soc. Sci. Comput. Rev. **12**(4), 561–587 (1994)

16. Schrodt, P.A., Gerner, D.J., Yilmaz, Ö.: Conflict and mediation event observations (CAMEO): An event data framework for a post Cold War world. In: Bercovitch, J., Gartner, S. (eds.) International Conflict Mediation: New Approaches and Findings. Routledge, New York (2009)
17. Schrodt, P.A., Van Brackle, D.: Automated coding of political event data. In: Subrahmanian, V.S. (ed.) Handbook of Computational Approaches to Counterterrorism, pp. 23–49. Springer, New York (2013)
18. Solaimani, M., Gopalan, R., Khan, L., Brandt, P.T., Thuraisingham, B.: Spark-based political event coding. In: BigDataService, pp. 14–23. IEEE (2016)

Measuring Perceived Causal Relationships Between Narrative Events with a Crowdsourcing Application on Mturk

Dian Hu(✉) and David A. Broniatowski

Department of Engineering Management and Systems Engineering,
The George Washington University, Washington, DC 20052, USA
{hudian,broniatowski}@email.gwu.edu

Abstract. The computational study of narrative is important to multiple academic disciplines. However, prior research has been limited by the inability to quantify each subject's comprehension of the causal structure. With the aid of big data technology and crowdsourcing tools, we aim to design a new approach to analyze the content of narratives in a data-driven manner, while also making these analyses scientifically replicable. The goal of this research is therefore to develop a method that can be used to measure people's understanding of the causal relationships within a piece of text.

Keywords: Crowdsourcing · Narrative · Network · Causal relationships

1 Introduction

The study of narratives is uniquely important to various groups including, but not limited to, public health officials, public relations professionals, market analysts and policy makers. In recent years, the emergence of various social media provides new platforms for the dissemination of narratives. With these new platforms, narratives reach more people simultaneously and the effects of narratives last longer [13, 22]. However, the new platforms have also introduced new problems. One study shows that the spread of information online has accelerated the spread of misinformation with negative implications for public understanding of science and public health [7]. Furthermore, in the field of understanding vaccine hesitation, several studies have found that misconceptions are often transmitted in the form of narratives on social media [3, 8]. These misconceptions have hindered the public's acceptance of factually-accurate public health messages. Thus, to increase the public's familiarity with, and acceptance of, factually-accurate message among the public, we must first understand how the public comprehends these messages, what forms of messages are more acceptable for them and how the public perceives the internal causal relationships within an article.

Many studies have done extensive work to explain how information is spread [2, 6, 16]. However, most prior work has focused on describing the underlying mechanisms, and the external variables affecting the spread. Less work has analyzed how the content of a narrative affects its spread. Merely analyzing the external variables and the spread pattern is not enough to understand why certain misinformation can be

D. Lee et al. (Eds.): SBP-BRIMS 2017, LNCS 10354, pp. 349–355, 2017.
DOI: 10.1007/978-3-319-60240-0_43

more popular than others. More importantly, understanding the dissemination mechanism alone cannot help to improve the quality of messages: the factors that lead people to understand the gist of the message. Effective communications can be better achieved if we can understand both factors: how publics comprehend messages and how publics transmit messages. In this project, we focus on one key internal feature: A narrative's causal structure, since prior work has shown that causal structure is a key component of narrative comprehension [19, 20]. Each reader may comprehend the same piece of text differently. Therefore, the causal structure generated by individual readers is a direct lens through which we can see their perception of the internal logic within a text. Furthermore, the extracted causal structure could provide a new variable that can be used in other studies of information diffusion. The goal of this current project is to develop a new method that can conveniently collect readers' perceptions of a narrative's internal causal structure. We have also proposed a systematic approach to measuring causal structures from a group using graph theory.

Although essential for future work, the tool presented in this paper is still preliminary. In the future, using data collected from this tool, we would first explore what factors lead people to disagree on the existence of causal links. These factors would include both variables from the individual subjects (e.g., individual's education level, familiarity with background information and common sense) and variables from the text messages (e.g., grammar, writing style, conjunction and reasoning effectiveness). Such analyses cannot be achieved without this work: a convenient web-based application that can collect causal graphs from multiple subjects.

2 Background: Representing Text as Causal Networks

The study of narrative reading comprehension has been an active field since the 1970s. Since that time, the field had identified three generations of research and investigations [21]. In the first generation, researchers focused on what readers remember after they read the text [12, 20, 21]. The second generation of research focused on the process by which readers generate products during readings [9, 15, 17, 18]. Nowadays, the third generation of research is still actively constructing theories to explain the findings from the first and the second generations [21]. Among these studies, causal network theory emerged as a replicable way to describe readers' mental presentations of text [19, 20]. When describing a narrative using causal network theory, each event in the narrative is represented by a node, and each causal relationship between events is represented as a link [20]. The rules to establish the causal networks is derived from psycholinguistic literature and rigorous grammars. Traditionally, the graph is constructed by trained judges in this discipline [18, 20]. A simplified rule to determine whether a causal link exists from Event A to Event B is to determine "if A did not occur, then B would not have occurred" [20]. Table 1 shows an example of labeled events in a simple narrative while Fig. 1 shows the associated causal network.

Furthermore, prior studies have shown that readers will be more likely to recall events that have a high number of connections when they perform a recall task after reading the narrative [19, 20]. However, several concerns still exist surrounding these findings: Can average readers generate repeatable causal networks that are as precise as

Table 1. Labeled events in a simple narrative

Event index	Text
1	Daniel arrived in his aunt's house
2	He knocked at the door and
3	rang the doorbell
4	Daniel's cousin was waiting at home
5	He opened the door for Daniel then
6	two dogs ran out to greet Daniel

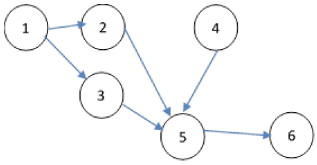


Fig. 1. Causal network associated with the narrative presented in Table 1.

those generated by trained scholars? Do readers compare each pair of potential causal relationships following the complicated grammar rules? If each reader has a causal graph in mind, would it be possible that each of them is indeed recalling the events that are highly connected in the unique graph in his/her own mind? To address these questions, we propose another approach:

Building upon this foundation of causal network theory, we aim to assess if a large group of readers can identify narrative events and construct these causal networks with a certain level of agreement, using our web-based crowdsourcing application.

3 Methods

Due to space limitations, we can only present the outline of the method in this paper. More detailed information on performance and sample outputs of the method section can be found in our prior working paper [11].

Identifying events can be a complicated and subjective process, especially when the task is assigned to a group of people. Thus, we built an event parser based on natural language processing using NLTK [1] to consolidate narrative events. Therefore, in this study, the events are fixed when they are presented to the subjects.

Historically, in off-line settings, creating these causal networks has been both time- and labor-intensive. However, several recent studies have demonstrated that online crowdsourcing is an effective way to perform participatory human research [4]. For example, Amazon's Mechanical Turk (Mturk) has been used in various studies to collect responses from a relatively large group of participants with a wide range of educational and cultural backgrounds [4, 5]. Therefore, we aim to gather causal networks from crowdsourcing workers on Mturk. The current version of this web application is based on HTML-5, CSS and JavaScript technology, a combination that is

especially suited to developing cross-platform application [10]. The tool can directly save the network information into standard JSON data, therefore granting swift responses and data integrity. To ensure that subjects understanding the task, we formulated specific instructions containing the following information: what the nodes represent, what the links represent and how to use the tool to draw and modify causal networks. A short demographic survey is also included at the end of the task.

4 Results from the Usability Study

On March 6, 2017, we launched our first study on Mturk using this tool. Turkers were eligible to complete the task if they had a HIT approval rate of 95% or higher and were in the United States. 20 subjects were recruited on Mturk without intervention or selection beyond these criteria. The first study used one narrative, borrowed from a prior study in narrative comprehension [19], with 22 events. The Mturk workers could refer to the instructions while finishing the task. In our first case study using this tool, we wanted to assess three objectives: (1) Can Mturk Workers understand our instructions? (2) How long does it take for Mturk Workers to accomplish the task? (3) To what level do the Mturk user agree with each other on each potential causal link? Are there links that exist (or don't exist) with statistical significance?

Among the 20 workers, the average time to finish the task (both drawing and survey) was 23 min, with a minimum of 9 min, and a maximum of 50 min, indicating that our tool can effectively reduce the drawing time for creating each graph. Even with a demographic survey, this is a 50% percent reduction compared to our pilot study on fellow researchers without the drawing tool. To analyze our causal network data, we first aggregated each subject's responses into a 22*22 matrix with one entry for each pair of events in this narrative. That is, we have a matrix where each cell represents the frequency of Mturk Workers who agree that there is a link between these two events. Table 2 shows the partial results of the first five nodes. For example, the number 11 means 11 out of 20 workers had agreed that there should be a causal relationship from Event 4 to Event 5. As we can observe from the partial results, there are many cells with value 0 inside the matrix. Theoretically, it means that 20 subjects all agree that there should be no link between these two events.

Table 2. Number Of Workers who agree that there should be a link between two events

	Event (From)	Event 2	Event 3	Event 4	Event 5
Event (To)	0	0	1	0	0
Event 2	**14**	0	0	0	0
Event 3	8	**3**	0	0	0
Event 4	1	1	**14**	0	1
Event 5	0	0	8	**11**	0

Figure 2 is a histogram of the frequency observed in Table 2. The data indicate that, when dealing with most pairs, participants are more likely to think there is no link between two events.

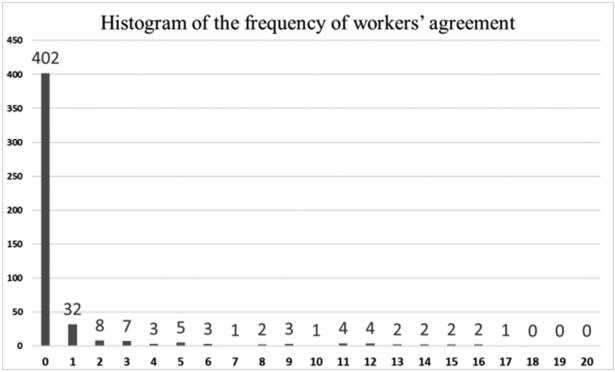


Fig. 2. Histogram of the frequency of workers' agreement

5 Discussion

To make more sense of the data, we performed a power analysis to assess whether the graph as a whole is statistically significant ($p < 0.05$). Between two ordered events, there can only be two scenarios: 1) link exists, or 2) link does not exist. Therefore, we can treat a causal network graph as a series of binary decisions. Thus, the binomial distribution would be appropriate to assess the graphs' significance. To reach $P(\text{graph}) < 0.05$ as a graph, given that we have 22 narrative events, and excluding self-links:

$$P(\text{link}) < \frac{P\ (\text{graph})}{\textbf{number of links}}, \textit{where} \frac{P\ (\text{graph})}{\textbf{number of links}} = \frac{0.05}{22^2 - 22} = 0.000108$$

Therefore, in this study, to ensure statistical significance, each link must ensure that

$$\binom{N}{n} p^n (1-p)^{N-n} < 0.000108$$

We assume that, absent information regarding causal relationships, people will randomly draw a connection between two events with a probability of 0.5 *a priori*. Assuming a binominal distribution with $p_1 = 0.5$, and $N = 20$ (number subjects), we need $n = 19$ or 20 to say that a link exists with statistical significance based on the result. Similarly, we need $n = 0$ or 1 to say that a link does not exist with statistical significance. However, in our current data, there is no pair where 19 or 20 out of 20 subjects had agreed that there is a link. Besides of the fact that we only have a small pool of subjects, another theoretical explanation is that the p_1 might not be 0.5 in our experimental settings. Since we are launching the task on Mturk, our experiment is limited by participants' effort in critical thinking. When Mturk workers are working on this project, they have the motivation to go through the task as quickly as possible to gain maximum profit per hour. Larger sample sizes will likely yield more significant results.

6 Discussion and Future Work

In this paper, we described the development of a new method that can be used to crowdsource the analysis of narratives. Moreover, we have examined the performance of the tool based on a usability study. We proposed a systematic and replicable approach to directly measure people's understanding of the causal relationships within online narratives. In the future, we plan to modify the tool as follows: we would present each pair of narrative events to the workers, asking them whether there is a causal relationship. After the workers have finished assessing all pairs, we would present the drawing panel to the workers again, giving them another chance to review their own decisions. Furthermore, network information saved by this tool can be integrated with other popular network analysis software. SNAP [14], for example, can import JSON based data and extract network properties for more in-depth analysis. We, therefore, aim to continue improving the use cases of this tool to make it more convenient and flexible for scholars and other professionals.

Acknowledgment and Disclaimer. This work was supported in part by the National Institute of General Medical Sciences under grant number 5R01GM114771. 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. Bird, S., et al.: Natural Language Processing with Python. O'Reilly Media, Inc. (2009)
2. Boyd, S., et al.: Randomized gossip algorithms. IEEE/ACM Trans. Networking (TON) **14** (SI), 2508–2530 (2006)
3. Broniatowski, D.A., et al.: Effective vaccine communication during the Disneyland measles outbreak. Vaccine **34**(28), 3225–3228 (2016)
4. Buhrmester, M., et al.: Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. **6**(1), 3–5 (2011)
5. Callison-Burch, C., Dredze, M.: Creating speech and language data with Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp. 1–12. Association for Computational Linguistics (2010)
6. Cowan, R., Jonard, N.: Network structure and the diffusion of knowledge. J. Econ. Dyn. Control **28**(8), 1557–1575 (2004)
7. Del Vicario, M., et al.: The spreading of misinformation online. Proc. Natl. Acad. Sci. **113** (3), 554–559 (2016)
8. Dredze, M., et al.: Zika vaccine misconceptions: a social media analysis. Vaccine **34**(30), 3441 (2016)
9. Gernsbacher, M.A., et al.: Investigating differences in general comprehension skill. J. Exp. Psychol. Learn. Memory Cogn. **16**(3), 430 (1990)
10. Heitkötter, H., Hanschke, S., Majchrzak, Tim A.: Evaluating cross-platform development approaches for mobile applications. In: Cordeiro, J., Krempels, K.-H. (eds.) WEBIST 2012. LNBIP, vol. 140, pp. 120–138. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36608-6_8

11. Hu, D., Broniatowski, D.A.: Designing a Crowdsourcing Tool to Measure Perceived Causal Relationships Between Narrative Events. http://sbp-brims.org/2016/proceedings/LB_129.pdf
12. Kintsch, W., Van Dijk, T.A.: Toward a model of text comprehension and production. Psychol. Rev. **85**(5), 363 (1978)
13. Kozinets, R.V., et al.: Networked narratives: Understanding word-of-mouth marketing in online communities. J. Mark. **74**(2), 71–89 (2010)
14. Leskovec, J., Sosič, R.: SNAP: a general purpose network analysis and graph mining library in C++, June 2014. http://snap.stanford.edu/snap
15. McKoon, G., Ratcliff, R.: Inference during reading. Psychol. Rev. **99**(3), 440 (1992)
16. Shutters, S.T., Cutts, B.B.: A simulation model of cultural consensus and persistent conflict. In: Proceedings of the Second International Conference on Computational Cultural Dynamics, pp. 71–78 (2008)
17. Singer, M., et al.: Minimal or global inference during reading. J. Mem. Lang. **33**(4), 421 (1994)
18. Trabasso, T., et al.: Logical necessity and transitivity of causal relations in stories. Discourse Processes **12**(1), 1–25 (1989)
19. Trabasso, T.: Causal cohesion and story coherence (1982)
20. Trabasso, T., Van Den Broek, P.: Causal thinking and the representation of narrative events. J. Mem. Lang. **24**(5), 612–630 (1985)
21. Van den Broek, P., et al.: The landscape model of reading: Inferences and the online construction of a memory representation. The construction of mental representations during reading, pp. 71–98 (1999)
22. Winterbottom, A., et al.: Does narrative information bias individual's decision making? A systematic review. Soc. Sci. Med. **67**(12), 2079–2088 (2008)

Author Index