Sun-Ki Chai
John J. Salerno
Patricia L. Mabry (Eds.)

LNCS 6007

Advances in Social Computing

Third International Conference on Social Computing, Behavioral Modeling, and Prediction, SBP 2010
Bethesda, MD, USA, March 2010, Proceedings

Springer

Lecture Notes in Computer Science 6007

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

Sun-Ki Chai John J. Salerno
Patricia L. Mabry (Eds.)

Advances in Social Computing

Third International Conference on Social Computing, Behavioral Modeling, and Prediction, SBP 2010
Bethesda, MD, USA, March 30-31, 2010
Proceedings

Springer

Volume Editors

Sun-Ki Chai
University of Hawaii, Department of Sociology
2424 Maile Way, Honolulu, HI 96822, USA
E-mail: sunki@hawaii.edu

John J. Salerno
Air Force Research Laboratory, Rome Research Site, AFRL/RIEF
525 Brooks Road, Rome, NY 13441, USA
E-mail: john.salerno@rl.af.mil

Patricia L. Mabry
National Institute of Health (NIH)
Department of Behavioral and Social Sciences Research
31 Center Drive, Bethesda, MD 20892-2027, USA
E-mail: mabryp@od.nih.gov

Library of Congress Control Number: 2010922361

CR Subject Classification (1998): H.3, H.2, H.4, K.4, J.3, H.5

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-12078-4 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-12078-7 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

Social computing is concerned with the study of social behavior and social context based on computational systems. Behavioral modeling provides a representation of the social behavior, and allows for experimenting, scenario planning, and deep understanding of behavior, patterns, and potential outcomes. The pervasive use of computer and Internet technologies by humans in everyday life provides an unprecedented environment of various social activities that, due to the platforms under which they take place, generate large amounts of stored data as a by-product, often in systematically organized form. Social computing facilitates behavioral modeling in model building, analysis, pattern mining, and prediction. Numerous interdisciplinary and interdependent systems are created and used to represent the various social and physical systems for investigating the interactions between groups, communities, or nation-states. This requires joint efforts to take advantage of the state-of-the-art research from multiple disciplines improving social computing and behavioral modeling in order to document lessons learned and develop novel theories, experiments, and methodologies to better explain the interaction between social (both informal and institutionalized), psychological, and physical mechanisms. The goal is to enable us to experiment, create, and recreate an operational environment with a better understanding of the contributions from each individual discipline, forging joint interdisciplinary efforts.

This volume comprises the proceedings of the third international workshop on *Social Computing, Behavioral Modeling and Prediction,* which has grown tremendously. In the first year we had 34 papers submitted; this year we had 78 (we accepted 26 papers to be presented during the main conference and 23 during the poster board session; 4 were withdrawn). The submissions were from Asia, Oceania, Europe, and the Americas. We were extremely delighted that the technical program encompassed keynote speeches, invited talks, and high-quality contributions from multiple disciplines. We warmly welcome all to actively participate in the interdisciplinary endeavors, and truly hope that our collaborative, exploratory research can advance the emerging field of social computing. This year we also introduced two new parts to the conference; four tutorials on Monday and a workshop on Thursday that included activities such as meetings with government program managers. The goal of these two new pieces was to continue the interaction and assimilation between the various disciplines that we believe need to come together to understand and pursue research in this area.

The accepted papers cover a wide range of interesting topics: (1) *social network analysis* such as social computation, complex networks, virtual organization, and information diffusion; (2) *modeling* including cultural modeling, statistical modeling, predictive modeling, cognitive modeling, and validation process; (3) *machine learning and data mining*, link prediction, Bayesian inference, information extraction,

information aggregation, soft information, content analysis, tag recommendation, and Web monitoring; (4) *social behaviors* like large-scale agent-based simulations, group interaction and collaboration, interventions, human terrain, altruism, violent intent, and emergent behavior; (5) *public health* such as alcohol abuse, disease networks, pandemic influenza, and extreme events; (6) *cultural aspects* like cultural consensus, coherence, psycho-cultural situation awareness, cultural patterns and representation, population beliefs, evolutionary economics, biological rationality, perceived trustworthiness, and relative preferences; and (7) *effects and search*, for example, temporal effects, geospatial effects, coordinated search, and stochastic search. It is interesting that if we traced these keywords back to the papers, we could find natural groups of authors of different papers attacking similar problems.

While it may seem at first glance that the topics covered by the papers are too disparate to summarize in any succinct fashion, there are certain patterns that reflect trends in the field of social computing. One is the increasing participation of the human and social sciences in social computing, and well as the active collaboration between such fields and science and engineering fields. Disciplines represented at this conference include not only computer science and electrical engineering, but also psychology, economics, sociology, and public health, a number of interdisciplinary and applied research institutions are also represented.

Another noticeable trend in the accepted papers is the sizable number of papers that address applied topics and/or examine empirical data of human behavior. The types of issues addressed include flu epidemics, the West Nile Virus, identity attitudes amongst Iranians, online knowledge-sharing sites, and social networks in LiveJournal. While there remain a healthy representation of papers that describe abstract models and methodologies (including several stemming from mainstream social science approaches) and software systems, work that is obviously essential to social computing, these are balanced by work that applies such models and systems to the problems of the real world.

A conference like this one cannot be run by a few. We would like to first express our gratitude to all the authors for contributing an extensive range of research topics showcasing many interesting research activities and pressing issues. The regret is ours that due to the space limit, we could not include as many papers as we wished. We thank the Program Committee members for helping review and provide constructive comments and suggestions. Their objective reviews significantly improved the overall quality and content of the papers. We would like to thank our keynote and invited speakers for presenting their unique research and views. We deeply thank the members of the Organizing Committee for helping to run the workshop smoothly; from the call for papers, the website development and update, to proceedings production and registration.

Last but not least, we sincerely appreciate the support from the Office of Behavioral and Social Sciences Research at the National Institutes of Health (NIH) along with the National Institute on Drug Abuse, the National Cancer Institute, and the National Institute on General Medical Sciences at NIH, Air Force Office of Scientific

Research (AFOSR), Air Force Research Laboratory (AFRL), Office of Naval Research (ONR), and the National Science Foundation (NSF). We also would like to thank Alfred Hofmann and Anna Kramer from Springer and Sonja Preston, Kelly Government Services. We thank all for their kind help, dedication and support that made SBP10 possible.

April 2010

Sun-Ki Chai
John Salerno
Patricia Mabry

Organization

Conference Chair: Patricia L. Mabry
Program Chairs: Sun-Ki Chai, John Salerno
Steering Committee: Huan Liu, John Salerno, Sun-Ki Chai
Advisory Committee: Rebecca Goolsby, Terrance Lyons, Patricia L. Mabry, Mary Lou Maher
Poster Session Chair, Lei Yu
Tutorial Chair, Sun-Ki Chai
Workshop Chairs: Fahmida N. Chowdhury, Bethany Deeds
Student Arrangement Chair, Yasmin H. Said
Publicity Chairs: Nitin Agarwal and Magdiel Galan
Proceedings Production Editor, Edward J. Wegman
Web Master: Shamanth Kumar

Workshop Committee

Olga Brazhnik	National Center for Research Resources, NIH
Jennifer Couch	National Cancer Institute, NIH
Yasmin H. Said	George Mason University

Program Committee

Nitin Agarwal, University of Arkansas at Little Rock
Edo Airoldi, Harvard University
Denise Anthony, Dartmouth College
Joe Antonik, AFRL (RI)
Gurdal Arslan, University of Hawaii
Chitta Baral, Arizona State University
Geoffrey Barbier, AFRL (RH)
Herb Bell, AFRL (RH)
Lashon Booker, The MITRE Corporation
Douglas Boulware, AFRL (RI)
Sun-Ki Chai, University of Hawaii
Xueqi Cheng, CAS, P. R.
David Chin, University of Hawaii
Carl Defranco, AFRL (RI)
Guozhu Dong, Wright State University
Richard Fedors, AFRL (RI)
Laurie Fenstermacher, AFRL (RH)
Tim Finin, UMBC
Anthony Ford, AFRL (RI)
Rebecca Goolsby, Office of Naval Research
Lisa Gould, AFRL (RI)
Jessica Halpin
Michael Hinman, AFRL (RI)
Jang Hyun Kim, University of Hawaii at Manoa
Terresa Jackson
Ruben Juarez, University of Hawaii
Byeong-Ho Kang, University of Tasmania
Anne Kao, Boeing
Douglas Kelly, AFRL (RH)
Masahiro Kimura, Ryukoku University
Irwin King, Chinese University of Hong Kong
Alexander Levis, George Mason University

Huan Liu, Arizona State University
Mitja Lustrek, Jozef Stefan Institute
Patricia L. Mabry, National Institutes of Health
Jim Mayfield, Johns Hopkins University
Janet Miller, AFRL (RH)
Hiroshi Motoda, Osaka University & AOARD
Keisuke Nakao, University of Hawaii at Hilo
Dana Nau, University of Maryland
Kouzou Ohara, Aoyama Gakuin University
Youngok Pino, AFRL (RI)
Bonnie Riehl, AFRL (RH)
Kazumi Saito, University of Shizuoka
John Salerno, AFRL (RI)
Antonio Sanfilippo, Pacific Northwest Nat Lab
Hessam Sarjoughian, Arizona State University
Jaideep Srivastava, University of Minnesota
Tracy St. Benoit, University of Central Florida
Gary Strong, Johns Hopkins University
V.S. Subrahmanian, University of Maryland
George Tadda, AFRL (RI)
Lei Tang, Arizona State University
John Tangney, Office of Naval Research
Ray Trechter, Sandia National Labs
Zhijian Wang, Zhejiang University
Rik Warren, AFRL (RH)
Edward J. Wegman, George Mason University
Graham Williams, Togaware
Xintao Wu, University of North Carolina at Charlotte
Ronald Yager, Iona College
Laurence T. Yang, STFX
Michael Young, AFRL (RH)
Philip Yu, UI Chicago
Lei Yu, Binghamton University
Mark Zappavigna, AFRL (RI)
Daniel Zeng, University of Arizona
Jianping Zhang, The MITRE Corporation
Jun Zhang, AFOSR

Table of Contents

Beyond Degree Distributions: Local to Global Structure of Social Contact Graphs

Stephen Eubank, Anil Vullikanti, Maleq Khan, Madhav Marathe, and Chris Barrett

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute at Virginia Tech
1880 Pratt Drive
Blacksburg, Virginia 24061, USA

Abstract. The structure and dynamical properties of networked systems are often characterized by the degree distribution of the underlying graph. The degree distributions of many real world networks have often been found to be power laws, and in this approach, many properties of dynamical processes occurring on these networks, such as attack rates of epidemics, have been related to the power law exponents. A potential problem in many of these results is that the variability among the space of graphs having a specified degree distribution is not usually taken into account.

We show by explicit construction that local properties such as the degree distribution or assortativity are not sufficient to characterize the global dynamics of diffusion processes, such as epidemics. We use a simple Markov chain based on "edge flips"' to generate random graphs with given degree and assortativity properties, and study epidemic properties of such graphs. We find that the "epidemic curves" in these graphs are significantly different, and the difference increases as the Markov chain gets further from its starting point.

Our work provides a cautionary note on the use of local random graph models to infer global network and dynamical properties. Furthermore, it introduces the edge flipping methodology as a tool for investigating the effects of a broad range of hypothesized local network structures on the global dynamics of any process.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, p. 1, 2010.

Endogenous Market-Clearing Prices and Reference Point Adaptation

Arnaud Z. Dragicevic

École Polytechnique ParisTech and CIRANO
2020, rue University
Montréal (QC) H3A 2A5 Canada
arnaud.dragicevic@cirano.qc.ca

Abstract. When prices depend on the submitted bids, *i.e.* with endogenous market-clearing prices in repeated-round auction mechanisms, the assumption of independent private values that underlines the property of incentive-compatibility is to be brought into question; even if these mechanisms provide active involvement and market learning. In its orthodox view, adaptive bidding behavior imperils incentive-compatibility. We relax the assumption of private values' independence in the repeated-round auctions, when the market-clearing prices are made public at the end of each round. Instead of using game-theory learning models, we introduce a behavioral model that shows that bidders bid according to the anchoring-and-adjustment heuristic, which neither ignores the rationality and incentive-compatibility constraints, nor rejects the posted prices issued from others' bids. Bidders simply weight information at their disposal and adjust their discovered value using reference points encoded in the sequential price weighting function. Our model says that bidders and offerers are sincere boundedly rational utility maximizers. It lies between evolutionary dynamics and adaptive heuristics and we model the concept of inertia as high weighting of the anchor, which stands for truthful bidding and high regard to freshly discovered preferences. Adjustment means adaptive rule based on adaptation of the reference point in the direction of the posted price. It helps a bidder to maximize her expected payoff, which is after all the only purpose that matters to rationality. The two components simply suggest that sincere bidders are boundedly rational. Furthermore, by deviating from their anchor in the direction of the public signal, bidders operate in a correlated equilibrium. The correlation between bids comes from the commonly observed history of play and each bidder's actions are determined by the history. Bidders are sincere if they have limited memory and confine their reference point adaptation to their anchor and the latest posted price. S-shaped weighting mechanism reflects such a bidding strategy.

Keywords: auctions, incentive-compatibility, rank-dependence, reference point, heuristic, bounded rationality, correlated equilibrium.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, p. 2, 2010.

Modeling Effect of Leaders in Ethno-Religious Conflicts

Lingzhi Luo, Nilanjan Chakraborty, and Katia Sycara

School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. Many incidents of ethno-religious violence in recent history have been associated with the fall of an authoratarian regime or have been perpetrated by dictators. These incidents underline the importance of the roles of political and/or religious leaders in the context of ethno-religious conflicts. In this paper, we extend the computational model of ethno-religious conflicts (based on repeated prisoner's dilemma (PD) game in graphs) proposed in [1], to include the effect of the leaders of the different groups. We present simulation results showing some interesting emergent effects: (a) even when a high fraction of the population of the two groups are willing to compromise with each other, if the leaders are not willing to compromise, there is high potential of conflict between the two groups, and (b) when a majority of the two population groups are unwilling to compromise, even if the leaders are willing to compromise, there is still a high potential of conflict between the two groups. Our simulation results also show that the two groups can coexist in peace, i.e., there is low potential of conflict, when both the leaders and a large fraction of the population are willing to compromise.

1 Introduction

The end of the cold war has seen a meteoric rise in incidents of ethno-religious violence in Europe, e.g., Yugoslavia, Chechnya. Separate incidents have also occurred elsewhere, e.g., Sudan, Rwanda. Many incidents of ethno-religious violence have been associated with either fall of an authoratarian regime or have been perpetrated by an influential dictator. These incidents underline the importance of the effect of influential political and/or religious leaders in the context of ethno-religious conflicts. Apart from the effect of political or religious elites, other social and economic factors, responsible for ethno-religious conflicts, that are identified by empirical research in sociology and conflict resolution literature are [2]: spatial distribution of the different ethnic groups, territorial claims, history of conflict between the groups, competition for scarce natural resources, and international influences. Computational models for studying/analyzing ethno-religious conflicts can be helpful in designing more effective and efficient policies to anticipate and deal with them. In this paper, we discuss the computational modeling of leader effects in ethno-religious conflicts.

In [1,3], we presented an agent-based computational model of ethno-religious conflicts based on repeated PD games in graphs. The main features of this model

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 3–12, 2010.

are (a) there are two types of agents representing an individual or a collection of individuals from an ethno-religious group and the interaction graph within each group is modeled as a social network and (b) each member of one group interacts with a few members of the other group and the interaction is abstracted as a PD game (this is inspired by the study of Lumsden [4], where he experimentally showed in the context of the Cyprus conflict that the inter-group conflict can be modeled as a PD game). Our model took into account the ethno-religious identity of the population, spatial distribution of the population, and the existing history of animosity. In this paper, we extend our model to take into account the effects of leaders. Many measures of centrality have been developed to measure nodes' influence in a network, such as degree centrality, closeness centrality, betweenness centrality, and centrality based on neighbors' characteristics (prestige-, power, and eigenvector-related centrality) [5]. Based on the definition of nodes' payoffs and strategy updating rule in our model [1], it is natural to use the degree centrality as the measure of agents social influence. Thus, we define the highest degree node in each group as the group's leader node.

We present simulation results showing some interesting emergent effects: (a) even when a high fraction of the population of the two groups are willing to compromise with each other, if the leaders are not willing to compromise, there is high potential of conflict between the two groups, and (b) when a majority of the two population groups are unwilling to compromise, even if the leaders are willing to compromise, there is still a high potential of conflict between the two groups. Our simulation results also show that the two groups can coexist in peace, i.e., there is low potential of conflict, when both the leaders and a large fraction of the population are willing to compromise. On the other hand, as is intuitively obvious, if neither the population, nor the leaders are willing to compromise, our simulation results show that there is high potential of conflict between the two groups. These results may suggest a policy that in order to ensure peace in areas of ethno-religious diversity with history of animosity, it is not only essential that the leaders of the two groups are persuaded to compromise with each other, it is also required that campaigns are undertaken to make sure that the population of the two groups are willing to compromise with each other.

This paper is organized as follows: In Section 2, we present a brief overview of the literature related to computational conflict modeling and the effect of leaders in conflict modeling. In Section 3, we present the notations and definitions that are used in the paper. In Section 4, we present our basic computational model from [1] and in Section 5, we incorporate the leader effects in our model and present simulation results showing the effect of leaders. Finally, in Section 6, we provide our conclusions and outline problems to be addressed in the future.

2 Related Work

There has been substantial empirical research in sociology and conflict resolution literature on analyzing the causes of ethno-religious violence (see [2] and references therein). More recently, there has been focus on computational study

of civil conflict in societies (see [6,7,8] and references therein), some of which have also studied societies with multiple ethno-religious groups [9,7]. Although the proposed computational models vary in detail, the common features of existing agent-based computational models are: (a) the agents are assumed to be distributed on a grid and the interaction between agents are restricted to a neighborhood around their position (usually their Moore-neighborhood, i.e., nearest 8 neighbors) and (b) the interaction between agents is non-adaptive (except in [7]), e.g., agents will die with a certain probability if they are in the neighborhood of opponent agents or they will migrate towards agents of only their type. Since it is well established that the structure of a social network of interacting population is not like a grid [10], in [1], we model the interaction topology among the agents as a graph. Moreover, we model the interaction between agents as a repeated prisoner's dilemma (PD) game where the agents update their strategies. In [11], a model of social matching game with costly monitoring was developed, where although the interaction between two agents is also modeled using PD game, the way of game playing and strategy update are very different and its focus is the relation between inter-ethnic social order and in-group policing. There has also been studies that have specifically modeled the behavior of leaders and interaction between leaders and societies in conflict situations [12,13,14].

3 Preliminaries

Undirected graph: An undirected graph G is an ordered pair, $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ is a set of n nodes, and $E \subseteq V \times V$ is a set of edges. Two nodes v_i and v_j are called *neighbors* of each other if $(v_i, v_j) \in E$. The set $\mathcal{N}_i = \{v_j | (v_i, v_j) \in E\}$ is the set of v_i's neighbors, and $|\mathcal{N}_i|$ is defined as the *degree* of node v_i. Denote $\mathcal{N}_i^+ = \mathcal{N}_i \cup \{v_i\}$.

Scale-free network: A scale-free network is a graph where the degree distribution of nodes follow a power law [10], i.e., $N_d \propto d^{-\gamma}$, where N_d is the number of nodes of degree d and γ is a constant called the power law degree exponent (typically $\gamma \in [2, 3]$).

Prisoner's Dilemma Game: In its simplest form, the prisoner's dilemma game is a single-shot two-player game where the players have two available strategies – cooperate (C) and defect (D). The payoff's of the players is given by

	C	D
C	σ_1, σ_2	a_1, b_2
D	b_1, a_2	δ_1, δ_2

where the index 1 corresponds to the row player and the index 2 to the column player. The entries in the payoff matrix of each player should satisfy $b_i > \sigma_i > \delta_i > a_i, i = 1, 2$. Here, we assume $\sigma_1 = \sigma_2 = \sigma$, $a_1 = a_2 = a$, $\delta_1 = \delta_2 = \delta$, $b_1 = b_2 = b$. For repeated PD games an additional constraint is $2\sigma > a + b$. We further follow the convention in Nowak [15] and set $a = 0$.

PD Game in Graphs: A PD game in a graph is a repeated game where the n-players form the nodes of the graph and the game proceeds in two phases:

(i) game playing phase (ii) strategy update phase. The parameters that define different versions of PD games in graphs are: (a) topology of the graph (fixed or variable) (b) game playing and strategy update neighborhood (c) strategy update rule (d) assumptions on synchronous or asynchronous strategy update. The PD game model that is most relevant to this paper is defined below.

PD game in fixed graphs with synchronized strategy update is a repeated game where each iteration of the game proceeds in the following two phases: (a) In the game playing phase the players play the PD game with all their neighbors with a fixed strategy and compute their total payoff. (b) In the strategy update phase, each player compares the payoffs of all its neighbors (including itself) and chooses the strategy of its neighbor with the highest payoff for the next iteration. In other words, our strategy update rule is: *imitate your best/wealthiest neighbor*. In the game defined above, the agents do not have any group labels, and the game playing and strategy update neighborhoods are identical. As we discuss below, in our model, the agents have different group labels and different game playing and strategy update neighborhoods based on its group label.

4 Problem Model

In this section, we present our agent-based model for studying ethno-religious conflicts [1]. We model the whole multi-cultural population in a geographical region as a collection of agents. An agent represents an individual or a collection of individuals. Since an individual interacts with few other individuals in a society, we model the collection of interacting population as a graph where the nodes are the agents and the edges denote interaction between the agents. We assume that the population consists of two different ethno-religious groups (i.e., there are two different types of nodes in the graph). The interaction between agents in two different groups is modeled as the PD game. In this context, the strategy cooperate (C) implies the willingness of the agent to compromise with the other group whereas the strategy defect (D) implies unwillingness to compromise with the other group. Thus, the fraction of links between the two groups where both agents play D can be used as a measure of tension between the two groups.

Network Construction: We use an undirected graph $G = (V, E)$ to represent the agents of two groups and their connections. We construct G in two steps:

1. Construct the graphs $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$ for each group separately.
2. Construct the set of edges $E_3 \subseteq V_1 \times V_2$ such that each agent in one group is connected to at least one agent in the other group. The edges are added by picking two nodes from the two groups uniformly at random. Let the average number of edges connecting an agent in one group to agents in the other group be k (a parameter capturing connectivity between two groups).

Thus, we get the graph $G = (V, E)$, with $V = V_1 \cup V_2$ and $E = E_1 \cup E_2 \cup E_3$. Figure 1 illustrates the network structure of G. By construction, each agent i has two types of neighborhood in G: (a) neighborhood of agents of same type

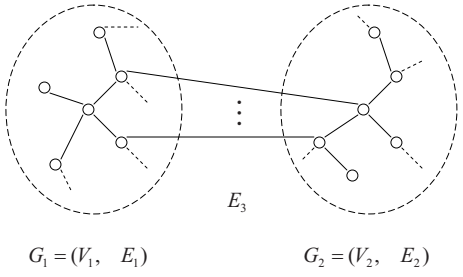


Fig. 1. The network structure G of the model. $G = (V_1 \cup V_2, E_1 \cup E_2 \cup E_3)$. On the left (or right) oval is a sub-graph of network G_1 (or G_2). Edges connecting nodes in the two ovals represent E_3, the set of edges between two groups.

$\mathcal{NS}_i = \{v_j | (v_i, v_j) \in E_1 \cup E_2\}$ and (b) neighborhood of agents of different type $\mathcal{ND}_i = \{v_j | (v_i, v_j) \in E_3\}$.

Phases in each round of the game: The PD game in graphs proceeds in rounds where each round consists of game playing phase and strategy update phase. We assume that the network structure is fixed and the strategy update is synchronous. Each agent plays the PD game with all agents of the other type in its neighborhood, i.e., $\mathcal{ND}_i$ is i's game playing neighborhood. Intuitively, this encodes the fact that we are interested in inter-group disputes and not in disputes within the group. Let $s_i(t)$ denote the strategy of agent i at round t, where $s_i(t) = 0$ implies cooperation (C), and $s_i(t) = 1$ implies defection (D). Each agent plays the same strategy with all agents in its game playing neighborhood. The aggregate payoff, $p_i(t)$, of agent i in iteration t can be computed by summing up the individual payoffs obtained from playing with agents in $\mathcal{ND}_i$.

$$p_i(t) = \sum_{j:v_j \in \mathcal{ND}_i} \Big(\sigma(1 - s_i(t))(1 - s_j(t)) + b s_i(t)(1 - s_j(t)) + \delta s_i(t) s_j(t) \Big) \quad (1)$$

In the strategy update phase, each agent i imitates the strategy of the agent with highest payoff at previous round from a set $\mathcal{C}_i^+ = \mathcal{C}_i \cup \{v_i\}$ where $\mathcal{C}_i$ is the strategy update neighborhood. We choose the strategy update neighborhood of an agent to be the neighborhood containing agents of the same ethno-religious type, i.e., $\mathcal{C}_i = \mathcal{NS}_i$. This encodes the assumption that an agent gives more importance to the opinions of neighbors of its own type regarding the opposite type than the opinion of the other type for its own type. If there is more than one agent with the highest payoff, an agent randomly selects one of the agents and imitate its strategy. Thus the strategy update for agent i can be written as:

$$s_i(t) = s_j(t-1) \text{ where } j = \arg\max_{k \in \mathcal{C}_i^+} (p_k(t-1)) \quad (2)$$

Model parameters: The PD game in graphs that we defined has a number of parameters: (a) The parameters of the payoff matrix σ, δ, a, and b; we set $a = 0$,

$\sigma = 1, \delta = 0.1$ [1], and $1 < b < 2$ represents the prevalent level of animosity between the two groups (e.g., due to historical reasons). (b) The initial fraction of cooperators in the PD game, f_c. (c) The number of agents in each group n_1, n_2, here we assume that $n_1 = n_2$. (d) The topology of the graphs and the average number of edges, k, from one group to another. In Section 5, we presents simulation results obtained by varying the parameters b and f_c. In [1], we showed that changing the number of agents n beyond a certain point does not change the conflict measure, so we set $n = 300$ in the simulations. We present results with $k = 5$. For other values of k the results are qualitatively similar.

Measure of Conflict Potential: The measure of conflict potential between two different groups should consider the interactions between agents in different groups, which can be expressed as the strategy pairs for two neighboring agents belonging to different groups (in the steady state). So we use the fraction of $D-D$ links (two neighboring agents both play defection) between two groups as a measure of the potential of conflict between the two groups, which can be computed as:

$$f_{dd} = \frac{\sum_{(v_i,v_j)\in E_3} s_i \cdot s_j}{|E_3|} \tag{3}$$

5 Effect of Leaders

In the basic model that we presented above, we did not incorporate the fact that different agents in a society may have different levels of influence. In any group of population, the strategies of some agents (leaders of the group) may have more effect on the society than the strategies of others. In this section, we incorporate the effect of leaders in our model. Since we model the population of each ethnic group as a scale-free network, we can use the degree of each node in the graph as a measure of its social influence. Consequently, high degree nodes can be thought of as the agents with high influence on the population. By multiplying the payoffs of the agents at each round by a factor proportional to the degree of their nodes, we can ensure that the strategies of the leader nodes have high payoffs. In other words, we calculate the payoff of each agent, $p'_i(t)$, using

$$p'_i(t) = p_i(t) * |\mathcal{N}_i| \tag{4}$$

where $|\mathcal{N}_i|$ is the degree of node i and $p_i(t)$ is the payoff for node i at round t without considering leaders' effect and is given by Equation 1. The strategy update rule is the same as in the basic model, i.e., *imitate your neighbor with highest payoff*, and is given by Equation 2.

According to our construction of edges between the two ethnic groups, each node in one group will have almost the same number of neighbors in the other group. Therefore, according to Equation 4, leader nodes with high degrees in their own group (i.e., high in-group degree) will have greater probability to get higher payoffs. Consequently, during the strategy update phase, the leaders will have more effect on the strategies of the whole population compared to low

degree nodes. In other words, the willingness of the leaders to compromise or not will play a significant role in the potential of conflict between the two groups.

To study the effect of leaders' strategies on the tendency of conflict between the two groups, we ran simulations with the leader as the highest degree node in each group. For each set of testing parameters, we randomly generated 500 graphs, and compared the average final fraction of *D-D* links for three different combinations of leaders' initial strategies: both leaders cooperate, one cooperate while the other defect, and both defect. In each graph randomly generated for the simulations, G_1 and G_2 are scale-free networks generated using the Barabasi-Albert algorithm [10]. The set of edges E_3 between nodes in V_1 and V_2 are generated randomly and ensure that each node in V_1 is connected to at least one other node in V_2 and the average number of edges between the two groups is k. The total number of iterations for each run was set at 30 and we verified that the system converged to a steady state.

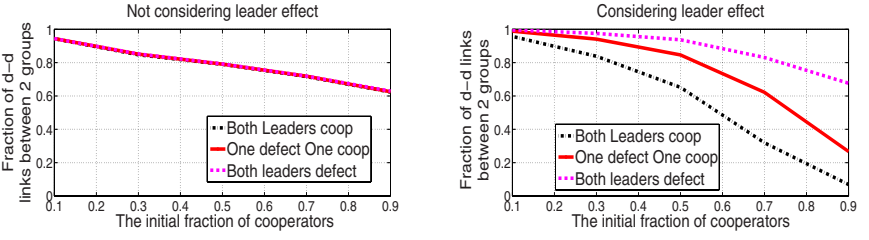


Fig. 2. The plot of f_{dd} as function of f_c using (a)the basic model without Equation 4, (b)the model incorporating leader's effect with Equation 4. The three curves in the fig corresponds to three different combinations of initial strategies of the two leader nodes with the highest degree in each group. $n = 300, b = 1.5, k = 5$.

Figure 2 shows the variation of the final fraction of *D-D* links f_{dd} with the initial fraction of cooperators f_c for different leader strategies (both leaders cooperating, both leaders defecting, and only one of them cooperating). The results in Figures 2 and 4 were generated with $f_c = 0.1, 0.3, 0.5, 0.7, 0.9$, and each data point is an average over 500 runs. Figure 2(a) shows that when we do not use Equation 4 for obtaining the payoff's of the agents, the strategy of the leaders does not affect the conflict measure (all the curves are almost identical), while 2(b) shows when we use Equation 4 for computing the payoffs, the conflict measure is dependent on the strategies of the leaders. A similar effect on the conflict measure for variation of b is shown in Figure 3. From Figure 2(b), we see that by changing the initial strategies of the leader nodes in each group from cooperate to defect, f_{dd} always increases. This is especially important when f_c is high, because this shows that although most of the population may be willing to cooperate, if the leaders are willing to defect then the conflict potential is high. For example, when $f_c = 0.9$, if both leaders cooperate, the average (over the 500

different runs) f_{dd} is about 0.1 while if both defect, f_{dd} is about 0.7. When both f_c is high and the leaders cooperate, the conflict measure is low. Apparently, it seems that when the majority of the population is willing to defect, the initial willingness of the leaders to cooperate does not have any significant effect (look at Figure 2(b) for $f_c = 0.1$). However, we note that apart from the leader, there are other high degree nodes present in each group. These influential agents may have less influence than the leader node, individually, but as a collective they may have more influence and may force the leaders to change their strategies.

To test this intuition, we also checked the fraction of leaders that changed their initial strategies. From Figure 4 it is apparent that when f_c is low, the leaders initially cooperating almost always change their strategies to defect. For example, when $f_c = 0.1$, leaders initially cooperating have a fraction of strategy switching as high as almost 1.0. On the other hand, for leaders initially defecting, the fraction of leaders switching strategy is uniformly lower than 0.03. Although the results in Figure 4 are for a constant $b = 1.5$, the nature of the curves remain same for other values of b, and there is no qualitative difference.

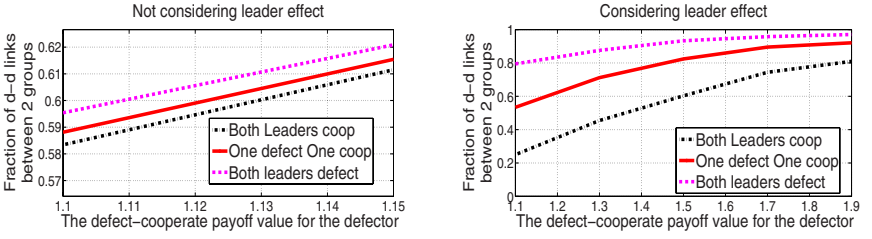


Fig. 3. The plot of f_{dd} as function of b using (a)the basic model without Equation 4, (b)the model incorporating leader's effect with Equation 4. $n = 300$, $f_c = 0.5, k = 5$.

In Figures 2 and 4, we have kept the value of the incentive to defect as a constant, i.e., $b = 1.5$. Now, we study the effect of varying b while keeping f_c constant. The results in Figures 3 and 5 were generated with $b = 1.1, 1.3, 1.5, 1.7, 1.9$, and each data point is an average over 500 runs. From Figure 3(b), we can see that irrespective of the value of b, there is more conflict potential if the leaders defect instead of cooperate. Moreover, even when b is low ($b = 1.1$), if the leaders initially defect, the potential of conflict is quite high ($f_{dd} = 0.8$). For high incentive to defect ($b = 1.9$), the potential of conflict is high even if the leaders are initially willing to cooperate ($f_{dd} = 0.8$). Different from before, this is not only due to the fact that the leaders switch strategies from cooperate to defect, since other influential leaders in the society are defecting (see Figure 5), but also due to the fact that the incentive of agents in both groups to defect is high.

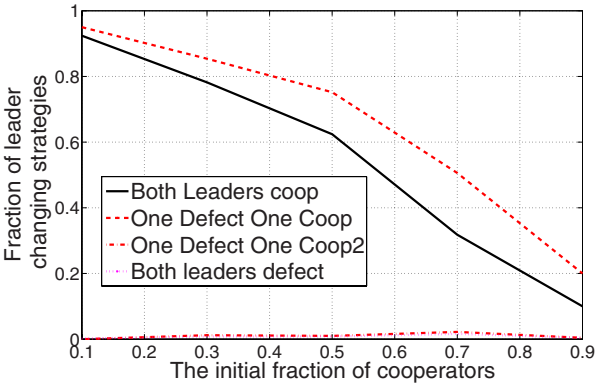


Fig. 4. Variation of fraction of cases where leaders switch their initial strategies to different final ones as a function of f_c; The four curves in the figure corresponds to three different combinations of initial strategies of the two leader nodes, where we distinguish the case when one leader cooperates and the other defects. $n = 300, b = 1.5, k = 5$.

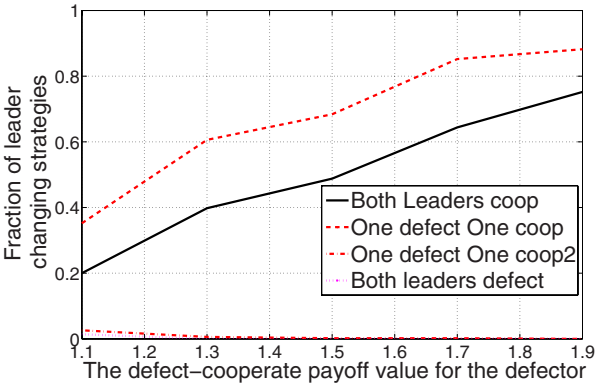


Fig. 5. Variation of fraction of cases where leaders switch their initial strategies to different final ones as a function of b. $n = 300, f_c = 0.5, k = 5$.

6 Conclusion

In this paper, we extended the agent-based computational model of ethno-religious conflict between two groups, proposed in [1], to incorporate the effect of leaders of the different groups. We presented simulation results showing the effect of various parameters of our model to the propensity of conflict in a population consisting of two ethno-religious groups. From the simulation results, we can see that the leader's effect is reflected only after the extension to our previous computational model. We also found some interesting emergent effects with our model, which showed that if either a high fraction of the population or the leaders are not willing to comprise, there is a high potential of conflict between the two groups. One relevant feature that is missing from our current model is the effect of the geographical distribution of economic and natural resources. In our future work, we plan to extend our model to take into consideration this effect. We also plan to validate our model using Sudan as a case study.

Acknowledgments

This research was partially funded by ONR MURI grant N000140811186 and by ARO MURI grant W911-NF-0810301.

References

1. Luo, L., Chakraborty, N., Sycara, K.: Modeling ethno-religious conflicts as prisoner's dilemma game in graphs. In: IEEE International Conference on Computational Science and Engineering, vol. 4, pp. 442–449 (2009)
2. Toft, M.D.: The Geography of Ethnic Violence: Identity, Interests, and the Indivisibility of Terrirtory. Princeton University Press, Princeton (2006)
3. Luo, L., Chakraborty, N., Sycara, K.: Prisoner's dilemma on graphs with heterogeneous agents. In: GECCO 2009: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference, pp. 2097–2102 (2009)
4. Lumsden, M.: The cyprus conflict as a prisoner's dilemma game. Journal of Conflict Resolution 17(1), 7–32 (1973)
5. Jackson, M.O.: Representing and Measuring Networks. In: Social and economic networks, pp. 37–43. Princeton University Press, Princeton (2008)
6. Epstein, J.: Modeling civil violence: An agent-based computational approach. PNAS 99(Suppl. 3), 7243–7250 (2002)
7. Goh, C.K., Quek, H.Y., Tan, K.C., Abbass, H.A.: Modeling civil violence: An evolutionary multi-agent, game theoretic approach. In: IEEE Congress on Evolutionary Computation, pp. 1624–1631 (2006)
8. Srbljinovic, A., Penzar, D., Rodik, P., Kardov, K.: An agent based model of ethnic mobilisation. Journal of Artificial Societies and Social Simulation 6(1) (2003)
9. Lim, M., Metzler, R., Bar-Yam, Y.: Global Pattern Formation and Ethnic/Cultural Violence. Science 317(5844), 1540–1544 (2007)
10. Barabasi, L.-A., Albert, R.: Emergence of Scaling in Random Networks. Science 286(5439), 509–512 (1999)
11. Nakao, K.: Creation of social order in ethnic conflict. Journal of theoretical politics 21(3), 365–394 (2009)
12. Bernard, M., Backus, G.: Modeling the interaction between leaders and society during conflict situations. In: 27th Intl. System Dynamics Conference (2009)
13. Phillips, C.L., Sokoloff, S.K., Crosscope, J.R., Geddes, N.D.: A validation process for predicting stratagemical behavior patterns of powerful leaders in conflict. In: Social Computing and Behavioral Modeling, pp. 155–162. Springer, US (2009)
14. Silverman, B.G., Bharathy, G., Nye, B., Eidelson, R.J.: Modeling factions for "effects based operations": part I–leaders and followers. Comput. Math. Organ. Theory 13(4), 379–406 (2007)
15. Nowak, M., Sigmund, K.: Game-dynamical aspects of the prisoner's dilemma. Appl. Math. Comput. 30(3), 191–213 (1989)

Calibrating Subjective Probabilities Using Hierarchical Bayesian Models

Edgar C. Merkle

Department of Psychology,
Wichita State University,
Wichita, KS 67260-0034
edgar.merkle@wichita.edu
http://psychology.wichita.edu/merkle

Abstract. A body of psychological research has examined the correspondence between a judge's subjective probability of an event's outcome and the event's actual outcome. The research generally shows that subjective probabilities are noisy and do not match the "true" probabilities. However, subjective probabilities are still useful for forecasting purposes if they bear some relationship to true probabilities. The purpose of the current research is to exploit relationships between subjective probabilities and outcomes to create improved, model-based probabilities for forecasting. Once the model has been trained in situations where the outcome is known, it can then be used in forecasting situations where the outcome is unknown. These concepts are demonstrated using experimental psychology data, and potential applications are discussed.

Keywords: Subjective probability, confidence, Bayesian methods, calibration, expert judgment.

1 Introduction

Subjective probability is commonly used to measure judges' certainties in decisions and forecasts. People are generally familiar with reporting such probabilities, making them a natural way to gauge certainty in many situations. This has led to a long line of psychology research devoted to understanding how individuals construct subjective probabilities, where it is often found that subjective probabilities tend to be larger than the true probabilities of the corresponding outcomes (e.g., [1,2,3]).

The intent of this paper is to study the use of a hierarchical logistic model for improving subjective probabilities. The model transforms individual subjective probabilities into predicted probabilities of an outcome's occurrence. Once the model has been fit to data with known outcomes, the model can be used to transform subjective probabilities and forecast unknown outcomes. A particularly-interesting aspect of the model is that it yields unique transformations for individual judges, accounting for individual differences in response styles while maintaining general trends present in the group of judges.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 13–22, 2010.

The model is related to the vast literature on combining expert judgments (see, e.g., [4,5]), where the goal is to take many subjective judgments as input and yield a single, aggregated prediction as output. The current paper differs from this literature in that it examines how *individuals'* subjective probabilities are related to true probabilities of particular outcomes. Thus, the model in the current paper may be used to improve individual expert probabilities prior to aggregating the probabilities (an idea advanced by [6]). A special case of the model may also be used in situations where only a single expert reports a probability.

In the pages below, I first define measures of the correspondence between subjective probabilities and outcomes, along with the measures' use in applications. I then define the model that is used to transform subjective probabilities. Next, I demonstrate the utility of the approach using data from a visual discrimination experiment. Finally, I describe how the model can be used in applications and consider other statistical methods that could be relevant.

1.1 Correspondence between Subjective Probability and Outcomes

Researchers have defined many measures of the correspondence between probabilistic forecasts and outcomes. One of the most intuitive measures is of the extent to which probabilistic forecasts match the long-term proportion of occurring outcomes. This can be defined mathematically as a measure of *bias*. Let $d_j \in \{0, 1\}$ be the outcome of event j ($j = 1, \ldots, J$),[1] and let f_j be a judge's subjective probability that $d_j = 1$. For the purposes of this paper, bias is then defined as:

$$\text{bias} = \overline{f} - \overline{d}, \tag{1}$$

where $\overline{f}$ is the mean of the f_j and $\overline{d}$ is the mean of the d_j ($j = 1, \ldots, J$). Biases close to zero reflect "good" forecasts, and biases far from zero reflect "bad" forecasts.

There exist a variety of other measures designed to examine other aspects of the correspondence between the f_j and the d_j; see [7]. Two of these measures are *slope* and *scatter*. Slope measures the extent to which forecasts differ for $d_j = 0$ and $d_j = 1$:

$$\text{slope} = \overline{f}_1 - \overline{f}_0, \tag{2}$$

where $\overline{f}_1$ is average subjective probability for events where $d_j = 1$ and $\overline{f}_0$ is average subjective probability for events where $d_j = 0$. Large slopes reflect good forecasts, and small slopes reflect bad forecasts.

Scatter reflects noise in the f_j that is unrelated to the d_j:

$$\text{scatter} = \frac{(n_1 - 1)s^2_{f_1} + (n_0 - 1)s^2_{f_0}}{n_1 + n_0 - 2}, \tag{3}$$

[1] *Outcome* has multiple meanings. It could refer to whether or not an event occurs, in which case we have 0=event does not occur, 1=event does occur. Alternatively, *outcome* could refer to whether or not a judge's prediction of an event's occurrence matches the event's actual occurrence. In this case, we have 0=judge's prediction was incorrect, 1=judge's prediction was correct.

where $s^2_{f_1}$ is the variance of the f_j for which $d_j = 1$, n_1 is the number of events for which $d_j = 1$, and $s^2_{f_0}$ and n_0 are defined similarly. Small values of scatter reflect good forecasts, and larger values reflect bad forecasts.

1.2 Use of the Measures

Decision researchers have tended to focus on the bias measure: bias is generally intuitive, and observed bias may be immediately compared to the "perfect" bias value of 0. In a variety of experimental tasks, researchers tend to find biases greater than zero; that is, judges' subjective probabilities tend to be larger than they should [7,8,9,10]. This implies that, in applied situations, subjective probabilities are suboptimal for guiding decisions. Less research has focused on measures other than bias (though see, e.g., [6,11]), which may be because it is generally impossible for judges to attain perfect values on these other measures. Thus, it is difficult to say whether a particular value of slope or scatter is good. It is still possible to compare relative magnitudes of slope and scatter, making them useful for comparing observed slope and scatter with model-predicted slope and scatter. These measures are used in the example that follows, but I first describe the specific model that is used to transform the subjective probabilities.

2 Model

Let i index judges and j index forecasts. To transform subjective probabilities, I consider a hierarchical logistic model with f_{ij} as a predictor variable and d_{ij} as a response variable. The basic model is given as:

$$\begin{aligned} d_{ij} &\sim \text{Bernoulli}(p_{ij}) \\ \log(p_{ij}/(1-p_{ij})) &= b_{0i} + b_{1i} f_{ij}, \end{aligned} \tag{4}$$

where p_{ij} is the probability that judge i is correct on forecast j. This probability is modeled using the judge's subjective probability, f_{ij}, as a predictor. The slope and intercept in the model vary for each judge i, allowing the relationship between p and f to differ from judge to judge. The hierarchical formulation of the model is obtained by assuming a joint normal distribution on the b_{0i} and b_{1i}:

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \sim \text{N}\left[\begin{pmatrix} B_0 \\ B_1 \end{pmatrix}, \mathbf{\Sigma}_b = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix} \right], \tag{5}$$

where B_0 is the mean of the intercepts, B_1 is the mean of the slopes, and $\mathbf{\Sigma}_b$ is the covariance matrix of the intercepts and slopes. The hierarchical normal distribution is traditionally used in this model, but a different distribution could be used if deemed useful or necessary. This flexibility in hierarchical distributions is an advantage of the Bayesian approach.

There are two other Bayesian advantages that led to the implementation of a Bayesian model here. First, the Bayesian model allows for incorporation of prior

knowledge about base rates of correct forecasts or about relationships between p and f. This could be useful for specific applications. Second, the Bayesian model allows for calculation of posterior predictive distributions of the p_{ij}. As will be shown below, this allows us to systematically transform judges' reported probabilities into conservative and/or liberal probabilistic predictions. To complete the Bayesian model, we require prior distributions on B_0, B_1, and the associated covariance matrix $\mathbf{\Sigma}_b$. These are given as:

$$B_0 \sim \text{N}(\mu_0, s_0^2) \tag{6}$$
$$B_1 \sim \text{N}(\mu_1, s_1^2) \tag{7}$$
$$\mathbf{\Sigma}_b \sim \text{Inv-Wishart}(\text{df}, \mathbf{\Sigma}_0), \tag{8}$$

where $\mathbf{\Sigma}_b$ follows an inverse Wishart distribution with $\text{df} > 1$ and scale matrix $\mathbf{\Sigma}_0$. These parameters can be set based on prior knowledge about the forecasting scenario, or they can be set to reflect the absence of prior knowledge. I consider the latter situation in the following example.

3 Example: Visual Discrimination

To demonstrate the potential applicability of these hierarchical models, I consider data from Experiment 1 of [12]. In this experiment, judges viewed images of asterisks randomly arranged in a 10×10 array. The number of asterisks was randomly drawn from one of two normal distributions, with the first distribution being $\text{N}(45, \sigma = 5)$ and the second being $\text{N}(55, \sigma = 5)$. For each of 450 trials, judges viewed an array and stated a probability that the asterisks arose from the second distribution. Judges were not told the distributions governing number of asterisks; they were required to learn the distributions by themselves. Reported probabilities were required to come from the set $\{.05, .15, .25, \ldots, .95\}$. Choices were inferred from the reported probabilities, and probabilities in the choices were then obtained (ranging from .55 to .95).

3.1 Model Details

The Bayesian hierarchical model described in the previous section was fit to data from 36 subjects across 40 experimental trials. The prior distributions on model parameters were taken to be noninformative:

$$\mu_0 \sim \text{N}(1, 1.0\text{E}5) \tag{9}$$
$$\mu_1 \sim \text{N}(0, 1.0\text{E}5) \tag{10}$$
$$\mathbf{\Sigma}_b \sim \text{Inv-Wishart}(2, \mathbf{I}), \tag{11}$$

where $\mathbf{I}$ is a 2×2 identity matrix. The model was estimated in OpenBugs [13] via Markov chain Monte Carlo methods, with relevant code appearing in the appendix. Three chains of parameters were sampled for $14,000$ iterations each, with the first $4,000$ iterations being discarded as burn-in.

OpenBugs was also used to obtain predicted probabilities for all judges across 20 trials that were not used during model fitting. To be specific, OpenBugs was used to sample from the posterior predictive distributions for these 20 trials. These distributions can be used to obtain conservative and/or liberal predicted probabilities. The extent to which this is useful is examined below.

3.2 Results

Results are presented in two parts. First, I make some general remarks about the fitted model and the probabilistic predictions. I then make detailed comparisons between the predicted probabilities and judges' reported probabilities.

Fitted Model. Before examining the predicted probabilities, a fundamental issue involves the extent to which reported probabilities (f_{ij}) are related to accuracy (d_{ij}). Within the model (Equation (4)), the hierarchical distribution on the b_{1i} informs this issue. This distribution is estimated as $\mathrm{N}(3.1, \widehat{\sigma_1^2} = 0.96)$, with a 95% posterior interval for the mean being $(2.24, 4.09)$. Because the interval is far from zero, we have evidence that the f_{ij} are indeed useful for predicting accuracy. Further evidence comes from the estimated b_{1i} for each judge. All 36 of these estimates are positive, with no 95% posterior intervals including zero.

Now that a predictive relationship between the f_{ij} and d_{ij} has been established, we can examine the extent to which the model's probabilistic predictions are an improvement over the f_{ij}.

Probabilistic Predictions. In this section, the model's probabilistic accuracy predictions, $\widehat{p_{ij}}$, are compared to the f_{ij} for the 20 trials that were excluded from model estimation. Figure 1 displays the observed f_{ij} versus the $\widehat{p_{ij}}$ for all 36 judges. The diagonal line in the graph is the identity line, reflecting instances where $f_{ij} = \widehat{p_{ij}}$. Considerable variability is observed in the mapping from f_{ij} to $\widehat{p_{ij}}$ for different judges. Further, the f_{ij} and $\widehat{p_{ij}}$ differ primarily for large f_{ij}: in these cases, the $\widehat{p_{ij}}$ are smaller. Thus, the model compresses the range of probabilities.

As stated previously, the model's posterior predictive distributions of $\widehat{p_{ij}}$ were obtained for the 20 trials excluded from model estimation. We can summarize these distributions in various ways to obtain predictive probabilities. A common summary involves taking the means of the posterior distributions. Alternatively, if we want more conservative predictive probabilities, we can take the 25^{th} percentile of these distributions, for example. Slope, scatter, and bias statistics were calculated for these two types of posterior summaries, along with statistics for the observed f_{ij}. These serve as measures of the extent to which the model predictions are improvements over the f_{ij}.

Figure 2 contains histograms of the difference between each judge's observed statistics and model-predicted statistics (using the means of the posterior predictive distributions). Values greater than zero reflect instances where a judge's observed statistic is greater than his/her model-predicted statistic. While there is judge variability, the graphs generally show that the model tends to yield

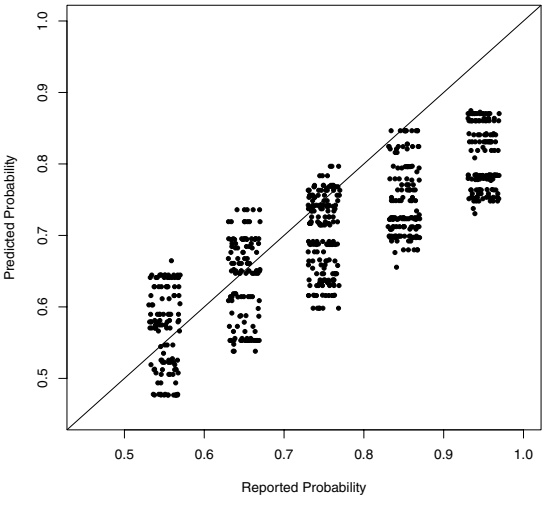


Fig. 1. Model mappings from reported probabilities to predicted probabilities. Points are jittered horizontally to reduce overlap.

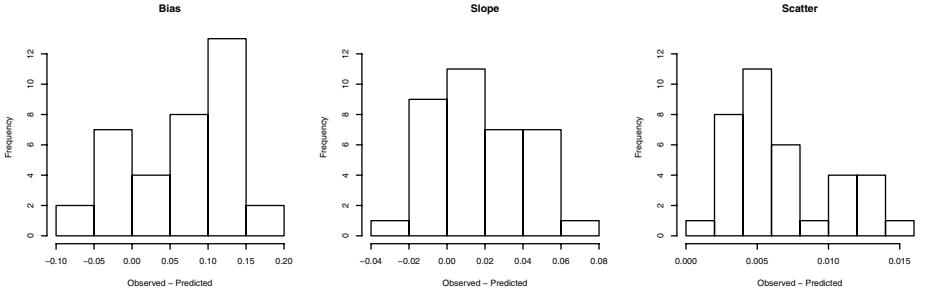


Fig. 2. Differences between observed probabilities and model predictions for each judge on the measures of bias, slope, and scatter

smaller bias and scatter statistics. These trends are supported statistically: 95% confidence intervals for the mean difference in bias and scatter are $(.039, .084)$ and $(.005, .007)$, respectively. While these are positive results for the model, the slope statistics reflect a negative result: the observed slopes tend to be larger than the model-predicted slopes, with the 95% confidence interval for the mean difference being $(.008, .022)$. I address this negative result in more detail below.

For model predictions using the 25^{th} posterior percentiles, results are similar: the predictions yield reductions in both bias and scatter, but they do not yield increases in slope. Comparing the two types of predictions (25^{th} percentile predictions and mean predictions), both slope and scatter are virtually the same, with mean differences of .001 and .0003, respectively. The conservatism of the 25^{th} percentile predictions is reflected in the bias statistic. Mean bias for the 25^{th} percentiles is $-.036$ and mean bias for the means is .006, with a 95% confidence for

interval for the mean difference being $(-.043, -.041)$. One may argue that the conservative predictions are too conservative, as the mean predictions display near-perfect bias statistics.

3.3 Discussion

The hierarchical logistic model transformed judges' subjective probabilities into predicted probabilities that were better calibrated (i.e., bias closer to zero) and contained less noise (i.e., reduced scatter). Importantly, the predictions were made on trials that were excluded from the model estimation. Thus, fitted models of this type can be used to predict probabilities of unknown outcomes, an attribute that is important for applications. More details appear in the General Discussion.

While the model improved bias and scatter, it did not improve slope. Stated differently, the model predictions were unable to better discriminate between correct and incorrect outcomes. This is partly due to the fact that the observed range of model predictions is smaller than the observed range of the f_{ij} (as shown in Figure 1). The result is also impacted by the fact that the $\widehat{p_{ij}}$ are increasing functions of the f_{ij}. This implies that the $\widehat{p_{ij}}$ follow the same ordering as the f_{ij}, which does not leave much room for improvement in slope. In the General Discussion, other statistical methods are considered that may improve slope.

4 General Discussion

Model-based corrections to subjective probabilities, such as those described in this paper, have the potential to be useful in many applied situations. Further, there exist many other statistical methods/models that can produce probabilistic predictions and that may yield improvements in slope. Both of these topics are considered below.

4.1 Applications

The model considered in this paper is applicable to situations where judges report many subjective probabilities in response to stimuli from the same domain, such as medical diagnoses and aircraft inspections. Focusing on the latter application, inspectors examine many areas of the aircraft and report the existence of defects. These reports are often made with considerable uncertainty, especially in nondestructive testing situations (e.g., [14,15]). For example, to check for cracks in bolt holes, inspectors sometimes rely on eddy current technology. An electrical charge is sent through the material around the bolt hole, and inspectors rely on a digital monitor to diagnose cracks. This occurs across a large number of bolt holes on the aircraft.

If inspectors report probabilities of cracks in each bolt hole, the hierarchical logistic model can be used to improve the reported probabilities of individual inspectors. In such a scenario, inspectors may first complete test inspections where the existence of a crack is known. The model can then be fit to these inspections, and the fitted model used to improve reported probabilities for cases where the existence of a crack is unknown.

4.2 Other Statistical Methods

The primary disadvantage of the hierarchical logistic model is that it fails to yield improvements in slope. This is likely to be a problem with any statistical model whose predictions are a linear function of the f_{ij}, because the ordering among the predictions will be the same as the ordering among the f_{ij}. As a result, it may be useful to study models or algorithms that utilize nonlinear functions of f_{ij}. There are at least two classes of methods that one may consider: statistical learning algorithms (e.g., [16]) and psychological/psychometric models of subjective judgment (e.g., [11,12,17,18,19]).

The main focus of statistical learning algorithms, such as boosting, is prediction. The algorithms can make predictions that are nonlinear functions of the inputs, meaning that they may more easily yield improvements in the slope measure (as opposed to the logistic model). A possible problem with the use of these algorithms is lack of data: the algorithms are often suited to data containing thousands of observations and hundreds of predictor variables. The applications considered here may contain hundreds of observations and two predictor variables (subjective probability, judge who reported the probability). In such cases, it is unclear whether the algorithms will result in improvements over more traditional statistical models.

Psychological models may also be used to transform subjective probabilities. These models often posit psychological processes contributing to the construction of subjective probabilities. As a result, the models often treat subjective probability as a response variable instead of a predictor variable. This may make it difficult to use subjective probabilities to predict accuracy. On the other hand, the psychological models may be modified to obtain distributions of accuracy conditioned on subjective probability. If the models truly describe psychological processes contributing to subjective probability, then their accuracy predictions may be better than the more general models/algorithms described earlier. In any case, Bayesian model formulations and Markov chain Monte Carlo are likely to be useful tools for studying these psychological models.

References

1. Dawes, R.M.: Confidence in intellectual vs. confidence in perceptual judgments. In: Lantermann, E.D., Feger, H. (eds.) Similarity and choice: Papers in honor of Clyde Coombs, pp. 327–345. Han Huber, Bern (1980)
2. Lichtenstein, S., Fischhoff, B.: Do those who know more also know more about how much they know? The calibration of probability judgments. Organizational Behavior and Human Performance 20, 159–183 (1977)
3. Wallsten, T.S., Budescu, D.V.: Encoding subjective probabilities: A psychological and psychometric review. Management Science 29, 152–173 (1983)
4. Cooke, R.M.: Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press, New York (1991)
5. O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T.: Uncertain judgements: Eliciting experts' probabilities. Wiley, Hoboken (2006)

6. Wallsten, T.S., Budescu, D.V., Erev, I., Diederich, A.: Evaluating and combining subjective probability estimates. Journal of Behavioral Decision Making 10, 243–268 (1997)
7. Yates, J.F., Curley, S.P.: Conditional distribution analyses of probabilistic forecasts. Journal of Forecasting 4, 61–73 (1985)
8. Keren, G.: On the ability of monitoring non-veridical perceptions and uncertain knowledge: Some calibration studies. Acta Psychologica 67, 95–119 (1988)
9. Lichtenstein, S., Fischhoff, B., Phillips, L.D.: Calibration of probabilities: The state of the art to 1980. In: Kahneman, D., Slovic, P., Tversky, A. (eds.) Judgment under uncertainty: Heuristics and biases, pp. 306–334. Cambridge University Press, Cambridge (1982)
10. Price, P.C.: Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. Organizational Behavior and Human Decision Processes 76, 277–297 (1998)
11. Dougherty, M.R.P.: Integration of the ecological and error models of overconfidence using a multiple-trace memory model. Journal of Experimental Psychology: General 130, 579–599 (2001)
12. Merkle, E.C., Van Zandt, T.: An application of the Poisson race model to confidence calibration. Journal of Experimental Psychology: General 135, 391–408 (2006)
13. Thomas, A., O'Hara, B., Ligges, U., Sturtz, S.: Making BUGS open. R News 6, 12–17 (2006)
14. Swets, J.A.: Assessment of NDT systems–Part I: The relationship of true and false detections. Materials Evaluation 41, 1294–1298 (1983)
15. Swets, J.A.: Assessment of NDT systems–Part II: Indices of performance. Materials Evaluation 41, 1299–1303 (1983)
16. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Springer, New York (2001)
17. Batchelder, W.H., Romney, A.K.: Test theory without an answer key. Psychometrika 53, 71–92 (1988)
18. Erev, I., Wallsten, T.S., Budescu, D.V.: Simultaneous over- and underconfidence: The role of error in judgment processes. Psychological Review 101, 519–527 (1994)
19. Ratcliff, R., Starns, J.: Modeling confidence and response time in recognition memory. Psychological Review 116, 59–83 (2009)
20. Gelman, A., Hill, J.: Data analysis using regression and multilevel/hierarchical models. Cambridge, New York (2007)

Appendix: OpenBugs Code for the Hierarchical Logistic Model

The code below takes a 36×40 matrix of accuracy data (0=incorrect, 1=correct) and a 36×60 matrix of confidence data, where rows reflect judges and columns reflect items. It simultaneously fits the hierarchical logistic model to 40 trials of data from each judge and yields posterior predictions for the final 20 columns in `corr`. The data file (not shown) contains the accuracy data matrix (`corr`), the confidence data matrix (`conf`), and a 2×2 identity matrix (`Iden`). More details on Bayesian hierarchical logistic models is found in, e.g., [20].

```
model{
  for (i in 1:36){
    for (j in 1:40){
      corr[i,j] ~ dbern(p[i,j])

      logit(p[i,j]) <- b[i,1] + b[i,2]*conf[i,j]
    }
    # Hierarchical distribution on bs
    b[i,1:2] ~ dmnorm(mu.b[], invS[,])
  }

  # Posterior predictions for new confidence judgments
  for (i in 1:36){
    for (j in 41:60){
      newp[i,(j-40)] <- b[i,1] + b[i,2]*conf[i,j]
    }
  }

  # Priors
  mu.b[1] <- dnorm(0,1.0E-5)
  mu.b[2] <- dnorm(0,1.0E-5)
  invS[1:2,1:2] ~ dwish(Iden[,],2)
}
```

State-Dependent Risk Preferences in Evolutionary Games

Patrick Roos[1,2] and Dana Nau[1,3,2]

[1] Department of Computer Science
[2] Institute for Advanced Computer Studies
[3] Institute for Systems Research
University of Maryland, College Park MD 20742, USA
{roos,nau}@cs.umd.edu

Abstract. There is much empirical evidence that human decision-making under risk does not correspond the decision-theoretic notion of "rational" decision making, namely to make choices that maximize the expected value. An open question is how such behavior could have arisen evolutionarily. We believe that the answer to this question lies, at least in part, in the interplay between risk-taking and sequentiality of choice in evolutionary environments.

We provide analytical and simulation results for evolutionary game environments where sequential decisions are made between risky and safe choices. Our results show there are evolutionary games in which agents with *state-dependent* risk preferences (i.e., agents that are sometimes risk-prone and sometimes risk-averse depending on the outcomes of their previous decisions) can outperform agents that make decisions solely based on the local expected values of the outcomes.

1 Introduction

Empirical evidence of human decision making under risk shows that humans are sometimes risk averse, sometimes risk seeking, and even behave in ways that systematically violate the axioms of expected utility [1]. Researchers have invested much effort into constructing utility functions that appropriately model human decision making under risk (e.g. [2,3,4]). Researchers have also constructed alternative descriptive theories of decision making that claim to correspond more closely to how humans make decisions involving risk, such as prospect theory [1,5], regret theory [6], and SP/A (Security-Potential/Aspiration) theory [7,8,9].

A question that has received much less attention is how behaviors corresponding to the above decision-making models, or any other empirically documented risk-related behavior that differs from expected value maximization, could have arisen or been learned in societies. We believe that one part of the answer to this question is the interplay between risk-taking and *sequentiality of choices*; and in this paper we present analytical and simulation results to support this hypothesis.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 23–31, 2010.

Our results demonstrate that depending on the game's reproduction mechanism, an agent that acts solely according to the local expected values of outcomes can be outperformed by an agent whose risk preference depends on the success or failure of its previous choices.

2 Evolutionary Lottery Games

We now describe a class of evolutionary games based on a finite, homogeneous population model in which agents acquire payoffs dispensed by lotteries. In each generation, each agent must make a sequence of n choices, where each choice is between two lotteries with equal expected value but different risks. One lottery has a certain outcome of payoff 4 (with probability 1), we call this the *safe* lottery. The other lottery gives a payoff of 0 with probability 0.5 and a payoff of 8 with probability 0.5, we call this the *risky* lottery. Both lotteries have an expected value of 4, the only difference is the payoff distribution.

Within this class, we can define different games by varying two important game features, both of which are discussed below: the number n of choices in the sequence, and the reproduction dynamics.

2.1 Number of Choices

We consider two cases: $n = 1$, i.e., at each generation the agents make a single, one-shot choice among the two lotteries; and $n = 2$, i.e., at each generation the agents make two sequential choices (i.e., $n = 2$).

When $n = 1$ there are two possible pure strategies, as shown in Table 1. When $n = 2$, there are six possible pure strategies, as shown in Table 2.

Table 1. All of the possible pure strategies when $n = 1$

Strategy	Choice
S	choose the safe lottery
R	choose the risky lottery

Table 2. All of the possible pure strategies when $n = 2$

Strategy	1st lottery	2nd lottery
SS	choose safe	choose safe
RR	choose risky	choose risky
SR	choose safe	choose risky
RS	choose risky	choose safe
R-WS	choose risky	choose safe if 1st lottery was won, risky otherwise
R-WR	choose risky	choose risky if 1st lottery was won, safe otherwise

2.2 Reproduction Dynamics

Our evolutionary model uses non-overlapping populations of agents. Once all lottery choices have been made and payoffs have been dispensed, all agents reproduce into the next generation (a new population). Reproduction does not necessarily mean biological reproduction, but can also be a treated as a model for the process of learning [10] or the social spread and adoption of cultural memes or behavioral traits [11], e.g. [12]. We consider two different variants of our games, using two widely used reproduction mechanisms: the replicator dynamic and an imitation dynamic.

The *replicator dynamic*, originating from biology, is the most widely used reproduction mechanism in the literature on evolutionary game theory. The payoffs received by agents are considered to be a measure of the agent's fitness, and agent types reproduce proportional to these payoffs [13,14]:

$$p^{new} = p^{curr} pay(agent_i)/\overline{pay} \tag{1}$$

where p^{curr} is the proportion of agents of type i in the current population, p^{new} is the corresponding proportion in the next generation, $pay(agent_i)$ is the average payoff an agent of type i received from all games played, and $\overline{pay}$ is the average payoff received by all agents in the population. An agent's type is simply the strategy it employs to make choices among lotteries.

Imitation dynamics are probably the second most widely used kind of reproduction mechanism, and are arguably more appropriate in modeling reproduction of strategies in the context of games played in societies [13]. We use the imitation process commonly referred to as *tournament selection* [15,16,17]. Here, each agent in the population is matched up with a randomly drawn other agent in the population and the agent with the higher acquired payoff is reproduced into the next generation. If the payoffs of the matched agents is equal, one of the two agents is chosen at random to reproduce.

3 Analytical Results

We now analyze how well the various strategies should perform under all four combinations of the following parameters: the number of sequential choices ($n = 1$ or $n = 2$), and the reproduction mechanism (imitation or replicator dynamic).

3.1 Case $n = 1$

Recall that for $n = 1$ (i.e., the single choice game) there are only two pure strategies, S and R. S will always receive a payoff of 4, while R will have a 50% chance to receive a payoff of 8 and a 50% chance to receive 0. Hence in each case, the expected value is 4. Thus under the replicator dynamic, by equation (1) we expect neither type of agent to have an advantage. Under the imitation dynamic, an R agent will have a 50% chance to beat an S agent and a 50% to lose, thus we expect neither agent to have an advantage here either.

Table 3. Payoff distributions for all agent types in the sequential lottery game

agent	*R-WS*	*R-WR*	*SR*	*RS*	*SS*	*RR*
payoff	**12 8 0**	**16 8 4**	**12 4**	**12 4**	**8**	**16 8 0**
probability	.5 .25 .25	.25 .25 .5	.5 .5	.5 .5	1	.25 .5 .25

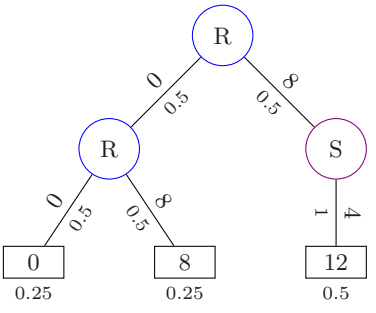


Fig. 1. Sequential lottery tree illustrating payoff distribution achieved by the *R-WS* strategy. Nodes represent lotteries (R for risky, S for safe). Edges are labeled with the payoff dispensed and the associated probability. Nodes are labeled with the final accumulated payoff and the probability for each.

3.2 Case $n = 2$

The situation is more complicated when $n = 2$. Recall from Table 2 that in this case there are six pure strategies. Table 3 gives, for each strategy, its possible numeric payoffs, and the probabilities of these payoffs. We can see by Figure 1 that the *R-WS* agent has a 50% chance of acquiring a payoff of 12, a 25% chance of acquiring a payoff of 8, and a 25% chance of acquiring 0.

Under the imitation dynamic, *R-WS* has an advantage over the other strategies because it has an increased probability of achieving a payoff at or above a certain reproduction threshold. This threshold is the payoff of a randomly drawn opponent, which has an expected value of 8 equal to the expected value of the lotteries. *R-WS* pays for this enlarged chance of being above the threshold through a small chance of doing much worse (payoff 0) than the summed expected values, which occurs when the first and the second risky choice is lost.

The replicator dynamic defines reproduction to be directly proportional to the amount by which the agent's payoff deviates from the population average. In this case the small chance of *R-WS* of being significantly below the expected value balances against the agent's larger chance of being slightly above it. Thus, under the replicator dynamic, the *R-WS* agents have no advantage. All six strategies have an expected value of 4 at each lottery choice, thus a total expected value of 8 for the sequence of two choices. Consequently, we would expect all six strategies to do equally well when using the replicator dynamic.

Since the imitation dynamic only considers whether or not the agent's payoff is better than another agent's in order to decide whether the agent reproduces, the *extent* to which the agent is better is not significant.

If we compare the payoff distribution of *SR* and *RS* with that of *R-WS*, we see that if agents of these strategies are matched up with each other under the imitation dynamic, there is an equal chance that either of the agent reproduces. But an *R-WS* has a significantly higher chance of beating an agent from the rest of the population. Against *SS* for example, *R-WS* has a 62.5% chance of winning: 50% of the time the payoff of 12 beats the sure payoff of 8 by *SS* and 1/2 of the time the two players are matched with equal payoff of 8 (25% chance), *R-WS* is favored. *SR* and *RS* on the other hand only have a 50% chance of winning against *SS*. Similar relations hold for *RR* and *R-WR*.

This shows an interesting dynamic of population-dependent success of agents:

- In an environment that contains *SR*, *RS*, and *R-WS* and no other strategies, all three should do equally well.
- In an environment that contains *SR*, *RS*, *SS* and *RR* and no other strategies, all four should do equally well.
- In an environment that contains *SR*, *RS*, *SS*, *RR*, and *R-WS*, *R-WS* will increase until *SS* and *RR* become extinct, at which point *SR* and *RS* and *R-WS* are at an equilibrium and remain at their current frequencies.

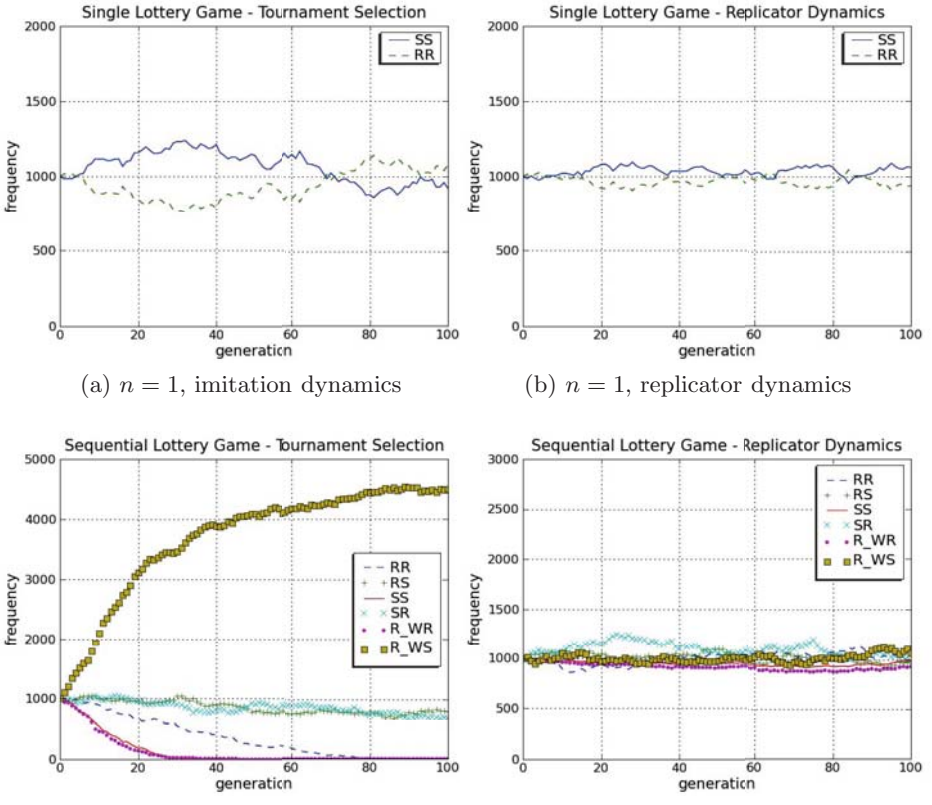
In the following section, we report on simulation results that confirm these predictions.

4 Simulation Results

To test the predictions at the end of the previous section, we have run simulations using all four combinations of the number of sequential choices ($n = 1$ or $n = 2$) and the reproduction mechanism (imitation or replicator dynamic). The types of agents were the ones described in Section 2.1. All simulations started with an initial population of 1000 agents for each agent type and were run for 100 generations, which was sufficient for us to observe the essential population dynamics.

Figures 2(a,b) show the frequency for each type of agent when $n = 1$. As we had expected, both *S* and *R* performed equally well (modulo some stochastic noise) regardless of which reproduction mechanism we used.

For $n = 2$ (Figures 2(c,d)), the results are more interesting and differ depending on the reproduction mechanism used. Under the replicator dynamic, all of the strategies performed equally well and remained at their frequency in the original population. But under the imitation dynamic, the conditional strategy *R-WS* outperformed the other strategies. *R-WS* rose in frequency relatively quickly to comprise the majority ($> 2/3$) of the population and remained at this high frequency throughout subsequent generations. Furthermore, the two unconditional strategies *SR* and *RS* remained, comprising the proportion of the population not taken over by *R-WS*.



(a) $n = 1$, imitation dynamics (b) $n = 1$, replicator dynamics

(c) $n = 2$, imitation dynamics (d) $n = 2$, replicator dynamics

Fig. 2. Agent type frequencies for all four simulations over 100 generations

5 Relations to Alternative Decision Making Models

The manner in which the *R-WS* strategy deviates from expected value maximization in our lottery game can be characterized as risk-averse (preferring the safe choice) when doing well in terms of payoff and risk-prone (preferring the risky choice) otherwise. Similar risk behavior is suggested by models such as prospect theory [1,5] and SP/A theory. In prospect theory, people are risk-seeking in the domain of losses and risk-averse in the domain of gains relative to a reference point. In SP/A theory [9], a theory from mathematical-psychology, aspiration levels are included as an additional criterion in the decision process to explain empirically documented deviations in decision-making from expected value maximization.

One explanation for the existence of decision-making behavior as described by such models is that the described behavioral mechanisms are hardwired in decision makers due to past environments in which the behaviors provided an evolutionary advantage [18]. Another interpretation, not necessarily unrelated,

is that the utility maximized by decision makers is not the payoffs at hand, but a different perhaps not obvious utility function. Along these lines, [19] proposes a model of decision making that includes probabilities of success and failure relative to an aspiration level into an expected utility representation with a discontinuous (at the aspiration level) utility function. Empirical evidence and analysis provided in [20] provide clear support for the use of probability of success in a model of human decision making. All these descriptive theories provide for agents to be sometimes risk-prone and sometimes risk-averse, depending on their current state or past outcomes, such as the *R-WS* in our simulations.

The sequentiality of choices in our game simulations allow for such state-dependent risk behavior to be explicitly modeled. One could theoretically model the sequential lottery game in normal form, i.e. reduce the choices to a single choice between the payoff distributions listed in Table 3. Doing so would provide essentially equivalent results except that the asymmetry in the payoff distribution of lotteries would be the determining factor of agent successes. In such a representation however, the analysis of risky and safe choices, and agents' preferences among them becomes blurred. In fact, we believe that a tendency towards modeling games in normal form often leads people to overlook the impact of sequentiality on risk-related behavior.

We believe our results show that imitation dynamics model an important mechanism that can lead to the emergence of risk-taking behavior with similar characteristics to that captured in alternative, empirical evidence-based models of decision making like the ones discussed above. Whenever the reproduction rate is not directly proportional to payoff (i.e., a reproduction mechanism other than the pure replicator dynamic),[1] risk propensities that differ from expected value maximization have the opportunity to be more successful than agents that solely consider expected value in their local choices. This suggests that there are many other reproduction mechanisms for which expected-value agents can be outperformed by agents that vary their propensities toward risk-taking and risk-averseness.

6 Conclusion

Our analytical and experimental results in several evolutionary lottery games demonstrate how sequentiality and reproduction can affect decision making under risk. Our results show that a strategy other than expected-value maximization can become prevalent in an evolutionary environment having the following characteristics:

– At each generation, the agents must make a sequence of choices among alternatives that have differing amounts of risk.

[1] We say "pure" here because the replicator dynamic can be modified to make reproductive success not directly proportional to payoff. For example, if a death rate (e.g. [21]) is implemented as a payoff-dependent threshold function, we might expect risk propensities to differ depending on whether an agent is above or below that threshold, similar to an aspiration level in SP/A theory.

– An agent's reproductive success is not directly proportional to the payoffs produced by those choices. We specifically considered an imitation dynamic known as tournament selection; but as pointed out in Section 5, we could have gotten similar results with many other reproduction mechanisms.

The most successful strategy in our analysis and experiments, namely the *R-WS* strategy, exhibits behavior that is sometimes risk-prone and sometimes risk-averse depending on its success or failure in the previous lottery. This kind of behavioral characteristic is provided for in descriptive theories of human decision making based on empirical evidence. It is not far-fetched to suppose that when human subjects have exhibited non-expected-value preferences in empirical studies, they may have been acting as if their decisions were part of a greater game of sequential decisions in which the success of strategies is not directly proportional to the payoff earned. Apart from a purely biological interpretation, in which certain behavioral traits are hardwired in decision-makers due to past environments, perhaps such empirical studies capture the effects of the subjects' learned habit of making decisions as part of a sequence of events in their daily lives.

Our results also demonstrate (see Fig. 2 and the last paragraph of Section 3) that the population makeup can have unexpected effects on the spread and hindrance of certain risk propensities. This may be an important point to consider, for example, when examining decision-making across different cultures or societies.

In conclusion, our simple lottery game simulations are a first step in exploring evolutionary mechanisms which can induce behavioral traits resembling those described in popular descriptive models of decision making. A specific related topic to explore is how the prospect-theoretic notion of setting a reference point may relate to evolutionary simulations with sequential lottery decisions. In general, there is much more opportunity for future work to use simulation for the purpose of exploring or discovering the mechanisms which induce, possibly in a much more elaborate and precise manner, the risk-related behavior characteristics described by prospect theory or other popular descriptive decision making models based on aspiration levels.

Acknowledgements

This work was supported in part by AFOSR grant FA95500610405, NAVAIR contract N6133906C0149, DARPA's Transfer Learning Program, DARPA IPTO grant FA8650-06-C-7606, and NSF grant IIS0412812. The opinions in this paper are those of the authors and do not necessarily reflect the opinions of the funders.

References

1. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. Econometrica 47, 263–291 (1979)
2. Friedman, M., Savage, L.J.: The utility analysis of choices involving risk. The Journal of Political Economy 56(4), 279–304 (1948)

3. Arrow, K.J.: Essays in the theory of risk-bearing. Markham, Chicago (1971)
4. Rabin, M.: Risk aversion and Expected-Utility theory: A calibration theorem. Econometrica 68(5), 1281–1292 (2000)
5. Tversky, A., Kahneman, D.: Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and Uncertainty 5, 297–323 (1992)
6. Loomes, G., Sugden, R.: Regret theory: An alternative theory of rational choice under uncertainty. The Economic Journal 92, 805–824 (1982)
7. Lopes, L.L.: Between hope and fear: The psychology of risk. Advances in Experimental Social Psychology 20, 255–295 (1987)
8. Lopes, L.L.: Re-modeling risk aversion. In: von Furstenberg, G.M. (ed.) Acting under uncertainty: Multidisciplinary conceptions, pp. 267–299. Kluwer, Boston (1990)
9. Lopes, L.L., Oden, G.C.: The role of aspiration level in risky choice: A comparison of cumulative prospect theory and sp/a theory. Journal of Mathematical Psychology 43, 286–313 (1999)
10. Harley, C.B.: Learning the evolutionarily stable strategy. Journal of Theoretical Biology 89, 611–633 (1981)
11. Dawkins, R.: The Selfish Gene. Oxford University Press, New York (1976)
12. Hales, D.: An open mind is not an empty mind: Experiments in the meta-noosphere. Journal of Artificial Societies and Social Simulation 1(4) (1998)
13. Hofbauer, J., Sigmund, K.: Evolutionary game dynamics. Bulletin of the American Mathematical Society 40(4), 479–519 (2003)
14. Gintis, H.: Game Theory Evolving: A Problem-centered Introduction to Modeling Strategic Behavior. Princeton University Press, Princeton (2000)
15. Hales, D.: Evolving specialisation, altruism, and group-level optimisation using tags. In: Sichman, J.S., Bousquet, F., Davidsson, P. (eds.) MABS 2002. LNCS, vol. 2581, pp. 26–35. Springer, Heidelberg (2003)
16. Riolo, R.L., Cohen, M.D., Axelrod, R.: Evolution of cooperation without reciprocity. Nature 411, 441–443 (2001)
17. Hales, D.: Searching for a soulmate - searching for tag-similar partners evolves and supports specialization in groups. In: Lindemann, G., Moldt, D., Paolucci, M. (eds.) RASTA 2002. LNCS, vol. 2934, pp. 228–239. Springer, Heidelberg (2004)
18. Houston, A.L., McNamara, J.M., Steer, M.D.: Do we expect natural selection to produce rational behaviour? Philosophical Transactions of the Royal Society B 362, 1531–1543 (2007)
19. Diecidue, E., Ven, J.V.D.: Aspiration level, probability of success and failure, and expected utility. International Economic Review 49(4), 683–700 (2008)
20. Payne, J.W.: It is whether you win or lose: The importance of the overall probabilities of winning or losing in risky choice. Journal of Risk and Uncertainty 30, 5–19 (2005)
21. Nowak, M.A., Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit for tat in the prisoner's dilemma game. Nature 364, 56–58 (1993)

Social Learning and Cumulative Innovations in a Networked Group

Thomas N. Wisdom and Robert L. Goldstone

Department of Psychological and Brain Sciences, Indiana University
1101 East 10th St., Bloomington, Indiana 47405 USA
{tnwisdom,rgoldsto}@indiana.edu

Abstract. We used a simple problem-solving game task to study imitation and innovation in groups of participants. Guesses were composed of multiple elements with linear and interactive effects on score, and score feedback was provided after each of a number of rounds. Participants were allowed to view and imitate the guesses of others during each round, and the visibility of score information accompanying others' guesses was manipulated in two conditions. When scores were not visible, social learning was impeded; participants were less efficient in their searching of the problem space and achieved lower performance overall. When scores were visible, results indicated a more equitable sharing of exploration among participants within groups as a result of selective imitation and cross-participant cumulative innovations, which were associated with higher performance.

Keywords: Social learning, innovation, imitation, problem solving, innovation diffusion.

1 Introduction

In a typical day, there are few activities that humans participate in that do not depend in some way on knowledge obtained from others. This is evident upon casual reflection about how people gather information and make choices about restaurants or movies, a candidate for a job or political office, a new city to live in or a large household purchase, not to mention explicitly collaborative tasks in many contexts. Such "social learning" has been defined broadly as "the acquisition of behavior by observation or teaching from other conspecifics" [1]. Social learning is a well-studied phenomenon in non-human animals, including foraging choices in starlings [2], food preferences in various rodent species [3], and mate choices in black grouse [4]. Humans' rare talent among animals for direct and flexible imitation has been called "no-trial learning" [5], because it is even faster than the one-trial learning observed in animals with a strong built-in tendency to form certain associations (e.g. between the taste of a food and a subsequent stomach ache). This talent allows an imitator to add new behaviors to his or her repertoire without the costs of trial-and-error learning.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 32–41, 2010.

1.1 Social Learning Strategies

Tendencies toward individual and social learning depend on the availability and reliability of information in the environment, including other learners. Laland [6] reviews strategies for *when* social learning is chosen, and *who* social learners choose to imitate. The first class of strategies (when to imitate) often uses the relative cost or uncertainty of asocial learning as criteria. For example, learning about predators on one's own can be very dangerous, so many animals have adapted to learn predator responses from others; in at least one instance this learning has occurred across species [7]. The second kind of strategy (who to imitate) often relies on absolute or relative performance of candidate solutions (such as *copy the best* or *copy if better* strategies, respectively), or their relative popularity (such as the *copy the majority* strategy); each of these strategies has been shown in several species [6].

1.2 Consequences of Social Learning

Rogers [8] performed simulations showing that in a temporally unstable environment, the extent to which random imitation is helpful depends on how recently the target of imitation has directly sampled the environment. Therefore, the addition of random social learners (information scroungers) to a population of asocial learners (information producers) does not improve the overall fitness of the population, because the costs of learning avoided by imitators will be offset by costs resulting from the use of outdated and inaccurate information. Boyd and Richerson [9] and Kameda and Nakanishi [10] confirmed and extended these results to show that when social learners can imitate selectively (e.g. choosing whom to imitate or imitating when individual exploration is relatively costly) the overall fitness of the population can increase, because individual learning can become more accurate or less costly.

The benefits for social learners (and thus average benefits for their group) in temporally stable environments are often assumed to be evident [11], but the mechanisms by which these benefits accrue are not necessarily clear. If social learning is essentially scrounging that only benefits imitators, then creating obstacles to social learning will only decrease the average performance of imitators. However, the results of previous experiments [12] have given us reason to believe that imitators are often also explorers, and that social learning serves as a vital component of cross-participant cumulative improvements. Thus impeding social learning is predicted to lead to decreases in the performance of all participants.

1.3 Experiment Overview

The following experiment investigates both the causes and consequences of social learning. We employ a task in which participants in groups consisting of between one and nine persons are instructed to individually build solutions consisting of multiple elements chosen from a larger set of elements over a series of rounds. These solutions are evaluated according to a score function that takes into account both individual element values and interactions between them. Groups of participants play simultaneously, and each can view the tentative solutions of all others. In one condition, participants may view fellow participants' scores alongside their solutions, and in another condition fellow participants' scores are invisible.

1.4 Predictions

We made the following predictions. When evaluative information about peer solutions was unavailable, participants would be unable to be sufficiently selective in imitation, and thus participants employing highly imitative strategies would have relatively lower scores. Imitation strategies in both conditions would be biased toward peers with solutions similar to the imitator's, and toward adopting solution elements that were more popular among peers, but these effects would be more pronounced in the invisible-scores condition in order to compensate for the lack of direct evaluative information. Mean scores would be lower for all participants (including successful asocial learners) in the invisible-scores condition because participants would not be able to easily take advantage of good solutions found by others through selective imitation and further improve upon them.

2 Methods

Participants were recruited from the Indiana University Psychological and Brain Sciences Department undergraduate subject pool, and were given course credit for taking part in the study. Participants populated each session by signing up at will for scheduled experiments with a maximum capacity of 9 persons. 209 individuals participated in the experiment, distributed across 61 sessions as shown in Table 1.

Table 1. Distribution of participants across group sizes

Group size	1	2	3	4	5	6	7	8	9
# Sessions	16	8	11	11	5	2	3	3	2
# Participants	16	16	33	44	25	12	21	24	18

2.1 Task Details and Instructions

We implemented the experiment using custom software written in Java and Flash and run in a web browser (a version of the task can be run as "Creature League" at http://groups.psych.indiana.edu/). Each participant used a mouse to interact with the experimental game. All participants' computers communicated with a game server, which recorded data and updated scores and team information for participants at the end of each round. In the game itself, participants attempted to maximize the number of points earned by their chosen subsets ("teams") from a set ("league") of creature icons over 24 rounds. The display included an area for the participant's own current team, another area that could be toggled to show the participant's previous round team or their best-scoring team up to that point in the game (along with its associated score), a league area which showed all of the icons (potential team members) that were available for selection, and indications of the current round in the game and the amount of time remaining in the current round. If a session included more than one participant, each participant's display also showed the team and, in the visible-scores condition, the associated score for each other participant in the previous round. Icons

could be copied from any part of the display to a participant's current team by dragging and dropping them with the mouse, except for those already on the participant's current team. The current team could be replaced entirely by another team by clicking the score box above it as a "handle" and dragging it to the current team area. A screenshot of the participant interface is shown in Figure 1.

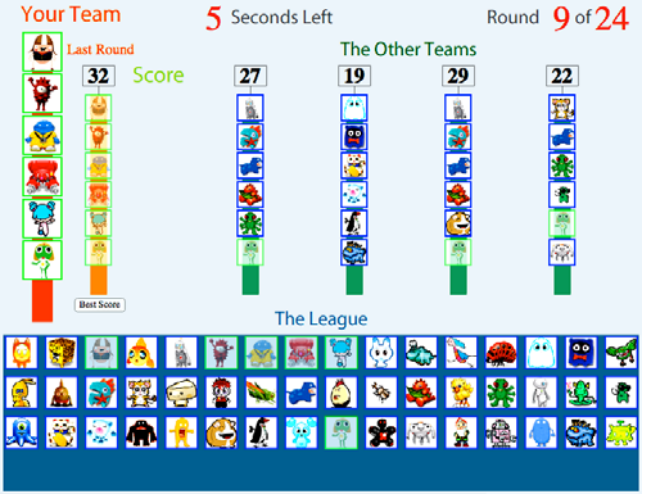


Fig. 1. Example of experiment task display

At the beginning of each session, players were given a hands-on demo of the game (including the various ways to move creatures to one's current team), and further informed about the mechanics of the game and what to expect in the remainder of the experiment session, including the following information. Each game consisted of 24 rounds, and each round was 10 seconds long. Score feedback was given after each round: if the participant's score had improved from the previous round, the numerical score display turned green and counted up to the new score, and if it had worsened, the display turned red and counted down to the new score. At the end of each game, the display showed the player's final score, along with a table of the scores of each player in each round of the game, sorted by average score. The player's own score was highlighted to show their relative performance without placing competitive emphasis on it. Players were instructed to do their best to maximize their teams' scores over all 24 rounds. At the beginning of each game, each player's team was a random selection of creature icons from the league. Each group played 6 games; in 3 of the games, other participants' scores were visible, and in the other 3 they were not. These were called the visible-scores and invisible-scores conditions, respectively, and were played in random order in each session.

In each game, each icon was associated with a certain positive number of points, and several special pairs of icons were associated with separate score bonuses or

penalties that captured interactions between icons. The score for a team was computed by summing the individual point values for each icon, and then adding or subtracting the value of any special pairs present. The pairs did not overlap, and the distribution was designed to be challenging: pairs which gave large positive bonuses were distributed among icons with small individual point values, and pairs which gave large negative penalties were generally found among icons with large individual point values. There was a greater number of positive interactions than negative ones, to give the score distribution a larger upper tail. For ease of comparison and analysis, all scores were normalized to the range [0,1] according to the minimum and maximum possible scores. The combinations of individual and pair values described above resulted in the probability distribution of scores among all possible teams shown in Figure 2. Participants were not given explicit information about the maximum score, the score distribution, or the position of the interaction terms. The icons' display position and associations with the point distribution were shuffled randomly for each game, so that their appearance and placement in the display did not give clues as to their point values during the course of an experiment session.

2.2 Dependent Variables and Definitions

In each round, the following data were automatically recorded for each player: the icons on the current team at the end of the round, the *source* of each icon, and the resulting score. The *source* information indicated whether an icon was unchanged from the previous round (Retained), copied from the player's own previous round team after initially being removed (Returned), copied from the player's own best-scoring team so far (Retrieved), chosen from the league display (Innovated), or copied from another player's team (Imitated). When Imitation was chosen, the persistent identifier of the copied player was recorded to allow further analyses of imitation decisions. In the case of a player replacing the entire team with Imitated icons, only the choices that were not already present on the team were counted as Imitated. Similar criteria applied to replacement of an entire team with Retrieved icons, or removing an icon and then putting it back on the team via an Innovation choice.

Choice similarity was defined as the proportion of icons that two teams have in common. An *improvement* was defined as an instance of a participant obtaining a score higher than all prior scores of all players within a particular game. Each participant's normalized *improvement share* was defined as their individually achieved proportion of the total improvements achieved by all participants in a condition, multiplied by the number of participants. A value of 1 indicated a "fair" share, e.g. a participant achieved one third of the improvements in a three-person session. A participant's *score rank* in a particular round was defined as the rank of their score (one being the best) among all scores in the group in that round; individuals with the same score had the same rank. *Guess diversity* for a group in a particular round was defined as the proportion of icons in the league represented on one or more participants' teams in that round. This value was normalized by the average expected value of this proportion for each participant group size, generated by a Monte Carlo simulation.

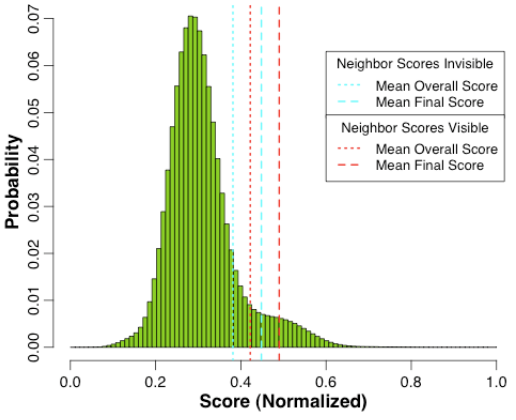


Fig. 2. Distribution of scores for all possible teams

3 Results

3.1 Differences in Performance

Participants achieved mean overall (across all rounds) and final normalized scores of .381 and .448 in the scores-invisible condition, and significantly higher scores (.422 and .490) in the scores-visible condition, as shown in Figure 2. Linear mixed-effects models were used to examine trends across rounds for score and guess diversity, with a random effect of participant group. Analysis of score versus round showed a strong positive trend in the scores-visible condition ($F(1,1402)=264.69$, $p<.0001$, $B=.520$, mean increase=0.184), and a slightly weaker positive trend in the invisible-scores condition ($F(1,1402)=223.05$, $p<.0001$, $B=.601$, mean increase=0.140; see Figure 3). Guess diversity showed a similarly strong decrease across rounds in the scores-visible condition ($F(1,1034)=263.40$, $p<.0001$, $B=-.448$, mean change=-0.460), and a weaker decrease in the scores-invisible condition ($F(1,1034)=80.28$, $p<.0001$, $B=-0.461$, mean change=-0.254; see Figure 3).

Mean proportions of each choice source for improvement and non-improvement guesses in each condition are shown in Table 2. In both conditions, the proportion of Innovation choices was higher for guesses that yielded improvements relative to non-improvements (scores-invisible: $t(690.5)=-13.13$, $p<.0001$; scores-visible: $t(825.9)=-15.58$, $p<.0001$). In the scores-invisible condition, the proportion of Imitation choices was significantly lower for improvements than non-improvements ($t(896.9)=11.39$, $p<.0001$), while in the scores-visible condition, the proportion of Retention choices was significantly lower for improvements than non-improvements ($t(832.1)=8.86$, $p<.0001$). Overall there was significantly higher Retention in the scores-visible condition ($t(360)=-2.218$, $p=.027$, indicating that guesses changed more slowly.

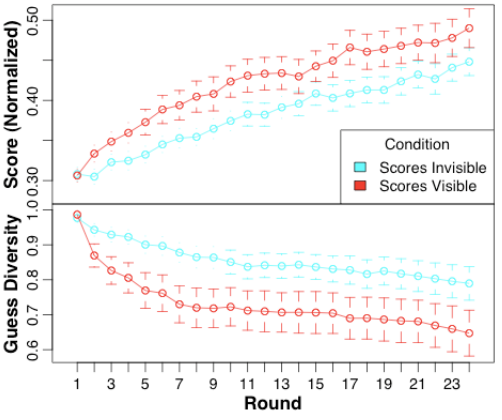


Fig. 3. Change in score and guess diversity across rounds in each condition

Table 2. Mean choice source proportions for (non-)improvement guesses in each condition. Significant differences within a condition are shown in boldface, and significant differences between condtions are shown in italics.

Condition	Improvement?	Imitate	Innovate	Retain	Retrieve
Scores Invisible	No	**.090**	**.139**	*.715*	.044
	Yes	**.034**	**.218**	*.706*	.037
Scores Visible	No	.088	**.117**	***.762***	.023
	Yes	.082	**.195**	***.693***	.022

Analyses of relationships between mean individual score and mean individual choice source proportions showed a strong negative correlation in both conditions between score and prevalence of Innovation choices (scores-invisible: $F(1,181)=55.77$, $p<.0001$, $B=-0.485$; scores-visible: $F(1,181)=138.5$, $p<.0001$, $B=-0.658$) and a strong positive relationship between score and Retention (scores-invisible: $F(1,181)=14.66$, $p=.0002$, $B=0.274$; scores-visible: $F(1,181)=59.96$, $p<.0001$, $B=0.489$), while a strong positive relationship was shown for Imitation only in the scores-visible condition ($F(1,181)=8.65$, $p=.0037$, $B=0.214$), and a strong positive relationship was shown for Retrieval only in the scores-invisible condition ($F(1,181)=15.2$, $p=.0001$, $B=0.278$).

Histograms of normalized improvement share showed a relatively equitable distribution of improvements within groups in the scores-visible condition, with the distribution peaked near a "fair" share of 1. In the scores-invisible condition, however, the distribution had a strongly inequitable skew, with a modal share of zero (see Figure 4). Mean overall score showed a strong positive correlation with improvement share in the scores-invisible condition ($F(1,137)=49.17$, $p<.0001$, $B=0.369$), but this relationship was not evident in the scores-visible condition.

3.2 Differences in Strategy

The score rank of imitated participants, as well as the difference in score between imitators and those they imitated, are shown in Table 3. These figures confirm that participants in the scores-visible condition were strongly biased toward imitating those with absolutely and relatively higher scores, while imitation decisions in the scores-invisible condition were essentially random with regard to the performance of peers' solutions.

A comparison between the mean choice similarity of participants' most recent guesses to those whom they imitated, and to those whom they did not imitate, revealed a slight but significant positive difference in the scores-visible condition: a similarity value of .557 for imitated and .515 for non-imitated guesses ($t(4584)=-5.6371$, $p<.0001$). The opposite was true in the scores-invisible condition: .316 for imitated and .338 for non-imitated guesses ($t(3534)=2.9523$, $p=0.003$). In other words, when scores were visible, imitation was biased toward similar guesses, and when scores were invisible, imitation was biased toward dissimilar guesses.

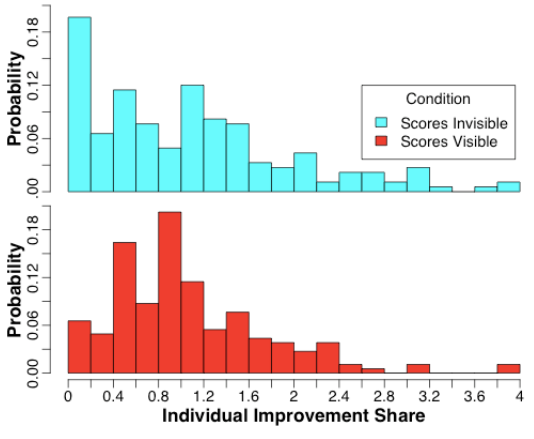


Fig. 4. Histograms showing relatively equitable achievement of improvements within groups in the scores-visible condition, and an inequitable distribution in the scores-invisible condition

Table 3. Mean proportions of imitations of each score rank among imitated players, and mean proportions of positive and negative score differences between imitated and imitating players in each condition

	Imitated Score Rank			Imitated Score Difference	
Condition	1	2	3+	> 0	< 0
Scores Invisible	.255	.225	.520	.544	.415
Scores Visible	.787	.113	.099	.883	.095

In order to measure the bias of participants to choose an icon according to its frequency in peers' teams, we tallied the number of players in the group whose teams included each icon in the previous round (N_{R-1}), as well as the number of the

remaining players who added it to their team in the current round via Imitation. To convert these figures to normalized frequencies, the first number was divided by the participant group size (N), and the second number was divided by the number of participants who did not possess the icon in the previous round ($N - N_{R-1}$). A linear mixed-effects analysis of imitation probability versus choice frequency showed a positive frequency bias that was significantly greater than chance in the scores-visible condition ($F(1,1128)=1648$, $p<.0001$, $B=.300$) and significantly below chance in the scores-invisible condition. These results indicate that for the scores-visible condition, the probability of imitating an icon increased significantly beyond chance as an increasing number of a participant's fellow participants possessed the icon.

4 Discussion

When scores were visible, participants were heavily biased toward imitating higher-performing peers, as would be expected, and performance was correlated with the average amount of Imitation in a participant's choices. Participants also showed a bias toward imitating solution elements that were possessed by more than half of their fellow participants, similar to the *copy the majority* strategy discussed in [6]. Another bias evident in the score-visible condition was toward imitating similar guesses, which allowed the imitator to make use of social learning while keeping a solution partially compatible with previous solutions and existing knowledge of the problem space, a phenomenon explored in studies of innovation propagation [13].

As expected, hiding other participants' scores strongly impeded social learning – when others' scores were not visible, the choice of whom to imitate was fairly random as expected, and performance was correlated with the average amount of Retrieved information on a participant's team, showing the incentive to focus on previously-acquired information rather than that of others. Unexpectedly, however, participants seemed to distrust the popularity of solution elements on fellow participants' teams as a cue to their value; as the prevalence of an icon among peers' teams increased, it was less likely to be imitated than expected by chance. This is consistent with Laland's [6] *copy when uncertain* strategy (or its corollary, *don't copy when certain)* in that, having been encouraged to focus on individual exploration by the low returns to imitation, participants trusted the information they had gained through asocial learning and avoided the uncertain social information conveyed by the adoption frequency of peers' solution elements. Participants in the score-invisible condition also seemed to be slightly biased against peer solutions that were similar to their own, perhaps indicating a bias toward novelty, which would help explain the overall decrease in individual Retention in this condition.

As seen in the increasing score and decreasing guess diversity trends across rounds, average performance increased via the convergence of group members on regions of the problem space that contained high-quality teams. This convergence combined with a small amount of individual exploration caused such regions to be explored more thoroughly and still better solutions to be found. However, in the scores-invisible condition, when imitation was not focused on a small group of better-performing neighbors (through a bias to select high scoring teams, because scores were not available), popular solution elements (through frequency bias), or similar

guesses, this convergence happened much more slowly, search was more diffuse and less efficient, and lower performance resulted.

The significant correlation of improvement share with mean scores in the score-invisible conditions shows that individuals who were relatively more successful at individual exploration were rewarded with proportionately better overall scores compared to others, because their fellow players could not easily copy their improvements and achieve their scores. In the score-visible conditions this relationship disappeared, but mean scores increased significantly such that all participants generally did better. In other words, when social learning was unimpeded in the scores-visible condition, high and low individual achievers had approximately the same payoffs, but absolute payoffs were higher for all compared to the scores-invisible condition. This is because imitators were not merely scroungers; the high proportion of Imitation present in improvements shows that imitated guesses were often the basis for further cumulative innovations.

Acknowledgements. The authors would like to thank Xianfeng Song, Zoran Rilak, and Todd Gureckis for their help in designing and programming the experiments. This work is funded by National Science Foundation REESE grant 0910218.

References

1. Boyd, R., Richerson, P.J.: The Origin and Evolution of Cultures. Oxford University Press, New York (2005)
2. Templeton, J.J., Giraldeau, L.-A.: Vicarious Sampling: The Use of Personal and Public Information by Starlings Foraging in a Simple Patchy Environment. Behav. Ecol. Sociobiol. 38, 105–113 (1996)
3. Galef Jr., B.G., Giraldeau, L.-A.: Social Influences on Foraging in Vertebrates: Causal Mechanisms and Adaptive Functions. Anim. Behav. 61, 3–15 (2001)
4. Höglund, J., Alatalo, R.V., Gibson, R.M., Lundberg, A.: Mate-Choice Copying in Black Grouse. Anim. Behav. 49, 1627–1633 (1995)
5. Bandura, A.: Vicarious Processes: A Case of No-Trial Learning. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. II. Academic Press, New York (1965)
6. Laland, K.N.: Social Learning Strategies. Learn. Behav. 32, 4–14 (2004)
7. Krause, J.: Transmission of Fright Reaction Between Different Species of Fish. Anim. Behav. 65, 595–603 (1993)
8. Rogers, A.R.: Does Biology Constrain Culture? Am. Anthropol. 90, 819–831 (1988)
9. Boyd, R., Richerson, P.J.: Why Does Culture Increase Human Adaptability? Ethol. Sociobiol. 16, 125–143 (1995)
10. Kameda, T., Nakanishi, D.: Does Social/Cultural Learning Increase Human Adaptability? Rogers's Question Revisited. Evol. Hum. Behav. 24, 242–260 (2003)
11. Kameda, T., Nakanishi, D.: Cost-Benefit Analysis of Social/Cultural Learning in a Non-Stationary Uncertain Environment: An Evolutionary Simulation and an Experiment with Human Subjects. Evol. Hum. Behav. 23, 373–393 (2002)
12. Wisdom, T.N., Song, X., Goldstone, R.L.: The Effects of Peer Information on Problem-Solving in a Networked Group (manuscript in preparation)
13. Rogers, E.M.: Diffusion of Innovations, 5th edn. Free Press, New York (2003)

Understanding Segregation Processes

Elizabeth Bruch

Assistant Professor of Sociology and Complex Systems
University of Michigan

Abstract. There is growing consensus that living in neighborhoods of concentrated poverty increases the likelihood of social problems such as teenage parenthood, drug and alcohol use, crime victimization, and chronic unemployment. Neighborhood inequality is also implicated in studies of enduring race/ethnic health disparities, and there are recent moves to broaden the definition of health care policy to policies targeting social inequality (Mechanic 2007). Residential segregation affects health outcomes in several different ways. First, income, education, and occupation are all strongly related to health (Adler and Newman 2002). Segregation is a key mechanism through which socioeconomic inequality is perpetuated and reinforced, as it hinders the upward mobility of disadvantaged groups by limiting their educational and employment opportunities. Second, segregation increases minority exposure to unhealthy neighborhood environments. Residential segregation creates areas with concentrated poverty and unemployment, both of which are key factors that predict violence and create racial differences in homicide (Samson and Wilson 1995). Neighborhood characteristics, such as exposure to environmental hazards, fear of violence, and access to grocery stores, affect health risks and health behaviors (Cheadle et al. 1991). Tobacco and alcohol industries also advertise their products disproportionately in poor, minority areas (Moore, Williams, and Qualls 1996). Finally, residential segregation leads to inequalitie in health care resources, which contributes to disparities in quality of treatment (Smedley, Stith, and Nelson 2002).

One strategy for reducing health disparities among ethnic groups, particularly among blacks and whites, would be to reduce or eliminate racial residential segregation and the corresponding geographic concentration of poverty. Currently, we have a number of programs in place that are designed, directly or indirectly, to affect the socio-demographic characteristics of neighborhoods (e.g., FHA low-income mortgage loans, Hope VI mixed income housing developments, and Section 8 housing vouchers). However, despite almost a century of neighborhood research, social scientists do not know what forces maintain segregation, what the respective strengths of these forces are, or how they combine. In the absence of theoretical and empirical knowledge of how segregation works, it is difficult to effectively evaluate the short-run effectiveness of policies, let alone their long-run equilibrium consequences. Moreover, without some understanding of the global impact of these housing policies, interventions may be ineffective, or worse yet produce unintended consequences that exacerbate problems of concentrated poverty. Examples of unintended consequences include the exploding crime rates in mid-sized cities which have been linked to the anti-poverty programs Hope IV and Section 8 (see Rosin 2008; Galster 2003; Bostic and Lee 2008).

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 42–43, 2010.

Neighborhood formation and change is one example of a process in which collective dynamics emerge from the behavior of individual agents, mediated through interaction and aggregation. People who exit a neighborhood because they cannot tolerate its ethnic composition are both responding to and modifying neighborhood ethnic composition. This leads to a "spiral" or "domino effect" whereby small changes in neighborhood composition amplify and result in unexpected and extreme segregation. This is true even when people have relatively mild preferences for their own group (Schelling 1971, 1978). A number of health problems share these features of dynamic interdependence and feedback, including the spread of obesity, the development of drug-resistant strains of tuberculosis and Staph infections, and the recent increase in HIV infection rates following the development of more effective antiviral treatments (see La Berge 2008; Elford 2006; Fong and Drlica 2003; Palumbi 2001; Lightfoot et al. 2005). Existing methods for studying this type of process have not been successful in developing effective mechanisms for change. Such processes have been dubbed "policy resistant," as attempts at interventions often fail due to the response of the system to the intervention (Sterman 2006).

My work blends empirical, analytic, and computational methods in an effort to develop a body of theoretical knowledge and corresponding agent-based model that can identify a few mechanisms central to maintaining segregated neighborhoods. The model incorporates information about how individuals discover and evaluate potential destinations, the relationship between decisions about where to live and aggregate patterns of segregation, and an understanding of how policies aimed at reducing segregation affect mobility behavior. The research is iterative, with model validation providing clues into what might be misspecified or missing from the behavioral or agent-based models, resulting adjustments of the behavioral and agent-based models, and subsequent reevaluation. Provided the agent-based model can reproduce key features of neighborhood turnover in different cities, my hope is that it will be useful in the development and evaluation of alternative policies for reducing segregation. Along the way, the model will provide useful information including how people respond to their environment under conditions of change and uncertainty, the relationship between mobility patterns, economic inequalities among ethic groups, and racial segregation, how segregation dynamics scale with population size, and the effect of government interventions such as public housing and housing vouchers on mobility patterns and the housing market.

Social Factors in Creating an Integrated Capability for Health System Modeling and Simulation

Paul P. Maglio, Melissa Cefkin, Peter J. Haas, and Pat Selinger

IBM Research – Almaden, San Jose, California
{pmaglio,mcefkin,peterh,patseli}@us.ibm.com

Abstract. The health system is a complex system of systems – changes in agriculture, transportation, economics, family life, medical practices, and many other things can have a profound influence on health and health costs. Yet today, policy-level investment decisions are frequently made by modeling individual systems in isolation. We describe two sets of issues that we face in trying to develop a platform, method, and service for integrating expert models from different domains to support health policy and investment decisions. The first set of questions concerns how to develop accurate social and behavioral health models and integrate them with engineering models of transportation, clinic operations, and so forth. The second set of questions concerns the design of an environment that will encourage and facilitate collaboration between the health modelers themselves, who come from a wide variety of disciplines.

Keywords: Health, Policy, Models, Simulation, Social Factors.

1 Toward a Science of Health Policy Decision Making

The health system of any nation is a complex system of systems. Decisions about comparative effectiveness or about investment in prevention or treatment programs may lead to complex interactions and have widespread consequences, many of which may be difficult to foresee. For example, the treatment of chronic diseases presents multi-faceted issues that the healthcare sector alone cannot address. Transportation, agriculture, housing, and education "have far-reaching health effects, but are not engaged or evaluated for those outcomes" [4]. Indeed, it is generally recognized that chronic diseases such as obesity reflect cultural, social, educational, political, and economic conditions as well as policies, practices, costs, and pricing in industries such as advertising, transportation, agriculture and others [8]. Certain sorts of behavioral modeling approaches may be appropriate for simulating some aspects of chronic disease [1,6], and various kinds of system modeling may be useful for simulating complex interactions of the effects of policies [7,13]. But a full understanding of such a complex system of systems – like the health system – can be enabled by modeling all relevant aspects of each constituent real-world system, probably by different experts using different modeling techniques, and then integrating the resulting models to "try out" alternatives. Though there have been some efforts at building frameworks that encompass data of various sorts to model and predict policy-level outcomes

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 44–51, 2010.

(e.g., [16]), there exists no overarching platform or framework with which to integrate disparate models based on distinct technologies and deep domain expertise.

To address this unmet need, we are developing a platform, method, and service to support such an integration of models. The *Smarter Planet Platform for Analysis and Simulation of Healthcare* – also known as *Splash!* – aims to enable the integration of independently created, deep models of health-related domains in an environment that is practical, flexible, cost-effective, and usable. Our goal is to have an impact on health at policy and investment levels, in understanding comparative effectiveness of treatments and preventions, in determining return on investment at an ecosystem level, and in understanding global consequences of decisions. Numerous technical and conceptual challenges must be addressed to integrate diverse models.

In this paper, we summarize some of the research challenges brought to the fore when considering social factors related to health and health policy, and the formation of an integrated modeling capability. We identify a dual set of challenges: (1) What particular issues must be faced in integrating social and behavioral models with statistical and deterministic models derived from other conceptual domains and data sources? And (2) how might the social conditions of different modelers and communities of experts themselves – their varying disciplinary assumptions, practices, and concerns – be addressed so that Splash! effectively enables collaboration that supports development of practical and meaningful results? We believe that the core research questions identified and raised in this examination represent important initial steps toward identifying opportunities to advance contributions of social and behavioral modeling to health and health policy. We also aim ultimately to provide insight on how joint efforts between modeling and policy communities – and multiple disciplines more generally – can continue to interact productively, forging innovative advances to knowledge and action.

2 The Splash! Approach: Architecture and Challenges

Currently, there are no means for usefully combining multiple independently created models to inform the kind of complex decision making demanded of health policy. There are many reasons why diverse models are rarely combined to create a comprehensive, detailed picture of any real system of systems. Different categories of models are constructed, maintained, and used by different people and organizations, each using distinct terms, conventions, and approaches. The challenges to creating integrated views are both technical and social, emerging in part from varied intellectual and scientific histories and practices.

2.1 Overall Architecture and Challenges for Model Integration

There are four main challenges to composing large-scale models for complex health ecosystems. First, *not all models can be combined in a sensible way*. The assumptions, time scales, capabilities, level of detail, and indeed the selection of the key aspects to represent may be quite different: What factors characterize the models that are compatible with one another? The challenge is to develop a deep understanding of model compatibility.

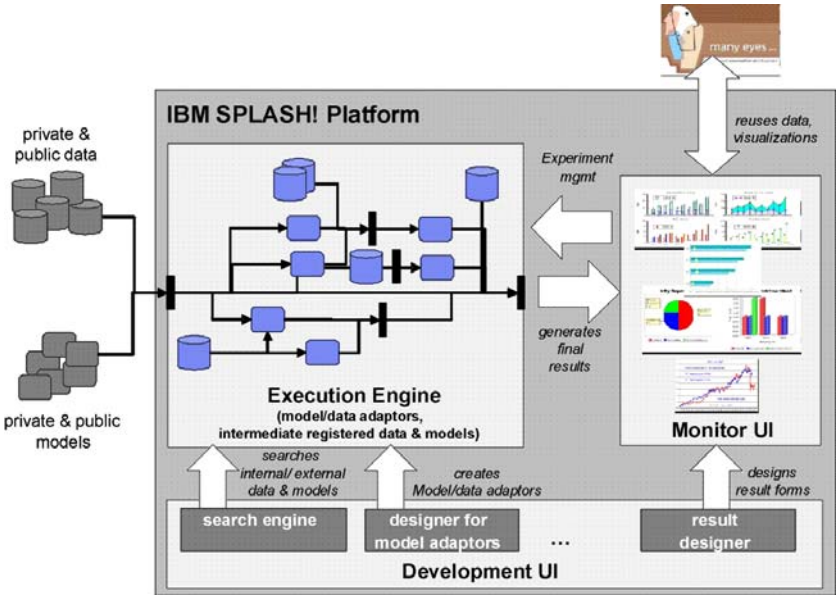


Fig. 1. Splash! will be an open community platform where proprietary and public models, data, and outcomes can be searched, combined, executed, visualized, and shared

Second, *there exists no standard way to describe models in sufficient depth to determine compatibility*. Here, the challenge is to create mechanisms and methods for describing models so that it is easy to determine how to integrate them into larger, more complex models of larger, more complex systems.

Third, *there are no tools or platforms to support the integration of independently created models in a simple, flexible, and useful way*. This adds the challenge of providing efficient mechanisms for searching and identifying applicable models, for establishing an appropriate execution environment, for automatically generating connectors between models and datasets, and for enabling reuse, result pruning, data transformations, flexible model transformations, experiment management, visualization, simulation output analysis, and so on (see Figure 1).

Fourth, *there is no targeted technology and set of practices to facilitate collaboration between the varied people and organizations that develop and use distinct domain models.* We envision an active community of participants contributing models and data, combining models, discussing models, exploiting previous results, and optionally sharing their models and modeling results. Participants in such an open community must have the means to (a) combine their proprietary models and data securely without risking intellectual property or violating privacy, (b) evaluate the quality of models and transformations used and communicate their findings to others, and (c) assess the trustworthiness of the outcomes produced. The final challenge is to develop a deep understanding of what is required for such an open integrated community system to successfully enable cooperation among all stakeholders.

Here, we ask whether and in what ways the inclusion of social computing approaches and social models into the integrated mix of models envisioned in Splash! presents particular challenges. What assumptions, forms of modeling, and language use, for instance, inform social and behavioral modeling in the health domain? We think that for the technology and supporting practices for Splash! to be useful, usable, and effective, they must be grounded in an understanding of the work and collaboration practices of the varied people and organizations that develop and use these models. So the fourth challenge identified above follows from the first three in that it is underlined by basic questions of the compatibility of the assumptions informing the models as well as the ways of describing them. These assumptions and languages, in turn, are informed by the varying social and intellectual histories of different scientific disciplines and other communities of experts. Historians of science and social scientists in the area of science and technology studies offer insight for consideration of what happens when different scientific and policy communities come together, suggesting ways that differences both challenge and create opportunities for greater advancement.

2.2 Challenges for Social and Behavioral Modeling

Consider the case of chronic disease management, such as obesity. Not only do numerous social factors inform underlying health conditions, but the interplay of social factors in determining impact of various forms of intervention is undeniable [9]. The number of factors affecting health outcomes is multilayered and highly complex (see Figure 2).

For instance, recent studies have examined how environmental factors contributing to access to food, as determined by availability and price, correlate with variable health outcomes. Findings have shown that lower food prices are associated with consumption of those food products; for instance, lower priced fruits and vegetables are associated with greater consumption of these products while lower priced fast food is associated with lower fruit and vegetable consumption. The ability to benefit from lower prices depends on the potential access to them to begin with, hence the shortage of markets selling lower priced fruits and vegetables in lower income areas is seen to contribute to higher rates of obesity in such communities [3]. Various policy interventions are possible, such as providing tax benefits to merchants for supplying lower-cost healthy food. What is needed is the ability to simulate the potential impact of such a move given the dynamic and non-deterministic dimensions of the social factors contributing to impact. In addition to being dynamic, finding appropriate means of defining the parameters of social factors introduces additional challenges. For instance, changing forms of ethnic identification and residence patterns (e.g., Pacific Islanders and traditionally African-American areas) and attendant shifts in consumer behavior must be considered.

Indeed social networks have been shown to reveal interesting patterns of obesity and weight gain and loss. Building off the data available through the longitudinal Framingham Heart Study, analysis of social networks conducted by Bahr et al. [1] identifies obesity clusters to be more prevalent at the second and third degrees of relationship – friends and friends of friends – rather than in familial or spousal units. They simulate the effect of certain social forces, such as advertising or taxation, on particular spots in the networks as predictive exercises in guiding potential policy.

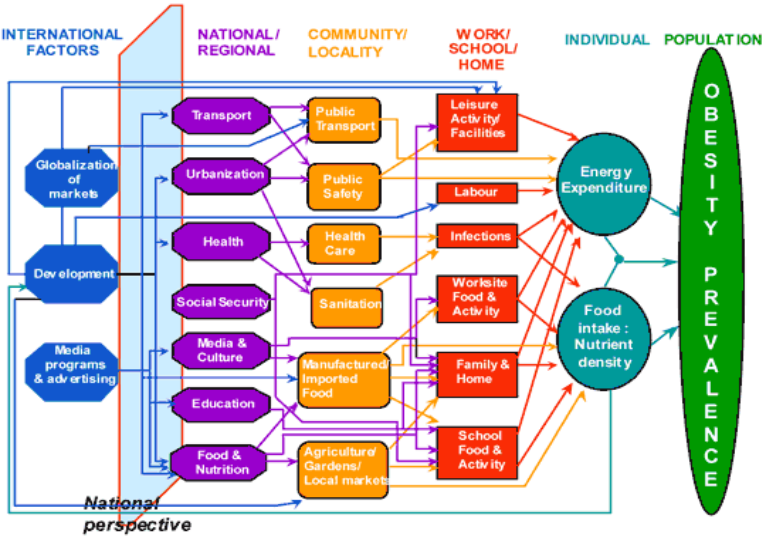


Fig. 2. Example of the complex system of systems related to obesity (from [8])

Social and behavioral factors inform not only health choices but likely responses to policy interventions at every turn, from preferences (e.g., consumer choices) to satisfaction (e.g., with healthcare treatment options) to forms of resistance (e.g., to government policies) and beyond. The challenge is to understand how to merge the kinds of models amenable to modeling social factors – social network or agent-based models, for instance – with other forms of deterministic models. The model-integration problem becomes even more challenging when merging these behavioral models with other types of models, such as transportation models that help determine access to supermarkets or clinics, models of health facility utilization, or econometric models (see, for instance, [14]).

2.3 Challenges in the Social Practices of Modelers

Questions raised by consideration of the integration of social and behavioral models in understandings of health and health policy with other model types points to a broader set of considerations around the worldviews and social practices that inform the underlying assumptions of the models and the expectations of their application. In practical terms, we aim to understand what it will take to bring together contributors in ways that will support fruitful collaboration across diverse communities of experts and that lead to the production of meaningful and useful outputs. What enabling technologies and sets of practices will support the kinds of knowledge-production and decision-making requirements of users? Here we broaden the lens from the particulars of the social and behavioral models themselves to focus more on the meta-level practices of modelers and those who aim to benefit from their results. Our interest parallels that of the conference itself: We are asking what the operational considerations of

the environment need to be to encourage experimentation and derive novel theories to help address the complex challenges of health and health policy.

Numerous investigators of the work of interdisciplinary innovation and development have identified the communicative, cognitive, and broader social challenges of creating meaningful exchange among scientific and other communities of experts. The notion of "trading zone" offers a compelling metaphor for consideration of how knowledge exchange across diverse scientific and technical communities occurs. Most famously associated with Galison's [5] description of how the distinctly different communities from engineering and science developed radar and particle detectors, the "trading zone" provides a means for recognizing how, despite potential misunderstandings (taking something to mean something other than what was intended) and mis-recognitions (thinking things mean the same when they do not), experimentation continues and communities manage to collaborate. Galison's and others' cases of how the trading zone operates highlight how languages developed by different communities of experts matter and must be accounted for. Star and Griesemer's [12] notion of "boundary object" adds a material and spatialized dimension of understanding to the picture. Boundary objects "are objects which are both plastic enough to adapt to local needs and constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual-site use. They may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable means of translations." [12, p. 393] This notion points to the fact that technical and scientific practice involves the manipulation of tangible *and* conceptual objects.

In considering how to establish productive interchanges between health and health policy researchers of varying backgrounds, and in particular where and how to integrate social and behavioral modeling approaches, we must recognize that the challenges do not concern only technicalities of modeling integration, but also concern establishment of the right environment for collaboration, including tools and terms of discourse. There are risks in not interrogating this aspect as we move ahead in model integration. Models and simulations are powerful social products in their own right, and while they can significantly advance understanding and compel action, they can also overwhelm and lead to ill-guided actions (see [15]). For example, Brailsford et al. [2] cite one example of a clash between communities in which healthcare clinicians objected to hospital operational models that were adapted from the manufacturing industry and retained too much manufacturing terminology; the clinicians felt that such models reduce people to "widgets in a production line" and are doomed to fail.

The fluency with and ability to critically engage with models and their results is sure to vary across participants in the health and health policy arenas. Based on an ethnographic study of climate modelers, Lahsen [11] argues that in many cases users of models are better able to maintain critical distance on the value and veracity of the model than the modelers themselves. Debates can emerge, for instance, when models aim toward providing answers to help refine existing policies (which may or may not be supported by the individuals and organizations engaged in the modeling efforts) as compared to when they aim to inform the creation of new policy. At stake are potential conflicts of interest and guarding of intellectual property, and such challenges are likely to play out through critical engagement with the very bases of the models.

A related set of challenges arises in considering the usage of an integrated system of diverse models such as Splash! Even assuming that the issues of differing vocabularies of discourse and conflicts of interest among modelers from different domains are overcome, a number of problems still remain. Suppose that Splash! succeeds in creating a cooperative platform with many diverse models cataloged in its repository. What factors and facilities are required for a newcomer to join in this community, integrate a set of models, and take action based on the results? What are the mechanisms for judging the value of individual models and specific model combinations? How is trust built among existing and newcomer community participants? One example mechanism is illustrated by the ManyEyes web site for data sets and visualizations [17]. In ManyEyes, participants may register, comment, blog, and provide ratings on data sets. Provenance of data is explicitly described in bylines associated with data.

Furthermore, even assuming the validity of each given model, how can a policy maker be assured that the outcome of an integration of trusted, expert models is itself scientifically valid and can be trusted sufficiently to make health system decisions or investments? That is, how can participant trust be transferred from the component models to combinations of such models? Can mechanisms such as peer review, rating systems, blog discussions, certification by trusted authorities, etc., play a role in building such trust? Can a cooperative platform leverage or enhance existing technologies for model verification and validation? We intend to explore the opportunities and vulnerabilities of collaborative modeling in addressing these challenges.

3 Summary

As a dynamic and complicated system of systems, health continues to demand increasingly sophisticated understanding to encourage hopeful experimentation and development toward improvement. Determination of health policy and the ability to monitor effect and outcome is similarly complex. We aim to provide scientific support to these efforts by enabling engineering, synthesis, and integration of models of real-world systems to provide a means to try out possible alternatives through simulation. But this means we must account for social factors that underlie and inform health-related actions and outcomes. This paper raises and explores two sets of questions for advancing social and behavioral modeling: (1) how do we effectively integrate social and behavioral models with other models to inform complex systems understanding? and (2) how do we create the appropriate social environment to encourage participation by diverse individuals and organizations in this integration?

Acknowledgments. The ideas for Splash! were developed by a team at IBM Research – Almaden, including Christopher Campbell, John Day, Susanne Glissmann, Kelly Lyman, Ben Shaw, Susan Stucky, and Steve Welch.

References

1. Bahr, D.B., Browning, R.C., Wyatt, H.R., Hill, J.O.: Exploiting social networks to mitigate the obesity epidemic. Obesity 17(4) (2009)
2. Brailsford, S.C., Bolt, T., Connell, C., Klein, J.H., Patel, B.: Stakeholder engagement in health care simulation. In: Proc. Winter Simulation Conf., pp. 1840–1849 (2009)

3. Chaloupka, F.J., Powell, L.: Price, availability, and youth obesity: evidence from Bridging the Gap. Prev Chronic Dis. 6(3) (2009)
4. Collins, J.L., Marks, J.S., Koplan, J.P.: Chronic disease prevention and control: coming of age at the Centers for Disease Control and Prevention. Prev. Chronic Dis. 6(3) (2009)
5. Galisoen, P.: Image & logic: A material culture of microphysics. The University of Chicago Press, Chicago (1997)
6. Hammond, R.A.: A complex systems approach to understanding and combating the obesity epidemic. Brookings Institution, Washington (2008)
7. Homer, J., Milstein, B., Wile, K., Trogdon, J., Huang, P., Labarthe, D., Orenstein, D.: Simulating and Evaluating Local Interventions to Improve Cardiovascular Health. Prev. Chronic Dis. 7(1) (2010)
8. Huang, T.T., Drewnowski, A., Kumanyika, S.K., Glass, T.A.: A systems-oriented multilevel framework for addressing obesity in the 21st century. Prev. Chronic Dis. 6(3) (2009)
9. Institute of Medicine and Robert Wood Johnson Foundation: Workshop on Unintended Consequences of Health Policy Programs and Policies (2001), http://books.nap.edu/openbook.php?record_id=10192&page=15#p200049679970015001
10. Kumanyika, S., Jeffery, R.W., Morabia, A., Ritenbaugh, C., Antipatis, V.J.: Public Health Approaches to the Prevention of Obesity (PHAPO) Working Group of the International Obesity Task Force (IOTF). Obesity prevention: the case for action. Int. J. Obes. Relat. Metab. Disord 26(3), 425–436 (2002)
11. Lahsen, M.: Seductive Simulations? Uncertainty Distribution around Climate Models. Social Stud. of Sci. 35 (2005)
12. Star, S.L., Griesemer, J.: Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social Stud. of Sci. 19(4), 387–420 (1989)
13. Sterman, J.D.: Learning from Evidence in a Complex World. Am. J. Public Health. AJPH.2005.066043 (2006)
14. Tu, T.T.: The Development of a Hybrid Simulation Modelling Approach Based on Agents and Discrete-Event Modelling. PhD thesis, University of Southampton (2008)
15. Turkle, S.: Simulation and Its Discontents. The MIT Press, Cambridge (2009)
16. van Meijgaard, J., Fielding, J.E., Kominski, G.F.: Assessing and forecasting population health: Integrating knowledge and beliefs in a comprehensive framework. Public Health Reports 124 (2009)
17. Viégas, F.B., Wattenberg, M., van Ham, F., Kriss, F., McKeon, M.: ManyEyes: A site for visualization at Internet scale. IEEE Trans on Visualization and Computer Graphics 13, 1121–1128 (2007)

A System Dynamics Approach to Modeling the Sensitivity of Inappropriate Emergency Department Utilization

Joshua G. Behr and Rafael Diaz

Virginia Modeling, Analysis & Simulation Center – VMASC
Old Dominion University, 1030 University Blvd., Suffolk, VA 23435
jbehr@odu.edu, rdiaz@odu.edu

Abstract. Non-urgent Emergency Department utilization has been attributed with increasing congestion in the flow and treatment of patients and, by extension, conditions the quality of care and profitability of the Emergency Department. Interventions designed to divert populations to more appropriate care may be cautiously received by operations managers due to uncertainty about the impact an adopted intervention may have on the two values of congestion and profitability. System Dynamics (SD) modeling and simulation may be used to measure the sensitivity of these two, often-competing, values of congestion and profitability and, thus, provide an additional layer of information designed to inform strategic decision making.

Keywords: System Dynamics, emergency department, hospital, non-urgent, inappropriate utilization, congestion, profitability, low-acuity, primary care physician, clinic, sensitivity analysis.

1 Introduction

Emergency Departments are conceived and administered to meet critical roles in the delivery of emergent- and trauma-based care. In addition, though, Emergency Departments have become the central point of contact for a diverse population seeking care for primary and non-emergent medical conditions. While widely recognized, explanations for this trend vary considerably and include structural forces relating to the affordability and availability of alternate service providers as well as individual-level forces rooted in cultural and social norms. The increased utilization of Emergency Department services for non-emergent conditions has been broadly labeled 'inappropriate usage.' This utilization has been attributed with increasing congestion in the flow and treatment of patients thereby conditioning the quality of delivered care and, by extension, impacting the revenue generated within the Emergency Department. Interventions designed to divert populations to more appropriate care may be cautiously received by operations managers due to uncertainty about the impact an adopted intervention may have on the two values of congestion and profitability. We use System Dynamics (SD) modeling and simulation to measure the sensitivity of these two, often-competing values of congestion and

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 52–61, 2010.

profitability and, thus, provide an additional layer of information designed to inform strategic decision making. Relying on unique data systematically gathered in a major urban trauma center, this research identifies the drivers manifest in the individual decision calculus to seek non-emergent services from the Emergency Department. This knowledge about the underlying logic employed in seeking emergency services allows us, by means of a SD approach, to identify the impact of alternate interventions among subpopulations.

2 Problem

Emergency Departments are frequently utilized to treat non-emergent conditions. Other treatment venues may have the staffing and resource capacity to manage such non-emergent conditions and, in a strict sense, may be viewed as more appropriate venues to address these non-emergent conditions. The failure to utilize venues other than the Emergency Department to manage non-emergent conditions results in a system that is sub-optimal. Specifically, three primary concerns stem from the current behavior of the system: non-emergent Emergency Department utilization is an expensive path to treat conditions relative other treatment venues; utilization of the Emergency Department for non-emergent conditions contributes to congestion; continuity of care is disrupted by non-emergent Emergency Department utilization.

2.1 Primary Motivation

Quality of care and longevity are impacted by venue utilization choice. Regularly seeking Emergency Department treatment for non-emergent conditions prevents the development of a substantive relationship with a medical home. A medical home delivers continuity in care that may result in better management of chronic disease and conditions and, thus, contribute to a decrease in suffering and an increase in longevity.

2.2 Twofold Objective

The objective of this research is twofold. First, we demonstrate the utility of a System Dynamics (SD) modeling approach to better understand the impact of competing interventions in the diversion of patients with non-emergent conditions to more appropriate treatment venues. System Dynamics is an appropriate approach to capture the complexity and feedback inherent in the dispersion of patients across treatment venues. Second, we provide an additional layer of predictive information relating to the potential impact of interventions among patient silos on the two competing values of Emergency Department congestion and profitability; increased knowledge relating to the dynamic nature of the flow of patients among health providers may spur adoption of interventions.

3 Dynamics of Patient Flow within the System

There are four core components that define the flow of patients among treatment venues that constitute the system: structural component, entity component, access

component, perceptual evaluation component. The structural component includes the variety of potential treatment venues for non-emergent conditions within the defined space of the study region. The entity component refers to the population that potentially may be consumers of health services offered at the treatment venues. The access component identifies both barriers and facilitators that condition the behavior of entities to seek the services from one or several treatment venues. The perceptual evaluation component addresses an individual's retrospective evaluation of the potential treatment venues. These latter two components (access and perceptual evaluation) comprise the factors that are weighed by individuals in the decision to seek treatment services.

3.1 Data and Conceptual Problems

The modeling of the flow of non-emergent patients within a system faces several methodological and conceptual hurdles. The validity and generalizability of a System Dynamics model is dependent upon the effective management of these difficulties. The following are among these obstacles:

1) Proper conceptual specification of the high-level causal loop relationships.
2) Informing the model with relevant knowledge and meaningful historical information relating to the behavior of entities under constrained conditions.
3) Capture of the linear and non-linear relationship among system components.

4 System Dynamics Modeling

System Dynamics (SD) has been extensively used as a reliable approach for an ample range of purposes that involve capturing and analyzing the dynamic complex interaction among dissimilar systems. It has been used to analyze the effects of time among a complex set of variables. The modeling strategy used by SD includes representing the structure of the studied system in terms of flows and stocks. As a result, components can be used to articulate the interactions found within subsystems as well as integrated to build a broader, system-wide model that captures dependencies and feedback among the system components. The SD framework acknowledges the complex interactions among many feedback loops, considers linear and nonlinear cause-and-effect, and requires an analytical perspective in which is considered the potential impact of effects on causes. Thus, SD allows for the revelation and quantification of unseen dynamics that might have a relevant affect on subsystems. SD transforms these components of causality into structured difference and differential equations.

Causal diagrams to describe complex systems and analyze the potential impacts of health policy interventions have been suggested as enabling techniques that allow researchers to capture the complexities of the healthcare system [3]. Both background and opportunities for using SD modeling techniques in public health have been analyzed [4]. Aspects related to awareness and support for systems thinking and modeling in the public health environment have been also investigated [5].

4.1 Value of System Dynamics Approach to Model a Health System

System Dynamics has been demonstrated to adequately represent and capture the defining factors that drive the behavior of numerous systems. The approach is flexible in that its utility has been demonstrated in representing the complex behavior of insurgencies, business cycles, and ecological evolution. The variety of structural arrangements and motivations of entities that define a health system make it inherently complex. Due to the scale of the system in terms of both dollars and number of entities served, any structural adjustment within the system has the potential to alter the delivery cost of the healthcare and impact the quality of healthcare, both of which are related to the suffering and longevity of entities within the system. Traditional methods to identify and address disparities or inequities within the healthcare system through targeted interventions, while perhaps successful in the proximate sense, have not been well suited to anticipate the second and third order consequences upon the larger system. The scope and complexity of a healthcare system, as well as a need to better understand the impact of potential adjustments on the broader cost and quality of care, make such a system a meaningful candidate for System Dynamics modeling.

5 Model Parameterization of Non-emergent Utilization

5.1 Informed Model Building

The model is informed and populated by data derived from the three recent studies. The first, a quality of life survey, interviewed 1,100 Hampton Roads residents. The sampling methodology included a stratified random sampling of 11 localities and an instrument that included both open- and closed-ended questions that gauged frequency and logic of treatment venue choice. A second survey, informed by the earlier quality of life survey, focused solely on health conditions and utilization of treatment venues. The sampling methodology included a stratified random sampling of 1,675 Hampton Roads residents and included both open- and closed-ended queries. The third study, concluded in 2009, extensively interviewed 1,500 non-emergent patients as they flowed through the treatment process at a major urban trauma center in Hampton Roads. Using a largely open-ended survey instrument, trained staff and Emergency Department physicians interviewed patients as they progressed through the registration-treatment-discharge process. Data were gathered addressing venue utilization patterns as well a wide-range of factors that are considered within the individual decision calculus to seek services from the Emergency Department for non-emergent conditions.

5.2 Model Parameterization: Treatment Venues

The terminal stocks modeled include the most common treatment venues within the region. These include the following:

- Emergency Department (ED)
- Primary Care Physician (PCP)

- Ambulatory Care Center (ACC)
- Clinic – Non-federally Assisted
- Community Health Center (CHC)
- Health Department
- Community Service Board (CSB)
- Urgent Care/Doc-in-a-Box
- Self Treat

5.3 Model Parameterization: Interventions

The interventions conditioning the flow of non-emergent patients within the system include those most commonly theorized to influence the individual decision calculus, including:

- Insurance Coverage
- Education Related to Venue Access
- Care Coordinator (HFIU)
- Financial Incentives/Disincentives
- Literacy
- Targeted Medical Regimen Compliance
- Expanded Hours
- Pre-consult Triage

5.4 Model Parameterization: The Decision Calculus

Predicting the impact of potential interventions on the flow of patients seeking treatment for non-emergent conditions necessarily requires an understanding of the factors that individuals consider when making the decision to seek medical services. The process of selecting which treatment venue to seek medical services (e.g., ED, PCP, CHC, etc.) involves the consideration of a wide variety of structural, situational, and cultural-historical factors. The combination of factors, and the relative weight given to each factor, forms an individual's decision calculus. Common, or recurring, factors weighted by individuals in deciding when and where to seek treatment include the following:

- Convenience/One-stop Shop
- Second Opinion
- Scheduling/Availability/Access
- Transportation
- Quality/Trust
- Tradition/Familiarity
- Culturally Conducive
- Literacy
- Medical Immediacy
- Cost Mitigation through Anonymity
- Medical Referral

- Institutional Connection
- Traveler/Tourist/Mariner
- Probation/Open Enrollment

5.5 Sensitivity

Sensitivity analysis is the process of altering a given set of assumptions that surrounds the value of parameters in the simulation model such that variation in resultant outputs may be examined [1]. Following the further refinement of the SD model (discussed below), a sensitivity analysis using VenSim software will be performed specifically to assess how responsive the model is to adjustments in the proposed intervention policies. The proposed model necessitates the examination of a large number of parameters, feedback structures, and interactions. The complexity of this model makes it unfeasible to investigate sensitivity simply by contrasting a given number of different simulations. Therefore, two appropriate simulation tools to investigate this behavior include the Monte-Carlo sensitivity approach and the Latin Hypercube approach. Following these approaches will yield sensitivity graphs that exhibit both the simulation traces and confidence bounds related to variable and parameter selection. Monte Carlo simulation performs a large number of simulations with parameter values sampled over a collection of values according to a given probabilistic distribution. Latin Hypercube sampling is a focused form of sensitivity investigation that allows for sensitivity testing more rapidly on large representations. The statistical technique was designed to generate a distribution of reasonable collections of set values from a multidimensional distribution whereby each sample is the only one in each axis-aligned hyper-plane clustering it. Traditionally, when sampling is performed over a function of a given number of variables, the range of each parameter is subdivided into a given number of equally probable intervals [2]. Thus, a number of incumbent points are subsequently located to satisfy the Latin Hypercube conditions. One of the most remarkable advantages of this approach is that random draws can be done one at a time while recollecting previously drawn samples. These approaches allow for behavioral boundaries to be properly understood and the robustness of model-based policies can be rigorously tested. For example, characterizing a set of given scenarios, sensitivity analysis can be performed to determine the effectiveness of the proposed education and insurance interventions on utilization of various treatment venues (e.g., ED, PCP, CHC, etc.).

Further, the impact of a particular intervention (or combination of interventions) is expected to be variable across groups; an intervention is not expected to equally change the behavior of all individuals within the system. That is, the behavior of one group may be more *sensitive* to an intervention relative another group. The classification, or grouping, of individuals may include any number of demographic or context-based characteristics. These grouping are referred to as “silos.” We may quantify the impact (measured in terms of flow to or from stocks) of changes in the intensity of an intervention on these silos. We measure the sensitivity of the following silos to the presence and intensity of the various interventions:

- Employment Status
- Race & Ethnicity

- High Frequency Inappropriate Utilizer (HFIU) Status
- Gender
- Age

6 Conceptualization

6.1 High-Level Causal Loop

The high-level conceptualization of the system involves a central feedback loop which captures the cycle of the population between two states: one where the population entities have a non-emergent need and one where the population entities do *not* have such a need. The "system" is conceptualized as this fundamental loop with multiple intervening elements that create feedback and dependencies. These elements have been generated and informed through our research on the drivers considered in the individual's decision calculus to seek treatment. The high-level causal loop does not include specification of treatment venues but, rather, is focused on the dynamical process that moves persons from a non-treated state to a treated state. In addition, this high-level conceptualization further includes broad concepts that condition the decision behavior of entities within the system. These concepts may be classified under the headings of social-cultural, economic-financial, and structural determinants that may either enhance or frustrate the movement of the population from one state to the next.

7 System Dynamics Model Formulation

The modeling of the flow of populations with non-emergent conditions among treatment venues within a region and the sensitivity of these flows to interventions are the central foci of this research. The initial development of our causal loop mapping required the identification of the social-cultural, economic-financial, and structural components that contribute to the probability of seeking treatment from one venue relative another and, thus, diagrams the movement of populations from the state of untreated condition to treated condition. The formulation of a System Dynamics model represents a conceptual and methodological evolution beyond the broadly defined causal relationships found in the high-level mapping. Correctly, the SD model incorporates our knowledge of causal relations as broadly understood in the causal loop mapping, but the SD model further includes the modeling of the various treatment venues as well as interventions that may either facilitate or frustrate the diversion of populations from one venue towards another. Existing empirical efforts guided our understanding of the many complex and interacting factors that individuals weigh in determining when and where to seek treatment. These earlier efforts have assisted us in better understanding the social and structural environment in which the decision to seek services is made. This understanding has been incorporated into the initial model version detailed below.

The Vensim simulation software is used to model the complex system. The simulation model is validated by manually tracing the mechanics of system by

traceable reference models that reflect current behavior and match well-known structures of Bass Diffusion models. Each simulation run represents the circulation of patients within the system for 3 years (approximately 1,000 time intervals). In addition, as in machine learning, a percentage of the survey data will be used for parameterization while other proportion will be employed for testing.

7.1 Stock and Flow

The model found in Figure 1 illustrates the fundamental organizing logic of the larger, more detailed stock and flow model which, due to the multiple treatment venues and multiple interventions, contains several layers. In this figure, though, we illustrate the logic by presenting just two of the treatment venues (Primary Care Physician and

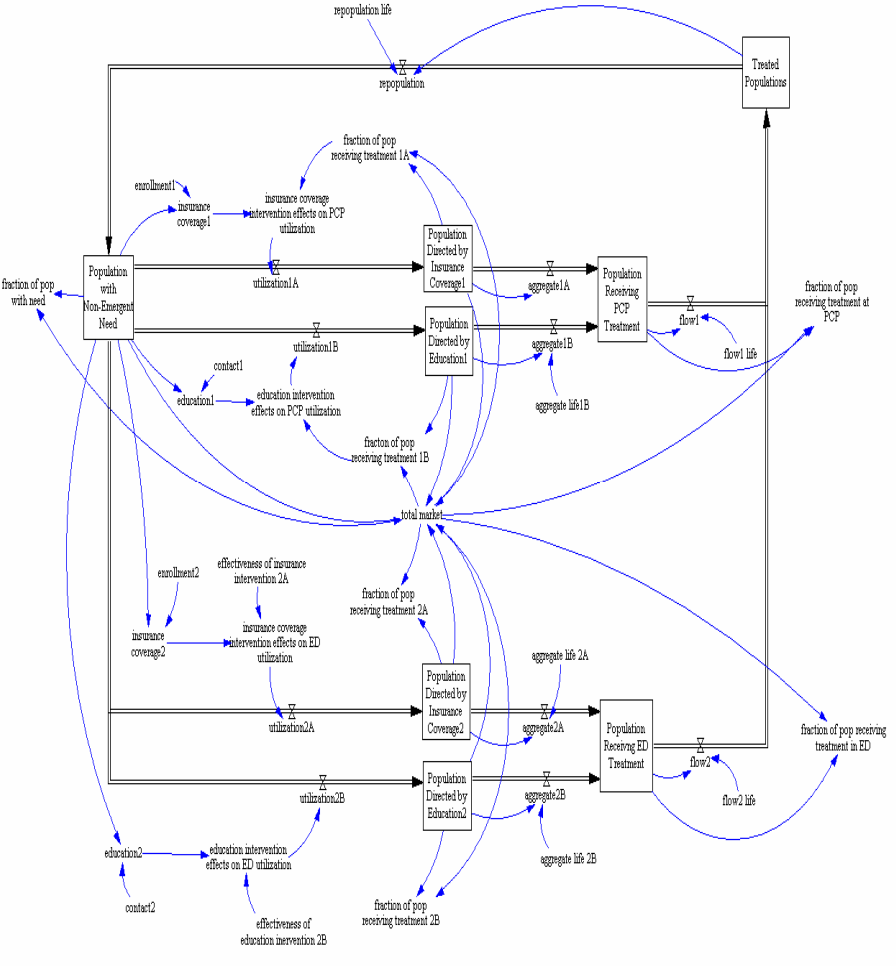


Fig. 1. Simplified Organizing Logic of Stock and Flow. This shows a system containing two treatment venues (PCP and ED) and two interventions (insurance coverage and education).

Emergency Department) and just two interventions (insurance coverage and education). The system model encompasses stocks and flow. The stocks represent the accumulation of a population at various stages in the cycle between the state of having a non-emergent condition in need of treatment (Population with Non-Emergent Need) and the state of no longer having that conditions (Treated Population). In the transfer from having a non-emergent medical condition to being treated, the population must travel through either one of the two treatment venues. The flows are represented by valves that regulate the transfer rate of the population from one stock to the next. In this illustration, the flow of population with a non-emergent need to the PCP treatment venue may be regulated by the interventions 'insurance coverage' and 'education relating to venue access.' Likewise, the rate of flow towards the Emergency Department is also regulated by the variability in both the insurance and education interventions. Note that the full model (not shown here) contains nine treatment venues; the flow of the population towards each of these venues is regulated by eight potential interventions.

8 Overview of Research Approach

The goals of this research are informed by awareness that quality of care and longevity may be either enhanced or frustrated by interactions among the structural, economic, and social aspects at work within a health system. It is recognized that Emergency Departments are often the venue in which are sought the treatment of non-emergent conditions. The collateral repercussions associated with this inappropriate utilization include the inefficient expenditure of resources to treat low-acuity patients, increased Emergency Department congestion, and decreased propensity to have a substantive relationship with a primary care home.

The central aim of this research is to demonstrate the application of System Dynamics as a meaningful approach to modeling the flow of entities with non-emergent conditions among various treatment venues and, secondly, to better understand the sensitivity of competing interventions in the diversion of low-acuity patients away from the Emergency Department. Due to the relative complexity of a system that includes multiple feedback loops, a System Dynamics approach is an appropriate analytical tool. The specification of the model has been realistically informed by data systematically gathered in a major urban trauma center as well as several in-depth surveys of consumers of the region's health services. This earlier research has allowed the identification of drivers manifest in the individual decision calculus to seek non-emergent services from the Emergency Department. Knowledge from causal loop mapping and the development of an informed modeling process will allow for the production of meaningful sensitivity analyses that will provide an additional layer of predictive information. This information may allow policy makers to assess the relative or combined impacts of potential interventions among defined patient silos. In this respect, low-performance interventions may be marginalized in favor of high return interventions.

References

1. Taha, H.: Operations Research, an Introduction, 7th edn. Prentice Hall, New Jersey (2002)
2. Robert, C., Casella, G.H.: Monte Carlo Statistical Methods. Springer, New York (1999)
3. Joffe, M., Mindell, J.: Complex Causal Process Diagrams for Analyzing the Health Impacts of Policy Intervention. Am. J. Publ. Health 96, 473–479 (2006)
4. Homer, J., Hirsch, G.: System Dynamics Modeling for Public Health: Background and Opportunities. Am. J. Publ. Health 96, 452 (2006)
5. William, M.T., Derek, A.C., Bobby, M., Richard, S.G., Scott, J.L.: Practical Challenges of Systems Thinking and Modeling in Public Health. Am. J. Publ. Health 96, 538 (2006)

Using Social Network Analysis for Spam Detection

Dave DeBarr and Harry Wechsler

George Mason University, Department of Computer Science,
Fairfax, VA 22030-4444
{ddebarr,wechsler}@gmu.edu

Abstract. Content filtering is a popular approach to spam detection. It focuses on analysis of the message content to identify spam. In this paper, we evaluate the use of social network analysis measures to improve the performance of a content filtering model. By measuring the degree centrality of message transfer agents, we observed performance improvements for spam detection in repeated experiments; e.g. a 70% increase in the proportion of spam detected with a false positive rate of 0.1%. We were also able to use anomaly detection to identify mislabeled messages in a publicly available spam data set. Messages claiming unusually long paths between the sender's message transfer agent and the recipient's message transfer agent turned out to be spam.

Keywords: Social Network Analysis, Degree Centrality, Spam Detection.

1 Introduction

Unwanted email, also known as spam, affects all email users on a routine basis. Some spam messages get delivered to the user's inbox. The content of these spam messages includes advertising for pharmaceuticals, jewelry, electronics, software, loans, stocks, gambling, weight loss, and pornography; as well as malware and phishing (identify theft) lures. Wanted email, also known as ham, may also be incorrectly junked as spam. And all users experience delays in message delivery as the email system works to determine which messages are spam and which messages are ham.

Content filters parse incoming messages looking for the presence of particular features in the message content to determine whether a message is ham or spam. For example the term "alert" may occur more frequently in ham (news alerts), while the term "http" may occur more frequently in spam (links). There are, of course, no magic terms that allow a content filtering system to perfectly separate spam and ham. Even if there were, spammers would quickly find these terms and adapt; i.e. start using the ham terms and stop using (or obfuscate) the spam terms.

2 Social Network Analysis

Social Network Analysis focuses on measuring various aspects of entities and the relationships between them. This includes identifying "central" nodes within a

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 62–69, 2010.

network and determining the distance between nodes. The entities of interest are often people and the relationships of interest are often social interactions, but the concepts are easily transferred to computers and the email connections between them.

3 Spam Detection

Many spam filtering systems augment their content filtering system with a block list, for known spam senders. Of course, the spammers have adapted to this system as well. By using malware to infect computer systems, spammers can then use the infected computers to send spam for them. These infected systems are called proxies, because they send the spam on behalf of the spammer. In order to hide the source of these messages, the proxies route the spam messages through "open relays." An "open relay" is a Message Transfer Agent (Simple Mail Transfer Protocol server) that will allow any Internet user to send email through it. Figure 1 shows a typical configuration between sender and recipient. For example, when user@example.edu sends an email to ddebarr@gmu.edu, user sends the message to mta.example.edu and mta.example.edu forwards the message to mta.gmu.edu. The focus of this work is to evaluate whether the degree centrality of the sender MTA is useful for helping to distinguish spam from ham, assuming that "open relays" are likely to have many more senders than other sender MTAs.

Fig. 1. Simple Network Illustrating Connections Between a Sender and Recipient

The connections between computers are recorded in message "headers", a portion of the message typically not seen by the sender or the recipient. The format of these headers vary from host to host, but they typically have the following form:

Received: from sender_mta_name ([sender_mta_address]) by recipient_mta_name

The sender_mta_address is recorded as an Internet Protocol (IP) address; e.g. "10.1.1.101", the network address of the sending computer. A typical email address will contain two "received" headers. The first recorded by the sender MTA, and the second recorded by the recipient MTA.

4 Features and Representation

Our proposed spam detection filter is trained with both content filtering features and our proposed social network analysis features. The content of the messages is divided into tokens, and the term frequency and inverse document frequency of the tokens are used to train the spam detection filter. For example, if the term "alert" is found 3 times in the content of a particular message and it occurs in 100 out of 10000 messages in the training corpus, this token is represented by the number 6 [3 * $\log_{10}(10000/100)$]. The token values for each message are then divided by the square root of the sum of squared token values, in order to mitigate the effect of differences in message length.

The proposed social network analysis features include the degree centrality of the sender's MTA as well as the path length between sender and recipient. To help identify possible open relays, we used the degree centrality of the sender's MTA. This was computed as the indegree of the sender MTA. For example, the following "received" headers form a link between "192.168.1.201" and "10.1.1.101":

Received: from sender ([192.168.1.201]) by sender_mta
Received: from sender_mta ([10.1.1.101]) by recipient_mta

To assess whether the sender_mta of an incoming message is an open relay, we used the "number of distinct sender subnets for the sender_mta during the previous week of observed email traffic" (indegree) as a feature for training the spam detection filter. The first 3 components of the IP address were used as a subnet identifier. For example, if sender_mta received email from 1.2.3.4, 1.2.3.5, and 6.7.8.9, the indegree would be recorded as 2 (for subnets 1.2.3.* and 6.7.8.*). We also used the number of "received" headers as a training feature, assuming that longer paths from sender to recipient are more likely to be spam related.

5 Classification

Machine learning algorithms are often used to train a computer to recognize patterns within data. For spam filtering, we provide the computer with examples of both spam and ham so that it can learn to distinguish between the two message categories. The machine learning algorithm known as LogitBoost is among the best in terms of constructing an effective spam detection filter. A trained model consists of a set of conditions and weights that are used to evaluate each message. For example, a condition might be "indegree for sender_mta > [learned threshold]" and the associated weight might be some positive value. Because LogitBoost is an additive logistic

regression algorithm, these weights are used directly to estimate the probability that a message is spam:

$$Probability(Spam) = \frac{1}{1 + \exp(-\sum weights)}$$

If the sum of weights is 0, then the Probability(Spam) is 50%; if the sum of weights is negative, then the Probability(Spam) < 50%; and if the sum of weights is positive, then the Probability(Spam) > 50%.

6 Experiments

The publicly available corpus from the "Spam Track" of the 2007 Text REtrieval Conference (TREC) was used for our experiments. It consists of 75,419 messages received by 2 email servers at the University of Waterloo, from April 8 to July 6 2007. Messages received from Monday April 16 to Sunday April 22 were used for training, while messages received from April 23 to July 6 were used for testing. Table 1 shows the number of messages used for training and testing.

Table 1. Train/Test Message Counts

	Spam	Ham
Train	5,162	2,171
Test	37,116	21,120

Because obtaining ham/spam labels for messages requires manual intervention, only a 1% sample of the training set was actually used for constructing a LogitBoost model. The training set was randomly divided into 5 subsets, then each subset of messages was clustered based on TF/IDF (term frequency, inverse document frequency) representation. The Partitioning Around Medoids (PAM) was used for clustering, and a total of 73 cluster prototypes from each of the 5 subsets were used to train LogitBoost models. Cross validation was used to optimize the parameters of the LogitBoost learning algorithm. The scaling coefficient was set to 0.5 and the number of iterations was set to 40 for each model. Each trained model was then evaluated using the entire test set.

Performance was evaluated using the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve is generated by changing the spam detection threshold from probability=1 to probability=0, and plotting the False Positive (FP) rate on the horizontal axis and the True Positive (TP) rate on the vertical axis. The FP rate is simply the percentage of ham messages predicted to be spam, while the TP rate is simply the percentage of spam messages predicted to be spam. The area under the ROC curve is the probability that the "probability of spam" assigned to a randomly selected spam message will be larger than the "probability of spam" assigned to a randomly selected ham message. Table 2 compares the

performance of the models with and without Social Network Analysis (SNA) features for each of the 5 training subsets. Each row compares the performance of the "Content Features Only" approach to the "Content + SNA Features" approach, using precisely the same training messages from each subset.

Table 2. Comparisons of Area Under ROC Curves

	Content Features Only	Content + SNA Features
Subset 1	93.2%	94.9%
Subset 2	94.3%	97.1%
Subset 3	97.2%	98.4%
Subset 4	95.9%	96.6%
Subset 5	95.0%	95.5%

Figure 2 shows the difference in the ROC curves for subset 3. The solid line shows the performance of the "Content Features Only" model, while the dotted line shows the performance of the "Content + SNA Features" model. Figure 3 zooms in on the upper left hand corner of figure 2, highlighting the difference in performance when low false positive rates are required. Table 3 compares the percentage of spam detected for selected false positive thresholds. For the lower false positive rates, the difference in performance is very significant. For example, with a false positive rate of 1 in 1000 ham messages (incorrectly identified as spam), the "Content + SNA Features" model detects 1.7 times as many spam messages.

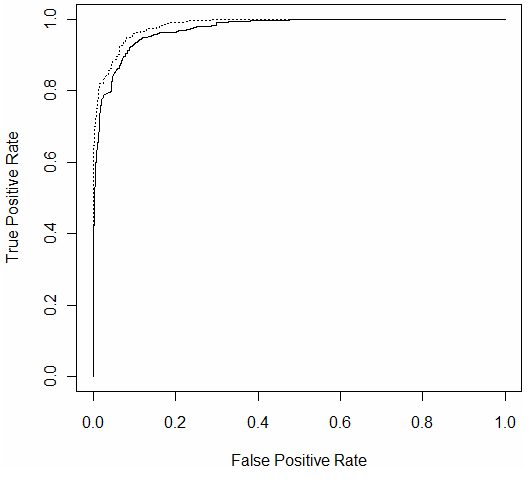


Fig. 2. Example ROC Curves

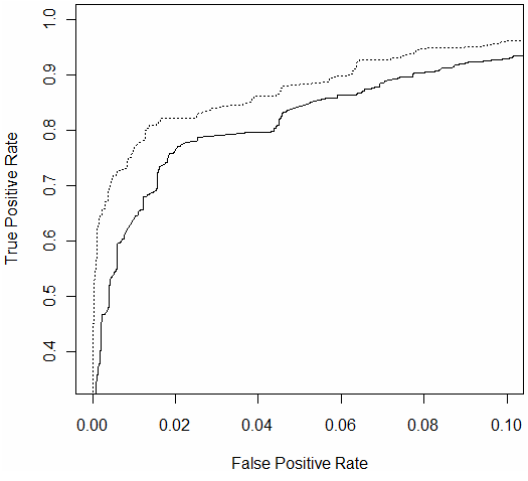


Fig. 3. Region of Interest for ROC Curves

Table 3. Comparison of Spam Detection Rates

False Positive Rate	Content Features Only	Content + SNA Features
0.1%	35.73%	60.64%
0.2%	44.37%	64.56%
0.3%	46.66%	66.96%
0.4%	51.52%	69.07%
0.5%	53.78%	71.79%
1.0%	62.26%	76.78%
2.0%	76.05%	82.03%
3.0%	78.72%	83.90%
4.0%	79.55%	86.12%
5.0%	83.57%	88.00%

The "Content + SNA Features" model shown in figure 3 assigns weight +2.0 if the indegree of the sender's MTA was at least 15 (i.e. the sender's MTA was used by 15 or more subnets within the previous week). The weight -1.55 was assigned otherwise.

While the path length (number of "received" headers) was not emphasized in the LogitBoost models, it was useful for identifying test set messages that were incorrectly labeled. For example, there were 14 messages that had 12 or more "received" headers; but only 2 of them were labeled as spam. A review of the messages labeled as "ham" shows that these messages have been mislabeled: 36527, 37057, 39208, 39846, 39904, 40652, 41486, 43113, 44498, 45275, 46296, and 47221.

Figure 4 shows the content of message 43113. It is an advertisement for Cialis and Viagra that has been sent to an email list. This emphasizes the difficulty of obtaining accurate labels for messages.

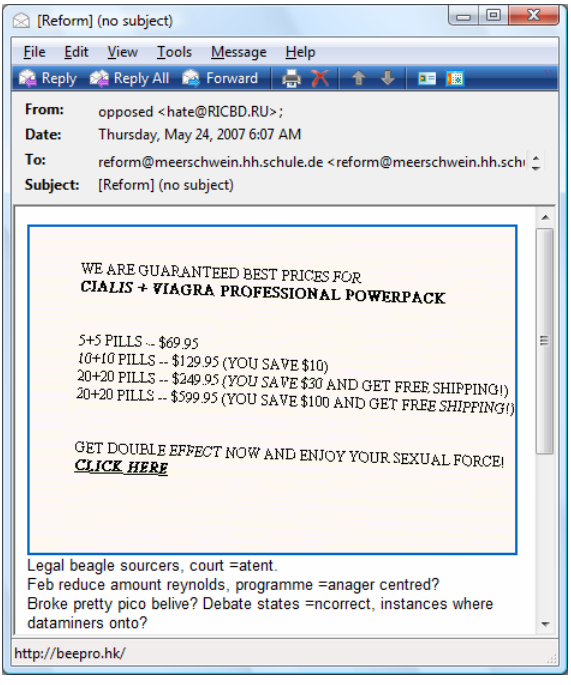


Fig. 4. A Spam Message Mislabeled as Ham

7 Conclusions

The use of social network analysis features significantly improved the performance of a content filter for spam detection. The use of indegree appears useful for identifying open relays, mail servers that will accept email traffic from anyone (allowing spammers to hide their identity). It's possible the open relays that predominantly send spam could be added to a network address block list, thus avoiding the overhead associated with content filtering. Future work includes evaluating the use of more complicated social network analysis features to characterize the network connectivity of mail transfer agents along the path between sender and recipient.

References

1. Calais, P., Guedes, D., Meria Jr., W., Hoepers, C., Chaves, M., Steding-Jessen, K.: Spamming Chains: A New Way of Understanding Spammer Behavior. In: Proceedings of the 6th Conference on E-Mail and Anti-Spam (2009), http://www.ceas.cc/papers-2009/ceas2009-paper-23.pdf

2. Cormack, G.V.: TREC 2007 Spam Track Overview. NIST Special Publication 500-274. In: The 16th Text REtrieval Conference, TREC (2007), http://trec.nist.gov/pubs/trec16/papers/SPAM.OVERVIEW16.pdf
3. Crocker, H.D.: Standard for the Format of ARPA Internet Text Messages. ARPANET Request for Comments (RFC) No. 822 (August 1982), http://www.ietf.org/rfc/rfc0822.txt
4. Fawcett, T.: An Introduction to ROC Analysis. Pattern Recognition Letters 27(8), 861–874 (2006)
5. Freeman, L.C.: Centrality in Social Networks: Concept Clarification. Social Networks 1(3), 215–239 (1979)
6. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: A Statistical View of Boosting. Annals of Statistics 28(2), 337–407 (2000)
7. Kaufman, L., Rousseeuw, P.J.: Partitioning Around Medoids. In: Finding Groups in Data, pp. 68–125. Wiley-Interscience, Hoboken (2005)
8. Manning, C.D., Raghavan, P., Schutze, H.S.: Term Weighting, and the Vector Space Model. In: Introduction to Information Retrieval, pp. 109–133. Cambridge University Press, Cambridge (2008), http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
9. TREC 2007 Public Spam Corpus, http://plg.uwaterloo.ca/~gvcormac/treccorpus07/

Literature Search through Mixed-Membership Community Discovery

Tina Eliassi-Rad and Keith Henderson

Lawrence Livermore National Laboratory,
P.O. Box 808, L-560, Livermore, CA 94551, USA
{eliassi,keith}@llnl.gov

Abstract. We introduce a new approach to literature search that is based on finding mixed-membership communities on an augmented co-authorship graph (ACA) with a scalable generative model. An ACA graph contains two types of edges: (1) coauthorship links and (2) links between researchers with substantial expertise overlap. Our solution eliminates the biases introduced by either looking at citations of a paper or doing a Web search. A case study on PubMed shows the benefits of our approach.

Keywords: Literature search, mixed-membership, community discovery.

1 Introduction

Given a research topic (e.g. reconstruction of the 1918 influenza virus) and a couple of seminal papers on that topic (e.g. [1] and [2]), how do we find authors who are conducting similar research? Traditional solutions to this problem include looking at the citations in the seminal papers and/or conducting Web searches on keywords associated with the chosen topic. Both of these commonly used solutions have biases that limit their effectiveness. For example, looking only at the citations of a paper provides a partial view of the domain (namely, the ones provided by the authors). Doing a Web search on keywords neglects the wealth of information embedded in social networks (such as co-authorship graphs).

In this work, we propose a new approach to the literature search problem that is based on finding mixed-membership communities on an augmented co-authorship (ACA) graph. We construct an ACA graph by fusing the information from a bipartite expertise-by-author graph into a co-authorship graph, which produces a denser and more structured version of the original co-authorship graph.

For the mixed-membership community discovery algorithm, we utilize our Latent Dirichlet Allocation for Graphs (LDA-G) [3]. LDA-G is a scalable generative model that adapts the Latent Dirichlet Allocation (LDA) [4] topic-modeling algorithm for use in graphs rather than text corpora. A simple post-analysis of LDA-G's communities provides a ranking of the most similar authors. In our

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 70–78, 2010.

experiments on PubMed[1] data, LDA-G produces better solutions than when it is applied to regular co-authorship graphs or bipartite expertise-by-author graphs. In addition to our qualitative results, we provide quantitative results based on link prediction performance of LDA-G's posterior estimate.

2 Mixed-Membership Community Discovery

We utilize our scalable generative LDA-G model [3] to find mixed-membership communities in large graphs. In this context, "mixed membership" means that nodes can belong to multiple communities with varying probabilities. Given a graph, LDA-G models each source node in the graph as a multinomial distribution over some set of communities Z. The cardinality of Z is unknown a priori and is learned via Bayesian inference from a Dirichlet prior. In LDA-G, each source node generates a series of communities from its multinomial; and each community is a multinomial distribution over target nodes. Any time a community is generated by a source node, that community generates a target node from its distribution. The distributions over source-node to community and community to target-node are learned using MCMC techniques (e.g., we use Gibbs sampling). To simplify inference, it is assumed that the behaviors of a node as a source-node and as a target-node are probabilistically independent. The generative model for LDA-G is as follows:

$$t_i \mid z_i, \varphi^{(z_i)} \sim Discrete(\varphi^{(z_i)}) \tag{1}$$

$$\varphi \sim Dirichlet(\beta) \tag{2}$$

$$z_i \mid \theta^{s_i} \sim Discrete(\theta^{s_i}) \tag{3}$$

$$\theta \sim Dirichlet(\alpha) \tag{4}$$

Equations 1 and 3 are the multinomial distributions from communities z to target-nodes t and from source-nodes s to communities z, respectively. Equation 2 and 4 are the prior distributions on target nodes with hyperparameter β and on communities with hyperparameter α, respectively.

Unlike most approaches to community discovery, LDA-G only requires present links (i.e., non-zero entries in the adjacency matrix). This property helps its runtime and space complexities. It has $O(NKM)$ runtime and $O(N(K+M))$ space complexity, where N is the number of nodes in the graph, K is the number of communities ($K << N$), and M is the average vertex degree in the graph ($M << N$).

We define link-prediction performance as a quantitative way of measuring the effectiveness of LDA-G in factoring a graph into communities. In particular, we compute area under ROC curve on the task of predicting links from held-out test-sets based on the (posterior) probability of a link between two nodes s and t. Equation 5 defines this probability.

[1] PubMed is a repository containing millions of citations from biomedical articles (http://www.pubmedcentral.nih.gov/).

$$p(s \rightarrow t) = \sum_{z \in Z} p(z|s)p(t|z) \quad (5)$$

There are a few scalable generative models that find community structure in graphs [3,5,6,7,8,9,10]; most of them extend LDA. The simplest adaptions are LDA-G and SSN-LDA [9]. There are also derivations that find communities in social networks with weighted links [8] or with categorical attributes on links [5]; find communities in textual attributes and relations [6,10]; and find communities in dynamic (time-evolving) graphs [7].

3 Augmented Co-authorship (ACA) Graph

An ACA graph is a denser and more structured version of a co-authorship graph. We construct an ACA graph by fusing the information from a bipartite expertise-by-author graph into a standard co-authorship graph. We advocate a two-step approach for the fusion. First, we prune the expertise-by-author multigraph[2] by removing links that appear less than r times (i.e., links with weights ¡ r). We pick the threshold r based on the distribution of weights on the expertise-by-author links. This step effectively removes "noisy" and "random" links from the expertise-by-author graph. Second, in the co-authorship graph, we add a link between any pair of authors that share an expertise in the pruned expertise-by-author graph. Hence, the ACA graph not only contains co-authorship links but also links indicating that two authors have substantial overlap in their expertise.

The intuition behind ACA graphs is that fusing data from different sources, especially introducing more structured data into less structured data, can be quite valuable during analysis. Figure 1 depicts the adjacency matrices for an expertise-by-author graph and a co-authorship graph extracted from PubMed and their associated ACA graph. Table 1 presents the basic statistics of these data graphs. The expertise nodes were extracted based on term frequency in PubMed abstracts. A link exists from an expertise node x to an author node y for every paper in which y is an author and x is a term appearing in the paper's abstract.

To generate the ACA graph, we need to select a threshold r to remove "noisy" and "random" links from the expertise-by-author graph. In other words, we want only the expertise-by-author relationships that are "significant" (because we are going to generate implicit co-authorship links between authors with significant expertise overlap). Figure 2 depicts the distribution of edge weights on our expertise-by-author graph. We used a threshold r of 12 for in our case-study. The probability of an edge weight being greater than or equal to 12 is 1.3%; hence, the links associated these weights do not exist because of chance or noise. Our threshold generated a pruned expertise-by-author graph with 1,310 authors (3.5% of the original authors), 117 expertise terms, and 1,565 links (1.3% of the original expertise-by-author links).

[2] Each time an author publishes in a given expertise, a link is created in the bipartite expertise-by-author graph.

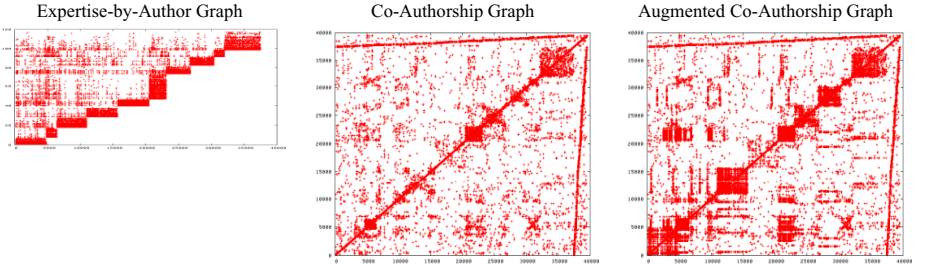


Fig. 1. Adjacency matrices for an expertise-by-author graph and a co-authorship graph extracted from PubMed and their augmented co-authorship graph. (The expertise-by-author graph's adjacency matrix is sorted by the order in which each author's expertise was added to the graph).

Table 1. Basic statistics on our PubMed-extracted data graphs (LCC is short for the largest connected component)

Data Graph	# of Nodes	# of Links	# of Components	% of Nodes in LCC	% of Links in LCC
Expertise-by-Author	117 (E) 37,483 (A)	119,443	1	100%	100%
Co-Authorship	37,227 (A)	143,364	4,556	23.54%	35.82%
Augmented Co-Authorship	37,227 (A)	339,644	4,389	30.40%	46.89%

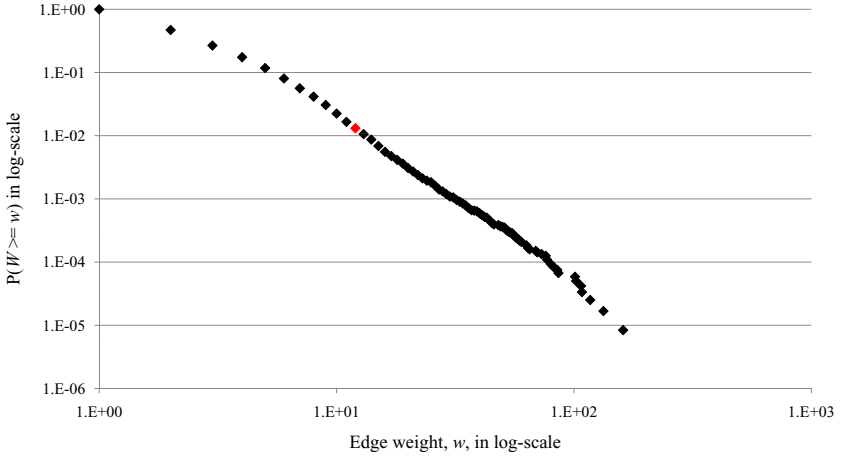


Fig. 2. Cumulative edge-weight distribution for the expertise-by-author graph. The chosen threshold ($r = 12$) is colored in red.

4 Experiments

Given the graphs depicted in Figure 1, we find mixed-membership communities on them with LDA-G, and then use the community structures to find authors that are performing similar research to authors of [1] and [2] (i.e. research on the reconstruction of the 1918 influenza virus). For the latter, we look for communities that are common between authors of [1] and [2]. In all three graphs, LDA-G finds communities that are common between authors of [1] and [2]. In the expertise-by-author graph, LDA-G finds four common communities (see Figure 3, top plot, communities #10, #20, #32, and #33). In the co-authorship graph, it uncovers one common community (see Figure 3, middle plot, community #6). In the ACA graph, it discovers three common communities (see Figure 3, bottom plot, community #3, #14, and #16). It is only in the ACA graph that LDA-G is able to find a common community with a significant overlap - specifically, 47% of authors of [2] and 30% of authors of [1] fall into community #3 of the ACA graph. Further inspection of this community reveals authors that have both similar co-authorship patterns and expertise as authors of [1] and [2]. We depict these authors and their expertise in the Figure 4. These authors have the highest percentage of membership in community #3 of the ACA graph, which is shared among authors of [1] and [2]. None of these authors were cited in [1] or [2]. We showed our findings to domain experts and received validation from them that we had indeed found the relevant researchers.

Figure 5 depicts the overlap in the expertise terms for authors of [1] and [2]. Even though both papers are on the reconstruction of 1918 influenza virus, the probability distribution on the expertise terms of their major author groups is different. In other words, simply conducting a keyword search on the (expertise) terms will not be sufficient for finding authors who are conducting similar research. LDA-G is able to effectively factor out a graph's community structure. Figure 6 plots the adjacency matrix and the resultant community-sorted matrix for the ACA graph. As it can be seen, LDA-G discovers nicely separated block-structure.

On link prediction, LDA-G's posterior estimates on the aforementioned graphs produce average area under the ROC curve (AUC) values of at least 0.918. (Recall that an AUC of 0.5 is a random guess.) Table 2 lists the AUC values on the PubMed graphs (averaged over 5 trials). As is standard in machine learning, we repeatedly divide the dataset into training and test sets, build a model on the training set, and examine its performance with respect to the chosen metric (e.g., AUC) on the held-out test-set. In particular, we use stratified random sampling to hold-out 1000 links from each graph. The remaining links are used to discover the latent communities. Then, the superiority of the discovery community structure is checked based on how well it predicts the existence of the held-out links as described in Equation 5. In [3,5], we present a comparative study on link prediction results (on these graphs) between LDA-G and five other community discovery approaches (including *Fast Modularity*, *Cross Associations*, and *Infinite Relational Models*). LDA-G's link prediction results either outperform or are competitive with the best performer.

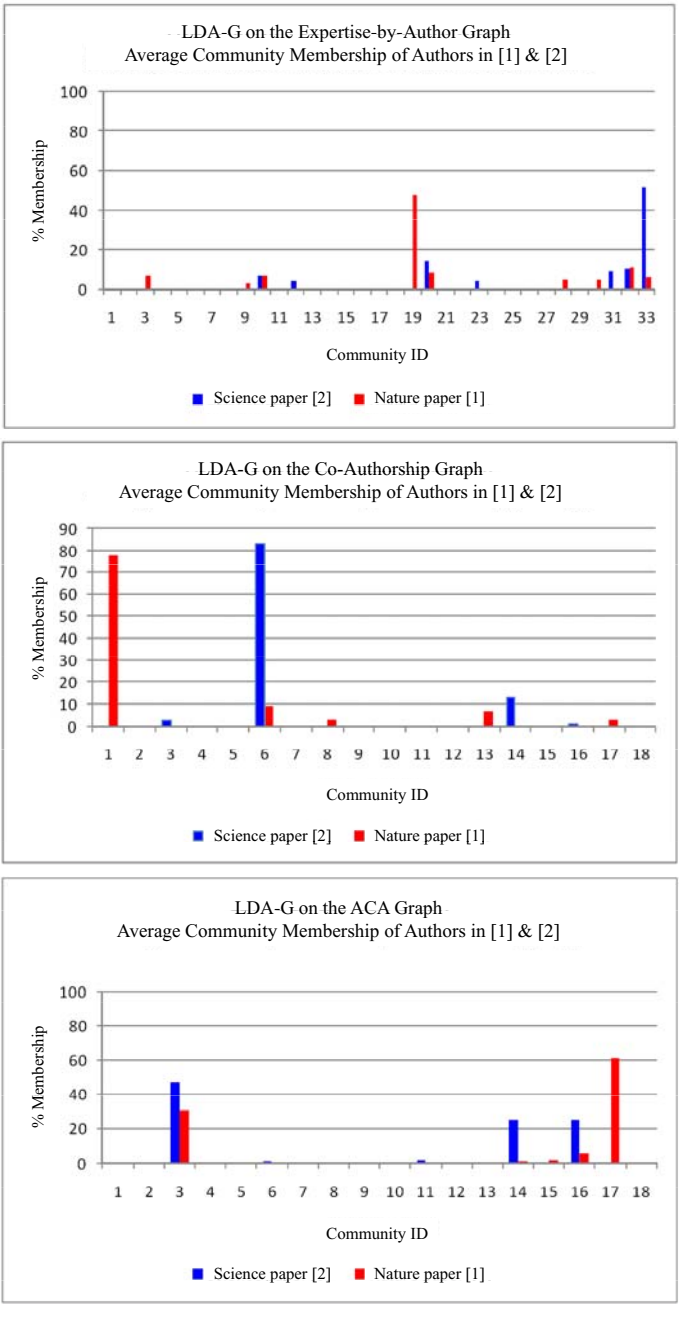


Fig. 3. Average community membership of authors of [1] and [2]. Only in the ACA graph do we find a common community (#3) with significant overlap between the authors of [1] and [2].

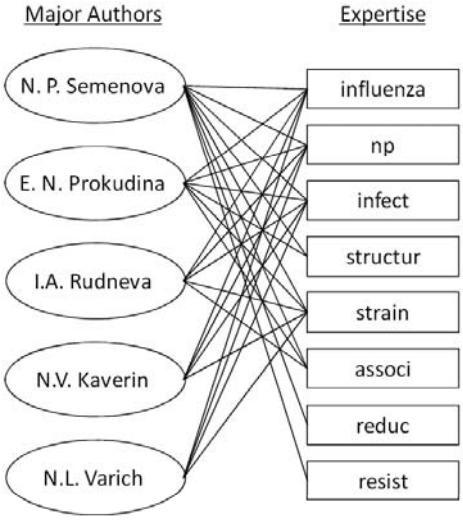


Fig. 4. Authors with the highest percentage of membership in community #3 of the ACA graph. These five authors and the authors of [1] and [2] share similar expertise and have topologically similar co-authorship neighborhoods.

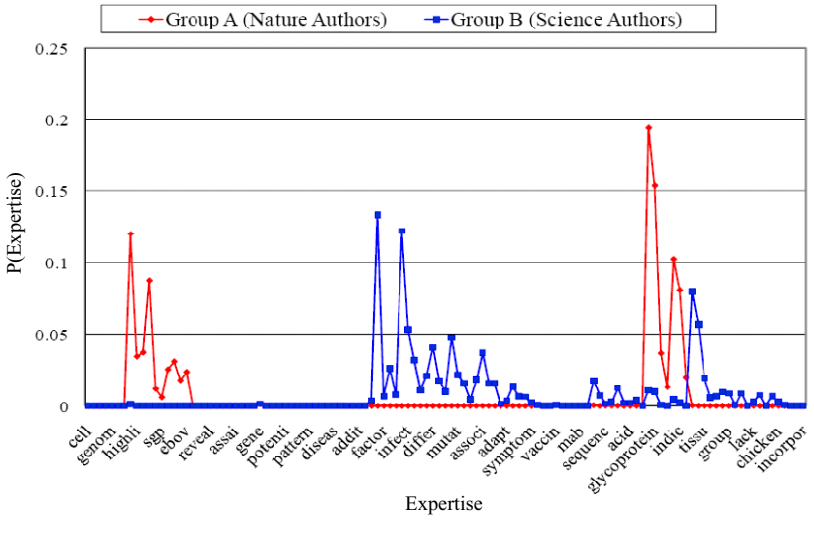


Fig. 5. LDA-G's qualitative results on the expertise-by-author graph. Plot shows the probability of expertise terms for major author groups of [1] in red and [2] in blue. Even though both papers are on reconstruction of the 1918 flu virus the authors' expertise terms does not overlap as much as expected.

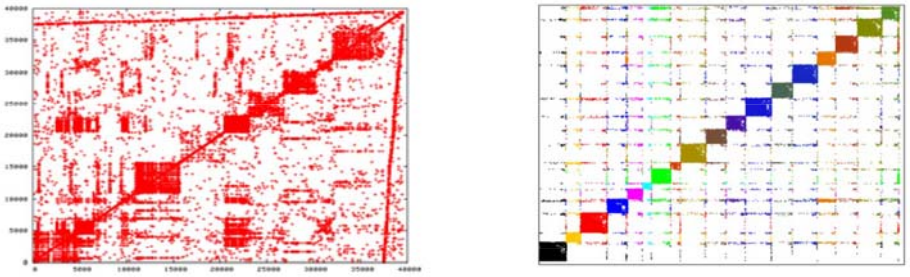


Fig. 6. The ACA Graph: Its adjacency matrix and its community-sorted matrix

Table 2. AUC values on link prediction averaged over 5 trials (default value is 0.5)

Data Graph	LDA-G's Posterior Estimates
Expertise-by-Author	0.955
Co-Authorship	0.925
Augmented Co-Authorship	0.918

5 Conclusions

We describe a new approach to the literature search problem, which involves finding mixed membership communities on augmented co-authorship (ACA) graphs with LDA-G (a scalable generative model). An ACA graph contains not only co-authorship links but also links between researchers with substantial expertise overlap. We evaluate our approach qualitatively and quantitatively on data from PubMed and present a successful case study.

Future work involves utilizing the distributed-inference, temporal version of our LDA-G on larger-scale dynamic graphs in order to track the delineation of scientific domains/communities.

Acknowledgements. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48 and No. DE-AC52-07NA27344.

References

1. Kobasa, D., Jones, S.M., Shinya, K., Kash, J.C., Copps, J., Ebihara, H., Hatta, Y., Kim, J.H., Halfmann, P., Hatta, M., Feldmann, F., Alimonti, J.B., Fernando, L., Li, Y., Katze, M.G., Feldmann, H., Kawaoka, Y.: Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. Nature 445(7125), 319–323 (2007)
2. Tumpey, T.M., Basler, C.F., Aguilar, P.V., Zeng, H., Solórzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenberger, J.K., Palese, P., García-Sastre, A.: Characterization of the reconstructed 1918 spanish influenza pandemic virus. Science 310(5745), 77–80 (2005)

3. Henderson, K., Eliassi-Rad, T.: Applying latent Dirichlet allocation to group discovery in large graphs. In: Proceedings of the 24th Annual ACM Symposium on Applied Computing (SAC 2009), Honolulu, HI, pp. 1456–1461 (2009)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Henderson, K., Eliassi-Rad, T., Papadimitriou, S., Faloutsos, C.: Hcdf: A hybrid community discovery framework. In: Proceedings of the 2010 SIAM Conference on Data Mining (SDM 2010), Columbus, OH (2010)
6. Li, H., Nie, Z., Lee, W.C., Giles, C.L., Wen, J.R.: Scalable community discovery on textual data with relations. In: Proceeding of the 17th ACM conference on Information and Knowledge Management (CIKM 2008), Napa Valley, CA, pp. 1203–1212 (2008)
7. Miller, K.T., Eliassi-Rad, T.: Continuous time group discovery in dynamic graphs. In: Notes of the 2009 NIPS Workshop on Analyzing Networks and Learning with Graphs, Whistler, BC, Canada (2009)
8. Zhang, H., Giles, C.L., Foley, H.C., Yen, J.: Probabilistic community discovery using hierarchical latent Gaussian mixture model. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI 2007), Vancouver, BC, Canada, pp. 663–668 (2007)
9. Zhang, H., Qiu, B., Giles, C.L., Foley, H.C., Yen, J.: An LDA-based community structure discovery approach for large-scale social networks. In: Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI 2007), New Brunswick, NJ, pp. 200–207 (2007)
10. Zhou, D., Manavoglu, E., Li, J., Giles, C.L., Zha, H.: Probabilistic models for discovering e-communities. In: Proceedings of the 15th international conference on World Wide Web (WWW 2006), Edinburgh, Scotland, pp. 173–182 (2006)

Predictability and Prediction for an Experimental Cultural Market

Richard Colbaugh[1], Kristin Glass[2], and Paul Ormerod[3]

[1] Sandia National Laboratories, Albuquerque, NM USA
rcolbau@sandia.gov
[2] New Mexico Tech, Socorro, NM USA
kglass@icasa.nmt.edu
[3] Volterra Consulting, London, UK
pormerod@volterra.co.uk

Abstract. Individuals are often influenced by the behavior of others, for instance because they wish to obtain the benefits of coordinated actions or infer otherwise inaccessible information. In such situations this social influence decreases the *ex ante* predictability of the ensuing social dynamics. We claim that, interestingly, these same social forces can *increase* the extent to which the outcome of a social process can be predicted very early in the process. This paper explores this claim through a theoretical and empirical analysis of the experimental music market described and analyzed in [1]. We propose a very simple model for this music market, assess the predictability of market outcomes through formal analysis of the model, and use insights derived through this analysis to develop algorithms for predicting market share winners, and their ultimate market shares, in the very early stages of the market. The utility of these predictive algorithms is illustrated through analysis of the experimental music market data sets [2].

Keywords: Social dynamics, prediction, theoretical analysis, empirical analysis.

1 Introduction

Enormous resources are devoted to the task of predicting the outcomes of social processes, in domains such as economics, public policy, popular culture, and national security, but the quality of such predictions is often quite poor. Consider, for instance, the case of cultural markets. Perhaps the two most striking characteristics of these markets are their simultaneous *inequality*, in that hit songs, books, and movies are many times more popular than average, and *unpredictability*, so that well-informed experts routinely fail to identify these hits beforehand. Examination of other domains in which the events of interest are outcomes of social processes reveals a similar pattern – market crashes, regime collapses, fads and fashions, and "emergent" social movements involve significant segments of society but are rarely anticipated.

It is tempting to conclude that the problem is one of insufficient information. Clearly winners are qualitatively different from losers or they wouldn't be so dominant, the conventional wisdom goes, so in order to make good predictions we should

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 79–86, 2010.

collect more data and identify these crucial differences. Research in the social and behavioral sciences calls into question this conventional wisdom and, indeed, indicates that there may be fundamental limits to what can be predicted about social systems. Consider social processes in which individuals pay attention to what others do. Recent empirical studies offer evidence that the *intrinsic* characteristics of such processes, such as the quality of the various options in a social choice situation, often do not possess much predictive power [3-10].

In order to understand this phenomenon more deeply, Salganik, Dodds, and Watts [1] conducted an elegant experiment in which over 14,000 participants were recruited to participate in an "artificial" music market and the impact of social influence on their choice of songs to download was examined. Briefly, the participants were presented with a web page displaying a selection of 48 songs by unknown bands and were asked to choose songs to listen to and download. As they arrived at the music market site they were randomly assigned to one of two experimental conditions: Independent, in which they saw only the names of bands and songs, and Social Influence, in which they were further divided into eight distinct "worlds" and could see (in addition to the bands and songs) the number of times each song had been downloaded by previous participants in their respective worlds. There were three main findings: 1.) song "quality" is only weakly related to market share success, 2.) the presence of social influence leads to "herding" behavior regarding song popularity, and 3.) increasing the strength of social influence increases both inequality and unpredictability of market outcomes.

The empirical analyses [3-10] and the experimental study [1] provide evidence that, for many social processes, it is not possible to obtain useful predictions using standard methods, which focus almost exclusively on the intrinsic characteristics of the process and its possible outcomes. We propose that useful prediction requires consideration of both intrinsics and the underlying *social dynamics*, and offer in [11] a new approach to predictive analysis which leverages this idea. The present paper applies this analytic framework to the experimental music market described in [1,2] and derives two main results. First, we develop a simple model for cultural markets (such as the music market [1]) which captures both process intrinsics and social influence dynamics. This model is employed to formally assess the predictability of market outcomes for various sets of candidate measurables and thereby identify measurables which possess predictive power. Second, using insights derived from this predictability analysis we formulate algorithms for predicting market share winners, and their ultimate market shares, very early in the market's evolution; the performance of these algorithms is illustrated through predictive analysis of experimental music market data sets obtained from [2].

2 Predictability Assessment

Basic idea. A defining characteristic of cultural markets is that participants are influenced by the behavior of others, for instance because they wish to obtain the benefits of coordinated action (e.g., enjoy the offering with friends) or infer otherwise inaccessible information (e.g., by observing people "in the know"). Processes in which observing a certain behavior increases an individual's probability of adopting that

behavior are often referred to as *positive externality processes* (PEP), and we use that term here. One hallmark of PEP is their apparent unpredictability: phenomena from hits in cultural markets to crashes in financial markets to political upheavals appear resistant to predictive analysis (although there is no shortage of *ex post* explanations for their occurrence!).

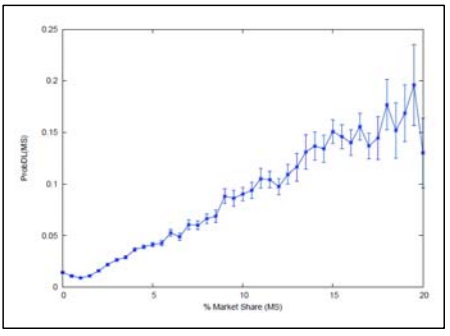
It is not difficult to gain an intuitive understanding of the basis for this unpredictability. Individual preferences and opinions are mapped to collective outcomes through an intricate, dynamical process in which people react individually to an environment consisting of others who are reacting likewise. Because of this feedback dynamics, collective outcomes can be quite different from those implied by simple aggregations of individual preferences; standard prediction methods, which typically are based (implicitly or explicitly) on and aggregation ideas, do not capture these dynamics and therefore are often unsuccessful. Interestingly, the feedback dynamics which reduces PEP predictability based on simple preference aggregation may *increase* the predictive power of very early measurements of these dynamics. Again the intuition is clear: early trends are reinforced through the positive feedbacks of PEP, suggesting the possibility that early rankings of alternatives may be informative concerning the ultimate outcomes. We now explore this intuition more formally.

Model. Consider an online market, such as the music market [1], in which individuals visit a web site, browse an assortment of available items, and choose one or more items to download. For simplicity, we focus on a market visited by a sequence of consumers, with each visitor choosing between two items {A, B}; generalizing this simple binary choice setting to any finite number of choices is straightforward [12,13]. We model this situation by supposing that agent i chooses item A with probability

Σ_{online} $\qquad P_i(A) = \beta\pi + (1-\beta)\, f$

where $f \in [0,1]$ is item A's current market share, $(1-\beta)$ quantifies the intensity of social influence (with $\beta \in [0,1]$), and π is the probability of an agent choosing A in the "no social influence" case (i.e., when $\beta=1$). Agent i selects item B with probability $1 - P_i(A)$. In this model, π can be interpreted as a measure of the "appeal" of item A (relative to B), f is the social signal, and β quantifies the relative importance of appeal and social influence in the decision-making process.

The model Σ_{online} is extremely simple, perhaps the simplest possible representation which captures the effects of both social influence and appeal in an online market. Nevertheless, this model reflects key behaviors observed in the music market [1] as well as in other cultural markets [e.g., 9]. For instance, Figure 1 provides a quantitative characterization of the roles of social influence and appeal for the music market [1]. The plot at the left of Figure 1 shows that in the Social Influence worlds song download probability (vertical axis) increasing with number of previous downloads (horizontal axis); the error bars represent two standard errors. The plot at the right of the figure shows ultimate market shares for all songs in the Independent worlds and indicates that, in the absence of social influence, some songs (e.g., song 25) are consistently appealing. Note that, in particular, the dependence of song download probability on previous downloads is approximately linear and capturing intrinsics in terms of download probability in the Independent condition appears reasonable.



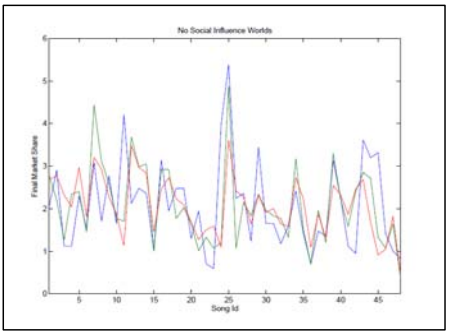


Fig. 1. Some characteristics of the music market dynamics. The plot at left shows that in the social influence worlds song download probability increasing with number of previous downloads. The plot at right depicts ultimate market shares for all songs in the no social influence worlds and indicates that, in the absence of social influence, a few songs are consistently appealing.

Moreover, simulations of Σ_{online} show that as social influence (SI) increases (β decreases) both inequality and unpredictability of market shares increase. Thus, despite its simplicity, Σ_{online} provides a useful starting point for studying predictability of online markets. Note that Σ_{online} can be written as a set of stochastic differential equations with state variables $x_1 = f$ and $x_2 = 1/(t+1)$ [12], so that the system's predictability properties can be evaluated using the methods presented in [11].

Predictability assessment. Consider the predictability of ultimate market share for the system Σ_{online}. We develop in [11] a mathematically rigorous approach to predictability assessment for a broad range of dynamical systems. Here we apply this assessment methodology in a somewhat informal, intuitive manner; the reader interested in a more formal analysis is referred to [11, 12]. Our main interest in this study is *eventual state (ES) predictability*. Roughly, a system Σ is ES predictable if qualitatively different outcomes (say hit or flop in a cultural market) have sufficiently different probabilities of occurring when Σ is initialized at similar starting configurations (see [11] for a more precise definition). A key aspect of this definition is that it depends upon the specification for "similar" initial configurations, which in turn depends on which system states and parameters are assumed measurable; this dependence permits ES predictability assessment to be used to *identify* measurables with predictive power.

The standard approach to market share prediction is to assume that item appeal is a relevant measurable, estimate appeal in some way, and use this estimate to predict market share. To examine the utility of this approach, we assess ES predictability of market share for items with identical appeal (π=1/2) and identical initial market shares (f(0)=1/2). If it is likely that the market will evolve so one or the other item dominates (f becomes large or small), then market dynamics is not very dependant on item appeal and therefore is unpredictable using the standard approach. In this case we should seek a different prediction method, perhaps based on other measurables. Alternatively, if market dominance by either item is unlikely then the market dynamics depends on item appeal in a more predictable way and the standard method may be useful.

We evaluate ES predictability using the assessment procedure proposed in [11]. Let X_{s1} and X_{s2} be two subsets of the state space of Σ_{online} corresponding to, respectively, $f \approx 1/2$ (approximately equal market share) and large/small f (market dominance by one or the other item). Define the set of similar initial configurations X_0 to be a small set surrounding $f(0) = 1/2$, the identical initial market share condition. Then, if both X_{s1} and X_{s2} are likely to be reached from X_0, the problem is ES unpredictable (and also unpredictable in a practical sense). See Figure 1 for a sketch depicting the basic setup.

As an illustration of the insights obtainable with such analysis, consider the high social influence (SI) case corresponding to small β in Σ_{online}. For a broad range of noise models, the analysis generates fairly high probabilities for reachability of both X_{s1} and X_{s2} from X_0. Thus two qualitatively different outcomes – market share equity (X_{s1}) and market shares dominance (X_{s2}) – are both likely, indicating that the system is ES unpredictable. This result is consistent with empirical findings for cultural markets [e.g., 1] and suggests that the standard approach to market share prediction is not likely to produce accurate forecasts.

Next consider the problem of searching for alternative measurables which provide better predictability properties in the high SI case. For example, it might be supposed that very early market share time series data would be useful for prediction when SI is high. The intuition behind this idea is that the "herding" behavior that can arise from SI, and which makes market prediction hard using standard methods, may lead to a lock-in effect, in which very early market share leaders become difficult to displace. To test this hypothesis, define X_0^* to be a small set surrounding $f(t^*) = 1/2$, where t^* is a small *but nonzero* time (see Figure 2). We compute, using the ES predictability assessment algorithm given in [11], the probability that Σ_{online} with π=1/2 will evolve from X_0^* to X_{s1} and X_{s2}. In this case, the analysis yields a high probability of reaching X_{s1} and low probability of reaching X_{s2} (typical probabilities are on the order ~0.9 and $\sim 10^{-3}$, respectively). Thus using very early time series data produces a more predictable situation, in which qualitatively distinct outcomes have qualitatively different probabilities of occurring.

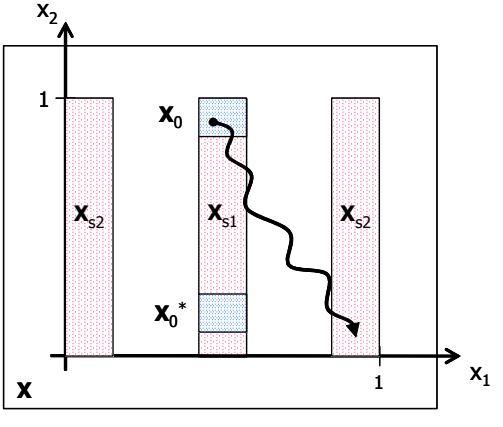


Fig. 2. Setup for online market predictability assessment

3 Market Prediction

This section leverages the insights obtained through predictability assessment to address two problems: 1.) identifying market share winners very early in the market's evolution and 2.) predicting ultimate market shares for these winning songs.

Identifying market share winners. The objective here is to identify a practically measurable "indicator" condition which enables successful songs to be recognized early in the music market's evolution. We focus on the ten high SI markets in the experiments [2], as these are the most unpredictable using standard methods [1]. Our method is simple and natural. The distribution of downloads in high SI markets is right-skewed, reflecting the PEP nature of these markets and in particular the tendency for market share "lock in" to occur early in the process. This observation suggests that when a market first exhibits signs of right-skew in market share distribution, a good prediction for the song that will ultimately win the largest market share is the one with leading market share at that point.

Consider the simple measure of right-skew $MM_i(t) = mean_i(t)/median_i(t)$, where $mean_i(t)$ is market share mean for the 48 songs in (high SI) world i at time t and $median_i(t)$ is the analogous median (time t is measured in market "ticks" [2]). The plot in Figure 3 shows that the dynamics of $MM_i(t)$ provides a reliable early means of distinguishing high SI markets from the low and no SI markets. Moreover, these data indicate that high SI markets reach $MM_i(t) \geq 1.1$ very early (i.e., at ~5% of the total market trajectory). Thus we propose the following method for identifying market share winners: for a given market i, predict as the ultimate market share winner the song with leading market share when $MM_i(t) \geq 1.1$ for the first time. Implementing this strategy with the music market data [2] yields the following results: 1.) the winning song is correctly identified in 90% of the high SI markets and 2.) this identification is made within the first ~5% of the market trajectory.

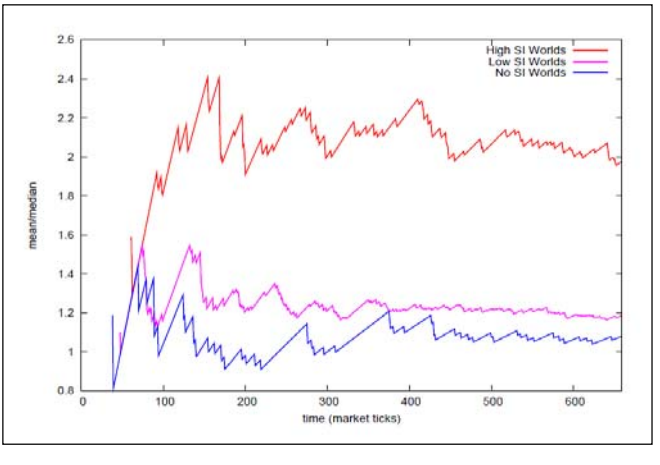


Fig. 3. Dynamics of $MM_i(t)$. The plot shows average $MM_i(t)$ for the high SI (red), low SI (magenta), and no SI (blue) music markets.

Predicting ultimate market shares. Consider next the problem of predicting the market shares for successful songs early in the market's evolution. More specifically, we wish predict the ultimate market share for the top five songs in a given market. Again the focus is on high SI markets, as these are unpredictable using standard methods [1]. We propose the following very simple prediction model:

$$\Sigma_{ms}: \quad ms_i(T) = \kappa + \alpha\, ms_{i,\,noSI} + \beta_1\, ms_i(\tau) + \beta_2\, ms_i(k\tau) + \beta_3\, ms_{i,\,mw}(a\tau, k\tau),$$

where $ms_i(t)$ is the market share of song i at time t, T is end time for the market under study, $ms_{i,\,noSI}$ is the mean ultimate market share for song i in the no SI markets, τ is the time at which this market first reaches $MM(t) \geq 1.1$, $ms_{i,\,mw}(t_1, t_2)$ is the mean market share for song i over the "moving window" $[t_1, t_2]$, $k \geq 1$ defines how much early market share time series is available to the prediction model, and $\{a, \kappa, \alpha, \beta_1, \beta_2, \beta_3\}$ are the model parameters. Note that obviously more sophisticated prediction models could be used; here the goal is to demonstrate useful performance with a simple linear regression predictor.

We now summarize some results of applying Σ_{ms} to the task of predicting the ultimate market share of 50 successful songs (the top five songs in each of the ten high SI markets). Note first that, for a broad range of early time series availability (i.e., k specified so that 5%–50% of market time series is used for prediction), all terms in Σ_{ms} (except κ) are statistically significant predictors of final market share ($p<0.05$). Next, consider the extent to which final market share can be predicted using only the intrinsic appeal of the songs, as measured by $\alpha\, ms_{i,\,noSI}$. As shown in Figure 4 (plot at left), this quantity has limited predictive power, explaining less than 50% of the variance of final market share over the 50 successful songs. In contrast, the most predictive dynamics term $\beta_2 ms_i(k\tau)$ can provide useful predictions even if only a small amount of early time series is available (e.g., this term explains almost 80% of final market share variance if 20% of early time series is available); see Figure 4 (plot at left).

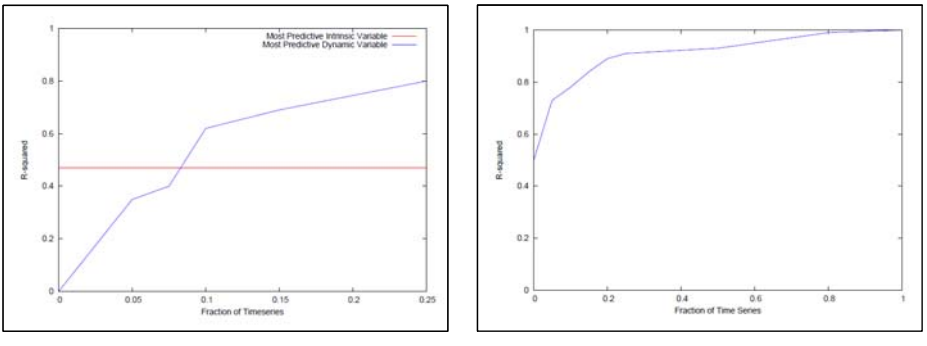


Fig. 4. Sample music market prediction results. The plot at left shows the fraction of final market share variance explained by the most predictive intrinsic alone ($\alpha\, ms_{i,\,noSI}$, red) and most predictive dynamics variable alone ($\beta_2\, ms_i(k\tau)$, blue), as a function of the fraction of early time series available. The plot at right depicts the fraction of final market share variance explained by Σ_{ms} as a function of the fraction of early time series available.

As expected from the predictability assessment summarized in Section 2, the predictive power of Σ_{ms} increases rapidly as a function of amount of early market share time series available to the model; this dependence is shown in Figure 4 (plot at right). Finally, a tenfold cross validation study shows that the model Σ_{ms}, although simple, provides good out-of-sample prediction performance. For example, with access to the first 15% of market share time series the model provides an out-of-sample prediction accuracy of ~82%.

Acknowledgements

This research was supported by the U.S. Department of Homeland Security, the U.S. Department of Defense, and the Laboratory Directed Research and Development program at Sandia National Laboratories.

References

1. Salganik, M., Dodds, P., Watts, D.: Experimental study of inequality and unpredictability in an artificial cultural market. Science 311, 854–856 (2006)
2. http://www.princeton.edu/~mjs3/data.shtml (accessed 2009)
3. Arthur, W.: Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116–131 (1989)
4. Bikhchandani, S., Hirshleifer, D., Welch, I.: Learning from the behavior of others. J. Economic Perspectives 12, 151–170 (1998)
5. Hedstrom, P., Sandell, R., Stern, C.: Mesolevel networks and the diffusion of social movements. American J. Sociology 106, 145–172 (2000)
6. Shiller, R.: Irrational Exuberance. Princeton University Press, Princeton (2000)
7. Rogers, E.: Diffusion of innovations, 5th edn. Free Press, NY (2003)
8. Walls, W.: Modeling movie success when 'nobody knows anything': Conditional stable-distribution analysis of film returns. J. Cultural Economics 29, 177–190 (2005)
9. Colbaugh, R., Glass, K.: Predictability and prediction of social processes. In: Proc. 4th Lake Arrowhead Conference Human Complex Systems, Lake Arrowhead, CA (April 2007)
10. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, Cambridge (to appear 2010)
11. Colbaugh, R., Glass, K.: Predictive analysis for social processes I: Multi-scale hybrid system modeling, and II: Predictability and warning analysis. In: Proc. 2009 IEEE Multi-Conference on Systems and Control, Saint Petersburg, Russia (July 2009)
12. Colbaugh, R., Glass, K.: Predictive analysis for social processes. Sandia National Laboratories SAND Report 2009-0584 (January 2009)
13. Ormerod, P., Colbaugh, R.: Cascades of failure and extinction in evolving complex systems. J. Artificial Societies Social Simulation 9(4) (2006)

Macroeconomic Analysis of Universal Coverage in the U.S.

Zhigang Feng*

ISB, University of Zurich,
Plattenstrasse 32, Zurich 8032, Switzerland
feng@isb.uzh.ch
http://www.isb.uzh.ch

Abstract. In this paper I employ a dynamic general equilibrium model to study macroeconomic effects and welfare implications of health policies for universal coverage in the U.S. The model is calibrated to the U.S. data. Numerical simulations indicate that adopting universal coverage has several important macroeconomic effects on health expenditures, hours worked, and increases welfare by improving aggregate health status, and removing adverse selection.

Keywords: Health care reform, Heterogeneous agents model, Welfare analysis.

1 Introduction

National health expenditures accounted for 16.3% of the U.S. GDP in 2007, compared to 5.2% in 1960 (Department of Health & Human Services, 2006). The rapid growth of medical costs leaves a large fraction of the population without health insurance.[1] The lack of insurance has serious negative consequences that include lack of access to needed care, declining health, and the possibility of crushing financial burdens. Uninsured adults are far more likely to postpone accessing health care or to forgo it altogether and are less able to afford prescription drugs or follow through with recommended treatments. A report by the Institute of Medicine (2003) states that the uninsured have a more rapid decrease in general health and a higher risk of dying prematurely than the insured. According to their estimation the cost for diminished health and shorter life span due to lack of insurance was between $65 and $130 billion in 2003. There are also financial externalities imposed by the uninsured on the third party through uncompensated care, whose costs were estimated to be $57.4 billion in 2008 [Hadley et al. (2008)].

These facts have stirred up various proposals for changing the U.S. health care system and to cover the uninsured. There are many empirical studies that explore

* I am very grateful to Adrian Peralta-Alva and Manuel Santos for their advice and encouragement. I also thank two referees for helpful comments. All remaining errors are mine.

[1] 17% of the nonelderly in the US was uninsured in 2007 according to Kaiser (2008).

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 87–96, 2010.

the impacts of health care reforms on individual's behavior such as crowding-out by public insurance [e.g. see Culter and Gruber (1996), Lo Sasso and Buchmueller (2004), Gruber and Simon(2008)], medical usage [Cheng and Chiang (1997)], and health status [Lurie et al. (1984), Currie and Gruber (1996), Hanratty (1996), Decker and Remler (2005)]. However, there is a paucity of economic models that address the macroeconomic and welfare implications of reforming the U.S. health care system.

A reform of the health insurance system could potentially affect macroeconomic variables by distorting the labor market through changes in tax rates, reducing the number of uninsured, and raising the aggregate health expenditure. Reforming the health insurance system will affect the household's demand for health insurance. Some individuals may shift from existing private insurance coverage to either the newly subsidized form of private coverage or to public coverage. This in turn alters the pool of agents insured, which affects insurance premiums. Similarly, different insurance decisions result in changing health status and labor productivity, which then will affect wages and hours worked. A change in the labor income tax may be required to fund the reform, which consequently will influence individual's labor supply decisions. A reform will also change agents' saving behavior (and thus the aggregate capital stock and factor prices) because health insurance may reduce precautionary saving motives. At the same time, better health implies longer life expectancy and thus a higher saving incentive. These complicated tradeoffs can only be fully captured in a general equilibrium framework.

The aim of this study is to analyze the macroeconomic impacts and welfare implications of alternative reforms to the health insurance system in the U.S. I employ a stochastic OLG framework developed by Feng (2009), which helps to capture general equilibrium effects of health care reforms on some important aggregate variables. I consider the expansion of Medicare to the entire population, which is one of the major reform proposals discussed in the U.S. I calibrate my model to the U.S. data. Then, I conduct several policy experiments to shed light on the costs and benefits of changing the health insurance system. My numerical experiments suggest that general equilibrium effects are substantial, and the impact of various reforms on the social welfare can be quite sizable.

The paper is organized as follows. Section 2 introduces the OLG model, while section 3 details reform proposal and presents all numerical results. The last section concludes.

2 The Dynamic Model

2.1 Demographics

This economy has overlapping generations of agents who live a maximum of three periods as *young*, *middle-aged*, and *old*. Let $g \in \{1, 2, 3\}$ denote the age. In the first period, the measure of newly born agents is normalized to 1. Individuals alive in period t survive to the next period with a certain probability. For old people this probability is always 0. For young and middle-aged people, the survival

probability is given by $\rho(h_g)$, which depends on the health status h_g at the end of age g as described below. The population of young individuals grows at a constant rate n, implying that the population of young in period t is $(1+n)^t$. I denote the relative size of age g to the population as μ_g, which is determined in the equilibrium.

2.2 Agent Types

All individuals enter the economy with the same level of health $\bar{h}_0$, an idiosyncratic endowment e_0, and an idiosyncratic health risk types i_h. Health risk type determines the probability of drawing a certain health shock $\varepsilon_t \in \Omega_\varepsilon = \{\varepsilon^1, ..., \varepsilon^{N_\varepsilon}\}$. The probability distribution of the shock is assumed to be age-type-dependent. Specifically, the probability of drawing $\varepsilon \in \Omega_\varepsilon$ by type i_h agent at age g is denoted by $p_{g,i_h}(\varepsilon)$, with $\Sigma_{\varepsilon\in\Omega_\varepsilon} p_{g,i_h}(\varepsilon) = 1$ for all (g, i_h). A typical history of shocks up to time t is denoted by $\sigma_t \equiv \{\varepsilon_0, ..., \varepsilon_t\}$, with $\sigma_{t+1} = \{\sigma_t, \varepsilon_{t+1}\}$. Agents are endowed with a fixed amount of time per period that can be allocated to leisure or labor. Agents participate in the labor market during the first two periods and receive a wage income $\tilde{w}e^{\zeta h}l$. Here ζ measures the effect of health on labor productivity.

During their work stage agents receive income in the form of wages and profit Π_t from the firm. They can also save a_g units of the consumption good using a storage technology with gross rate of return $R_{t+1} = 1 + r$. Retired agents have income through previous saving and profit, and consume all of their income at their last period of life.

2.3 Preferences

Preferences over stochastic sequences of consumption, leisure and health are given by

$$U = \mathbf{E}_t \left\{ \sum_{g=1}^{3} \beta^{g-1} \Pi \rho(h_{g-1}) \cdot u(c_g, L_g, h_g) \right\} \tag{1}$$

where β denotes the discount factor, ρ survival probability, c consumption, L leisure and h health status. $\mathbf{E}_t$ denotes the conditional expectation with the information available when the agent is born.

2.4 The Evolution of Health

I use the idea of health capital introduced by Grossman (1972). In the model, each agent chooses an optimal amount of medical consumption m to offset the negative effect of health shock ε on health and builds up health capital h. The accumulation process of health is given by:

$$h' = (1 - \delta_h)h + \frac{\varepsilon}{\exp\left[A_m m^{\zeta}\right]}. \tag{2}$$

where δ_h represents the natural deprecation rate of health and A_m measures the medical technology. I assume that technological progress in the production of medical service A_m is exogenously given. The price of medical care p_m is exogenously given so that each unit of consumption good can be transformed into $\frac{1}{p_m}$ units of medical care.

Conditional on being alive at the current age with end of period health stock h, agent will survive to the next period with probability $\rho(h)$. Death is certain when health falls below zero ($\rho(h) = 0$ if $h \leq 0$). I assume that $\rho'(h) > 0$. Deceased agents leave their savings a as an accidental bequest that is collected by the government as revenues.

2.5 Medical Expenses and Health Insurance

Non-elderly can choose one out of three possible insurance states labeled as $in = \{1, 2, 3\}$. To purchase private health insurance is $in = 1$, $in = 2$ denotes that the agent has Medicaid, and $in = 3$ indicates that the agent is uninsured. The out of pocket health expenditure will be $(1 - \tilde{q}(p_m m, 1))p_m m$ if the agent chooses to buy insurance and $(1-\tilde{q}(p_m m, 2))p_m m$ when he/she is covered by the government program Medicaid. [2] It will cost the entire expenditure $p_m m$ if the agent does not have insurance($\tilde{q}(p_m m, 3) = 0$). Here $\tilde{q}(p_m m, in)$ is function that represents the coinsurance rate and varies with the health insurance state in as we discuss in the following subsection. Agents take coinsurance rate as given and it is calibrated from the data. Retired agents are insured under Medicare.

2.6 The Representative Agent's Problem

A representative agent of generation $g = \{1, 2\}$ enters each period with characteristics $s_g = (i_h, x, h_{g-1}, i_{ma})$, where i_h is the risk type of the agent, x is the net wealth, h_{g-1} is the health status at the beginning of the period, and i_{ma} is the indicator function that signals the availability of the Medicaid benefit in the current period. Since all old agents are automatically enrolled in the Medicare program and leave the labor market, their characteristics simply are $s_3 = (i_h, x, h_2)$. The distribution of households over their state space is given by $f_g(s_g, \sigma_t)$, which is endogenously determined in the equilibrium and evolves over time.

Agents observe s_g at the beginning of the period. They take prices and taxes as given and make the insurance decision $in_g(s_g)$ and choose a set of state-contingent decision rules, $\{c_g(s_g, \varepsilon_g), a_g(s_g, \varepsilon_g), m_g(s_g, \varepsilon_g), L_g(s_g, \varepsilon_g)\}$, to solve the following problem.

$$\max_t \mathbf{E}_t \left\{ \sum_{g=1}^{3} \beta^{g-1} \Pi \rho(h_{g-1}) \cdot u\left[c_g(s_g, \varepsilon_g), L_g(s_g, \varepsilon_g), h_g(s_g, \varepsilon_g)\right] \mid \sigma_t \right\} \tag{3}$$

[2] To simplify the analysis, I only consider the Employer-Sponsored Health Insurance *(EHI)*. As a matter of fact, more than 90% non-elderly who have private insurance purchase through their employers.

subject to the budget constraint and a no-borrowing constraint

$$(1+\tau_c)c_1(s_1,\varepsilon_1)+[1-\tilde{q}(p_m m_1,in_1)]\cdot p_m m_1(s_1,\varepsilon_1)+\tilde{\pi}(in_1)+a_1(s_1,\varepsilon_1)$$
$$\leq e_0+\Pi_t+(1-0.5\tau_{mr})\left[\tilde{w}_t e^{\zeta h_1}l_1(s_1,\varepsilon_1)-1_{\{in_1=1\}}\tilde{\pi}(in_1)\right]-T(y_1) \quad (4)$$
$$a_1(s_1,\varepsilon_1)\geq 0 \quad (5)$$

when young;

$$(1+\tau_c)c_2(s_2,\varepsilon_2)+[1-\tilde{q}(p_m m_2,in_2)]\cdot p_m m_2(s_2,\varepsilon_2)+\tilde{\pi}(in_2)+a_2(s_2,\varepsilon_2)$$
$$\leq R_{t+1}a_1+\Pi_{t+1}+(1-0.5\tau_{mr})\left[\tilde{w}_{t+1}e^{\zeta h_2}l_2(s_2,\varepsilon_2)-1_{\{in_2=1\}}\tilde{\pi}(in_2)\right]-T(y_2) \quad (6)$$
$$a_2(s_2,\varepsilon_2)\geq 0 \quad (7)$$

when middle-aged; and

$$(1+\tau_c)c_3(s_3,\varepsilon_3)+[1-q_{mr}(p_m m_3)]\cdot p_m m_3(s_3,\varepsilon_3)+\pi_{mr}$$
$$\leq R_{t+2}a_2+\Pi_{t+2}-T(y_3) \quad (8)$$

when old.

2.7 Aggregate Production Function

The consumption goods are produced by a neoclassical production function. The aggregate production function takes a nested Cobb-Douglas specification in the following form.

$$Y_t=A_t E_t^{\alpha} \quad (9)$$
$$E_t=\sum_{g=\{1,2\}}\mu_g(t)\int\left[e^{\xi h_g}l_g(s_g,\varepsilon_g)\right]f_g ds_g \quad (10)$$

where A_t is a total factor productivity, and E_t is an aggregate efficiency labor input, which depends on individual worker's health status. The firm's profit maximization problem is

$$\max_{\{E_t\}} A_t E_t^{\alpha}-w_t E_t. \quad (11)$$

Profits Π_t are distributed back to households in a lump-sum payment.

2.8 The Government

I impose a government balanced budget constraint period by period. The government has three different types of outlays: general public consumption, Medicaid and Medicare expenses. The government collects revenues from various sources: income taxation according to a progressive tax function $T(\cdot)$, consumption taxation at rate τ_c, Medicare taxation at rate τ_{mr}, Medicare premium π_{mr}, Medicaid premium π_{ma}, and accidental bequests B collected from deceased agents.

2.9 Health Insurance Company

The health insurance company is competitive. Hence, in equilibrium the premium π_E is charged such that expected expenditures on the insured are precisely covered.

$$\pi_E = \frac{\sum_{g=\{1,2\}} \mu_g(t) \int \left[q_E(p_m m_g) p_m m_g \cdot 1_{\{in=1\}} \right] f_g ds_g}{\sum_{g=\{1,2\}} \mu_g(t) \int 1_{\{in=1\}} f_g ds_g} \tag{12}$$

Notice the coverage ratio function $q_E(\cdot)$ is taken as exogenously given.

3 Numerical Results

In this section, I conduct counterfactual experiments as in Feng (2009) to determine the effect of reforming the health insurance system. In line with Conesa and Krueger (1999), I measure the welfare effect of a reform by computing the consumption equivalent variation (CEV).[3]

Alternative sources of revenue to fund these reforms are also considered. I first consider supporting the reform by adjusting the income tax. I also conduct companion experiments where the government funds the reform through a labor income tax and through a lump-sum transfer separately.

3.1 To Fund Reforms by Income Tax

In this experiment the private health insurance and the Medicaid program are abolished. Non-elderly will be covered by a uniform health insurance program, which is sponsored by the government, with premium π_{mr} and coverage rate $q_E(\cdot)$. Specifically, non-elderly pay for a premium that equals 2.1% of the per capita GDP, which is cheaper than the counterpart in the benchmark.[4] A fraction $q_E(p_m m)$ of their health expenditure will be paid by the government.

I assume that the average price level for medical service p_m and medical technology A_m are constant and exogenously given. I can also consider a case in which the technology slows down (or speeds up) as a result of the reform.

Experiment results are summarized in Table 11. The top section displays some statistics of aggregate variables: the fraction of insured non-elderly, the Medicare tax rate, the average effective income tax rate, average hours worked, average effective working hours, and the health expenditure as a ratio of GDP. The lower section displays the welfare effects of each reform. *% w/ CEV > 0* indicates the fraction of agents in the benchmark that would experience a welfare gain (positive CEV) if the alternative reform is taken place.

[3] A $CEV(i_h, x, i_{ma})$ of -10% implies that if the given policy reform is put into place, then an individual of type (i_h, x, i_{ma}) will experience a welfare loss due to the reform equivalent to sacrifice 10% of his consumption in the initial steady state with leisure, health insurance and health expenditure constant at the initial choices.

[4] In the benchmark, 72.5% of non-elderly who purchase private insurance pay an actuarially fair premium π_E, which is about 10.9% of the per capita GDP.

Expansion of Medicare to the entire population achieves a universal coverage as shown in the fraction of insured non-elderly. The aggregate health expenditure as a ratio of GDP increases by 0.3%. This is attributed to the fact that those newly insured non-elderly will utilize more medical service and incur higher amount of health expenditure as the reform provides them with cheaper health insurance. The current reform needs to raise tax revenue to cover 15.2% of the non-elderly who would be uninsured in the benchmark and to pay for part of the expenditure of the previously insured, who pay a premium of π_{mr} after the reform, which is about 20.0% of the premium they paid before the reform. The reform also saves some tax revenues through changes in the arrangement in the health care sector. In the benchmark, the government provides Medicaid to the low incomes, which costs 2.2% of total GDP. It also subsidizes the purchase of group insurance and the total subsidy amounts to 0.8% of total GDP. Once the reform is implemented, the government can save these spending, since both Medicaid and private insurance are abolished. Put them together, the government raises the proportional income tax rate by 4.5%, which will discourage labor supply. At the meantime, the individuals have access to better health insurance. Average health has been improved, which brings workers higher productivity and incentive to work longer. Consequently, average hours worked decreases by 4.8% to 28.7 hours per week. Total output decreases by 2.0% as labor supply shrinks and average consumption decreases by 3.0%.

Now let's look at the saving behavior. The average health stock of the non-elderly increases, which implies a longer life expectancy and a stronger saving incentive. A decreased exposure to the health shocks lowers the demand for precautionary saving, but this effect is dominated by the previous one and the aggregate saving rate slightly increases by 0.8%.

Although agents are subject to a higher income tax after the reform is implemented, the cheaper health insurance program from the government is enough to compensate this cost for most agents. As shown in *% w/ CEV > 0,* 72.6% of agents would experience a welfare gain from this reform, and the average welfare effect is in the order of 2.6% in terms of consumption in all states. However, low income agents, especially those covered by Medicaid before the reform, will suffer from this policy because the new insurance program from such a reform is less generous than Medicaid. On average, low income individuals would experience a welfare loss equivalent to 4.3% of consumption. Compared to agents who have income above the poverty line have a welfare gain equivalent to 6.0% of consumption.

In order to understand whether there exists a Pareto-improving variation of the above reform, I also consider experiment A-2. This experiment is similar to A-1 except that all low income agents are covered by Medicaid program. Under such reform, the tax rate needs a bigger increase since the health insurance provided to the low income is more generous than the one in experiment A-1. Consequently, they will consume higher amount of medical services which drives aggregate health expenditure to rise. Nevertheless, the benefit from such a guaranteed Medicaid coverage cannot offset the loss due to a higher tax rate, which is

Table 1. Policy Experiment A

		income tax		labor tax		lump-sum tr.	
	Bench.	A-1	A-2	A-1	A-2	A-1	A-2
Insured non-elderly (in %)	84.8	100.0	100.0	100.0	100.0	100.0	100.0
Medicare tax (in %)	2.5	2.5	2.5	7.9	11.2	2.5	2.5
Ave. income tax (in %)	24.6	29.4	30.4	24.7	24.5	25.4	25.4
Ave. Working hrs.	30.6	28.7	28.5	28.9	27.8	30.5	30.6
Ave. Effective Working hrs.	61.1	57.3	57.0	57.7	55.8	60.9	61.2
Health exp. (in % of GDP)	16.6	16.9	17.7	16.8	17.7	16.7	17.5
π_E (in % of per capita GDP)	10.1	2.1	2.1	4.2	4.2	2.1	2.1
Output	100.0	98.0	98.1	97.2	95.1	99.3	98.6
Average health stock	46.6	46.9	46.8	46.9	46.8	46.8	46.9
Lifetime CEV after transition							
all (in %)	–	2.6	2.8	1.8	2.5	2.7	3.0
income $> Y_{ma}$ (in %)	–	5.9	4.9	5.5	4.4	6.0	4.9
income $\leq Y_{ma}$ (in %)	–	−4.3	−1.4	−5.6	1.3	−3.9	−0.9
% w/ CEV > 0	–	72.6	76.7	72.6	76.7	72.6	76.7

required to provide generous Medicaid program to low income agents. As shown in CEV from transition, agents with income lower than the poverty line still experience a welfare loss, but at a much smaller magnitude of 1.4%. The welfare gain of higher income agents decreases to 4.9% from 5.9% in experiment A-1. On average, agents have a welfare gain in the order of 2.8% in terms of consumption in all states. From this experiment, it seems possible to make expansion of Medicare a Pareto-improving program by appropriately funding the reform.

3.2 To Fund Reforms by a Labor Income Tax

In order to understand how the macroeconomic effects of these proposals change in response to how the government finances the reform, I also consider funding the reform by changing the Medicare tax τ_{mr}. Now, government expenditure G, consumption tax rate τ_c and the progressive part of income tax function $T(\cdot)$, as well as the proportional tax rate τ_y remain unchanged from the benchmark. The government adjusts the Medicare tax rate τ_{mr} to balance the budget.

As shown in average working hours in table 1, to fund the reform through labor income tax creates stronger distortions compared with income taxes.[5] Notice I change some policy targets in order to make the experiment meaningful. The Medicare premium doubles from 2.1% of GDP to 4.2%. Otherwise the labor income tax rate will skyrocket and partially crash the labor market as some agents will leave the market. To finance the reform with labor income tax requires

[5] There is no capital in my model. The profit Π is distributed back to the agent as a payment, which is inelastic supply to the individual. The interest rate is exogenous and the demand for saving is inelastic as well. Furthermore, the tax base of income tax is broader than labor income tax. These facts explain why taxing labor income creates more distortion than taxing gross income.

τ_{mr} to increases from 2.5% to 7.9%. As a consequence, average hours worked decreases by 5.6%. The welfare of a typical agent decreases compared to the case when the government finances the reform through the gross income tax.

3.3 To Fund Reforms by a Lump-Sum Transfer

The analysis so far indicates that the change in taxes may play a dominant role in how health care reforms affect the macroeconomy. In order to isolate the effect of tax distortion, I also conducted companion exercises in which the government funds the reform through a lump sum transfer. In the companion experiments, the tax rates are kept intact as in the benchmark. The government returns a lump sum transfer to each individual. The transfer is determined so that the government's budget is balanced.

Numerical results in Table 1 indicate that the labor supply effect of health care reforms is rather small. Medicare expansion increases welfare by improving health status and reducing adverse selection in the health insurance market.

4 Concluding Remarks

In this paper, I build up a micro-founded dynamic general equilibrium model to study the impact of alternative health care reforms on the aggregate labor supply, health expenditures, savings, welfare, and the fraction of uninsured population. In contrast to some papers in the literature, I consider a model with a labor-leisure choice as well as a health expenditure decision. These latter choices may change the demand for medical services, which in turn affects the individual's health status and labor productivity. Moreover, financing reform may create distortions on the labor supply by requiring additional tax revenues. The magnitude of the distortion depends on the details of the reform as well as the funding method. My results suggest that the aggregate health expenditure rises as the insured population increases. Funding the reform through payroll taxes does not seem promising because such a policy can heavily distort the labor market.

Welfare analysis from numerical simulation suggests that universal health coverage increases lifetime CEV. The primary sources of welfare gains include improved aggregate health status through more inclusive health insurance coverage, decreased adverse selection, and higher labor productivity associated with individual health. These results raise a question why is universal health coverage highly unpopular in the U.S.? The gap between this theoretical prediction and political antipathy to universal healthcare in reality is an interesting research topic. However it is beyond the scope of the current study. Another interesting extension is to look at the macroeconomic effects of health care reform in other countries or to conduct an international comparison. I leave these subjects for future research.

References

Anderson, G.F.: From 'Soak the Rich' to 'Soak the Poor': Recent Trends in Hospital Pricing. Health Affairs 26(3), 780–789 (2007)

Cheng, S.H., Chiang, T.L.: The Effect of Universal Health Insurance on Health Care Utilization in Taiwan. Results from a Natural Experiment. The Journal of the American Medical Association 278, 89–93 (1997)

Conesa, J.C., Krueger, D.: Social Security Reform with Heterogenous Agents. Review of Economic Dynamics 2, 757–795 (1999)

Currie, J., Gruber, J.: Health Insurance Eligibility, Utilization of Medical Care and Child Health. Quarterly Journal of Economics 111, 431–466 (1996)

Cutler, D.M., Gruber, J.: Does Public Insurance Crowd out Private Insurance. Quarterly Journal of Economics 111, 391–430 (1996)

Decker, S.L., Dahlia, K.R.: How Much Does Universal Health Insurance Reduce Socioeconomic Disparities in Health? A Comparison of the US and Canada. Applied Health Economics and Health Policy 3(4), 205–216 (2005)

Feng, Z.: Macroeconomic Consequence of Alternative Reforms to the Health Insurance System in the U.S. University of Miami Working Paper No. 0908 (2009)

Gruber, J., Kosali, S.: Crowd-Out Ten Years Later: Have Recent Expansions of Public Insurance Crowded Out Private Health Insurance? Journal of Health Economics 27, 201–217 (2008)

Hadley, J., Holahan, J., Coughlin, T., Miller, D.: Covering The Uninsured In 2008: Current Costs, Sources Of Payment, And Incremental Costs. Health Affairs 27(5), 399–415 (2008)

Hanratty, M.J.: Canadian National Health Insurance and Infant Health. American Economic Review 86, 276–284 (1996)

Institute of Medicine, A Shared Destiny: Community Effects of Uninsurance. National Academy of Sciences (2003)

Kaiser: The Uninsured: A Primer. Kaiser Commission on Medicaid and the Uninsured (2008)

Sasso, L., Anthony, T., Buchmueller, T.C.: The Effect of the State Children's Health Insurance Program on Health Insurance Coverage. Journal of Health Economics 23(5), 1059–1082 (2004)

Lurie, N., Ward, N.B., Shapiro, M.F., Brook, R.H.: Termination from Medi-Cal - does it Affect Health. New England Journal of Medicine 311, 480–484 (1984)

Wobus, D.Z., Olin, G.: Health Care Expenses: Poor, Near Poor, and Low income People in the United States Civilian Noninstitutionalized Population. Agency for Healthcare Research and Quality Working Paper No. 05016 (2005)

Projecting Sexual and Injecting HIV Risks into Future Outcomes with Agent-Based Modeling

Georgiy V. Bobashev, Robert J. Morris, and William A. Zule

RTI International,
3040 Cornwallis Rd. P.O. Box 12194,
Research Triangle Park, NC 27709

Abstract. Longitudinal studies of health outcomes for HIV could be very costly cumbersome and not representative of the risk population. Conversely, cross-sectional approaches could be representative but rely on the retrospective information to estimate prevalence and incidence. We present an Agent-based Modeling (ABM) approach where we use behavioral data from a cross-sectional representative study and project the behavior into the future so that the risks of acquiring HIV could be studied in a dynamical/temporal sense. We show how the blend of behavior and contact network factors (sexual, injecting) play the role in the risk of future HIV acquisition and time till obtaining HIV. We show which subjects are the most likely persons to get HIV in the next year, and whom they are likely to infect. We examine how different behaviors are related to the increase or decrease of HIV risks and how to estimate the quantifiable risk measures such as survival HIV free.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, p. 97, 2010.

Cultural Consensus Theory: Aggregating Continuous Responses in a Finite Interval

William H. Batchelder, Alex Strashny, and A. Kimball Romney

Institute for Mathematical Behavioral Sciences, School of Social Sciences,
University of California, Irvine, CA 92697
{whbatche,astrashn,akromney}@uci.edu

Abstract. Cultural consensus theory (CCT) consists of cognitive models for aggregating responses of "informants" to test items about some domain of their shared cultural knowledge. This paper develops a CCT model for items requiring bounded numerical responses, e.g. probability estimates, confidence judgments, or similarity judgments. The model assumes that each item generates a latent random representation in each informant, with mean equal to the consensus answer and variance depending jointly on the informant and the location of the consensus answer. The manifest responses may reflect biases of the informants. Markov Chain Monte Carlo (MCMC) methods were used to estimate the model, and simulation studies validated the approach. The model was applied to an existing cross-cultural dataset involving native Japanese and English speakers judging the similarity of emotion terms. The results sharpened earlier studies that showed that both cultures appear to have very similar cognitive representations of emotion terms.

Keywords: Cultural Consensus Theory, cognitive models, cross-cultural study.

1 Introduction

This paper proposes a new cognitive model for Cultural Consensus Theory (CCT) involving numerical responses in a finite interval. This case might involve an informant responding with a mark in a finite interval with poles at "unfavorable" and "favorable," expressing a degree of confidence in the truth of a proposition about some cultural belief, or probability estimates of the likelihood of various events [1], [2]. Further the model is intended to serve as an approximation to cases where only a finite number of ordered responses are available such as the assignment of course grades to an exam paper. There have been models by others that can be applied to ordinal or continuous response data without prior knowledge of the answers, e.g. [3], [4]. Our approach is different in that it attempts to provide a cognitively plausible model of the informant.

CCT is an approach to information aggregation developed since its inception in the mid 1980s (see [5], [6]). CCT models are parametric statistical models employed in situations where each of a set of informants who share a 'common culture' responds to each of a set of questions about some domain of their shared knowledge. The models assume that there are consensus (culturally correct) answers to the questions

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 98–107, 2010.

that characterize the nature of the shared knowledge; however, these answers are not known to the researcher a priori. The primary goal of applying a CCT model is to estimate consensus answers to the questions, assess the degree of confidence in these estimates, and evaluate the adequacy of the assumptions behind the model.

The paper is divided into four sections. After this introduction, the model will be described. Then in Section 3 Bayesian inference for the model will be developed using MCMC methods. The final section will apply the inferential methods to both simulated data and previously published cross-cultural data.

2 A CCT Model for Responses in a Finite Interval

We develop the model for responses that fall into the unit interval I=[0, 1]; however for other finite intervals a simple linear transformation can be use to code responses in I. We follow an approach often seen in models for psychophysical judgments [7] by postulating that there are latent (unobservable) random variables, $Y_{ik} \in I$, that capture the internal representation produced in informant i by item k, for i=1,2,…, N and k=1,2,…, M. These internal representations are postulated to correspond to the consensus answer z_k distorted by a zero mean additive error random variable. The error distribution is allowed to vary from informant to informant as a function of a parameter D_i which stands for informant i's cultural competency, where the larger the competency the smaller the error variance. Unlike the treatment of error in most classical test theory models [8], the error distribution must also depend on the value of z_k since the internal representation is required to be in the interval [0,1]. The manifest responses $\mathbf{X} = (X_{ik})_{NxM}$ are a function of the corresponding latent random variables Y_{ik}; however, the model includes a bias function that has the potential to distort the latent representation due to each respondent's response bias in using the scale.

2.1 The Core Axioms for the Model

There can be different parametric versions of the model depending on the distribution one assumes for the error as well as how bias is represented; however, all of the versions of the model share the following four axioms about the latent random variables, $\mathbf{Y}= (Y_{ik})_{NxM}$.

Axiom 1. (**Common Truth**)**.** There is a fixed answer key, $\mathbf{Z}= \langle z_k \rangle_{1xM}$, where $z_k \in I$ for k= 1, 2, …, M.

Axiom 2. (**Latent Representations**)**.** The latent representational random variables are given by $Y_{ik}= z_k + e_{ik}$, where the e_{ik} are continuous type random variables representing measurement error, with $-z_k < e_{ik} < 1 - z_k$, $E(e_{ik}) = 0$, and variances $\sigma_{ik}^2 > 0$, i = 1, 2, ..., N and k = 1, 2, ..., M.

Axiom 3. (**Inhomogeneous Competence**)**.** There is a function ϕ from I into the positive reals given by $\phi(z) = z(1-z)$ and competencies $\mathbf{D}= \langle D_i \rangle_{1xN}$, with $D_i > 0$, such that

$$\sigma_{ik}^2 = \phi(z_k)/(1 + D_i). \tag{1}$$

Axiom 4. (**Conditionally Independent Errors**). The e_{ik} are conditionally independent given the parameters $\mathbf{Z}= <z_k >_{1xM}$ and $\mathbf{D}= <D_i >_{1xN}$.

Axiom 3 expresses σ_{ik}^2 as a ratio of a term involving the consensus answer parameter z_k and a term depending on the cultural competence parameter D_i of the informant. The form of $\phi(z)$ is motivated by calculating the least upper bound for the variance of a random variable in the interval [0,1)] with mean z. This bound would be achieved if all the probability were located at the end points. The D_i in Eq. (1) acts like a precision term in the sense that larger values lead to smaller error variance. Axiom 4 treats the parameters as random variables consistent with the Bayesian inferential analysis of the model in Section 3 using MCMC methods.

2.2 Introducing an Error Distribution

It is desirable to introduce a parametric model for the e_{ik} in Axioms 2 and 3 in order to facilitate estimation of the z_k and D_i. A flexible family of models on the unit interval is the beta distribution family, with parameters $\alpha, \beta > 0$ and density given by

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\cdot\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1}, \tag{2}$$

for $0 \le x \le 1$, and $\Gamma(\cdot)$ the gamma function. The beta distribution has mean $\mu = \alpha/(\alpha+\beta)$ and variance $\sigma^2 = \mu(1-\mu)/(\alpha+\beta+1)$. In order to adapt the beta distribution to our situation, we assume that Y_{ik} has a beta distribution with mean $\mu_{ik} = z_k$, and the variance is $\sigma_{ik}^2 = z_k(1-z_k)/(1+D_i)$. These restrictions lead to a reparameterization of the beta for each combination of informant i and item k given by $\alpha_{ik} = z_k \cdot D_i$, $\beta_{ik} = (1-z_k)\cdot D_i$.

Axiom 5. (Beta Distributed Errors). The Y_{ik} have a beta distribution given by Eq. (2) with parameters $\alpha_{ik} = z_k \cdot D_i$ and $\beta_{ik} = (1-z_k)\cdot D_i$.

2.3 Introducing Response Bias

The model assumes that the manifest response may reflect a bias function applied to the latent representation, $X_{ik}=h(Y_{ik})$, depending on how an informant uses the response scale. In this paper we model the biases of expansion and contraction of the scale. Reasonable restrictions on such a bias function can be represented by a one-parameter family of functions H satisfying the conditions that each function $h \in H$ is continuous and increasing from [0,1] onto [0,1] and subject to four functional constraints: $h(0)= 0$, $h(1/2)= 0.5$, $h(1)= 1$, and for $0 < y < 1$, $h(y)= 1-h(1-y)$. There are several ways to implement these constraints, and we chose one for the applications in this paper. The approach we used is based on designing a quadratic "Bézier curve" on [0, 1] that obeys the restrictions on h(y). Bézier curves are used extensively in computer graphics. For example many computer type fonts are defined by Bézier curves. The curves were developed by Paul de Casteljau [9] and popularized by Pierre

Bézier [10]. We use a quadratic Bézier curve to define the bias function. We omit the details, but the result is the one parameter family of bias functions given by

$$h(y,b)=\begin{cases}\left(b-2y\cdot b-1+y+\sqrt{1-2b+b^2-2y+4y\cdot b}\right)\cdot(2b-1)^{-1} & \text{if} \quad y\le .5, b\ne .5\\ \left(3b-1+y-2y\cdot b-\sqrt{-1+2b+b^2+2y-4y\cdot b}\right)\cdot(2b-1)^{-1} & \text{if} \quad .5<y<1, b\ne .5\\ y & \text{if} \quad b=.5\end{cases}$$

Figure 1 shows three Bézier bias function for the points b=0, b=0.5, b=1. Note that b< 0.5 leads to an expansion of the scale and b> 0.5 leads to a contraction.

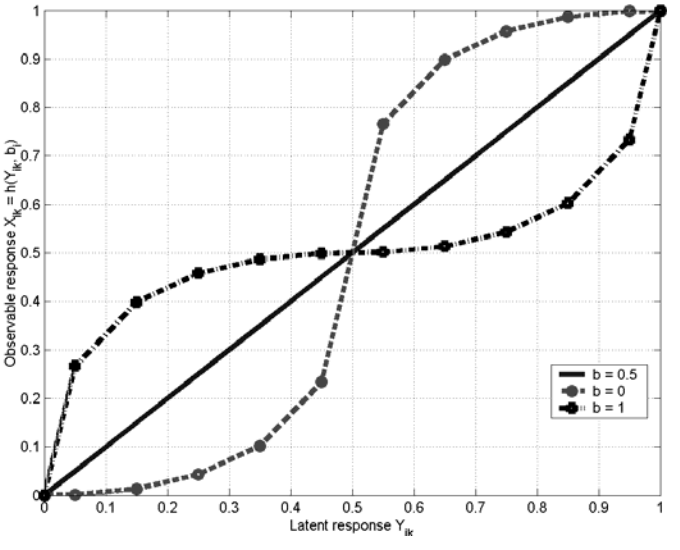


Fig. 1. Three Bézier bias functions satisfying for b= 0, 0.5, and 1

We complete the model by adding a sixth axiom covering bias as follows:

Axiom 6. (Bézier Bias). The manifest response random variables, X_{ik} , are given by $X_{ik} = h(Y_{ik};b_i)$, for parameters $b_i \in [0,1]$, i= 1, ..., N, k= 1, …, M.

2.4 The Distribution of the Manifest Responses

When we assume the beta error distribution in Axioms 5 and the bias process in Axiom 6, the beta distribution does not govern the distribution of the $\boldsymbol{X}_{ik} = h(\boldsymbol{Y}_{ik})$; however, it is easy to calculate the distribution of X_{ik} using standard methods for monotonic transformations of random variables of the continuous type, e.g. [11], and the result is a density of the X_{ik} given by

$$g(x_{ik};D_i,b_i;z_k) = f(h^{-1}(x_{ik};b_i);D_i,z_k)\cdot\left|\frac{dh^{-1}(x_{ik};b_i)}{dx_{ik}}\right|. \tag{3}$$

It is straightforward to compute the components of Eq. (3), and since we will need them for the model inference section they are presented next. The inverse of the bias function $y = h^{-1}(x,b)$ is

$$h^{-1}(x,b) = \begin{cases} \left(b - 2bx + x - \sqrt{2x - 4bx + b^2}\right)(2b-1)^{-1} & \text{if} \quad 0 < x \le 0.5, b \ne 0.5 \\ \begin{pmatrix} 3b - 2xb + x - 2 \\ +\sqrt{2 + b^2 - 4b + 4bx - 2x} \end{pmatrix}\cdot(2b-1)^{-1} & \text{if} \quad 0.5 < x < 1, b \ne .05 \\ x & \text{if} \quad b = 0.5 \end{cases} \tag{4}$$

Further, the appropriate Jacobian is

$$\left|\frac{dh^{-1}(x,b)}{dx}\right| = \begin{cases} \left(\sqrt{2x - 4bx + b^2}\right)^{-1} - 1 & \text{if} \quad 0 \le x \le 0.5, b \ne 0.5 \\ \left(\sqrt{b^2 - 4b + 2 + 4xb - 2x}\right)^{-1} - 1 & \text{if} \quad 0.5 < x \le 1, b \ne 0.5 \\ 1 & \text{if} \quad b = 0.5 \end{cases} \tag{5}$$

Then the likelihood function for the model is given from Eq. (3) by

$$L[< D_i >, < b_i >, < z_k >; (x_{ik})_{NxM}] = \prod_{i=1}^{N}\prod_{k=1}^{M} g(x_{ik};D_i,b_i,z_k)\ . \tag{6}$$

3 Estimation Theory for the Model

When the model is applied to response profile data, $\mathbf{X}=(x_{ik})_{NxM}$, for N informants responding to M items, there are N competence parameters, $\mathbf{D}=<D_i>_{1xN}$, N bias parameters, $\boldsymbol{\beta}=<b_i>_{1xN}$, and M answer key parameters, $\mathbf{Z}=<z_k>_{1xM}$, to be estimated from $N \cdot M$ data points. There are a variety of approaches to parameter estimation; however, because the number of parameters scales up with the number of data observations, standard methods based on maximizing the likelihood function in Eq. (6) can yield problematic estimators. We believe that an appropriate approach to estimating parametric CCT models is to use Markov Chain Monte Carlo (MCMC) methods. We adapted the standard Metropolis-Hastings (M-H Algorithm) as described in [13] to set up a Bayesian fixed-effects analysis of the model in Axioms 1-6. Several papers, e.g. [14], [15], have illustrated the value of using MCMC in estimating item response (IRT) models. CCT models can be viewed as an extension of IRT models where the correct answers are not known a priori but are instead parameters to be estimated. The first use of such methods for CCT models was [12], which provided a Bayesian fixed effects analysis of a model for dichotomous data.

To implement the algorithm, we only need to know the joint posterior density of the parameter vector, up to proportionality constant. This posterior density is proportional to the likelihood function, given in Eq. (6), multiplied by the prior. We used uniform priors for each z_k and b_i, we followed [13] for defining exponential priors for each D_i, and we assumed independence of all the model parameters in the joint prior. Thus, letting $\pi_B(z_k)$ and $\pi_B(b_i)$ be uniform distributions and $\pi_G(D_i)$ be an exponential distribution with mean 24, and using Eq. (6), the posterior distribution for the model is proportional to the following:

$$p[< z_k >,< D_i >,< b_i >| (x_{ik})] \propto \prod_{i=1}^{N}\prod_{k=1}^{M} g(x_{ik}; D_i, b_i, z_k) \prod_{k=1}^{M} \pi_B(z_k) \prod_{i=1}^{N} \pi_G(D_i)\pi_B(b_i) \quad (7)$$

We constructed a suitable auxiliary function (candidate generating function) and applied standard ways of assessing convergence again following the approach in [13]. In our case as in most applications of the Metropolis-Hastings to IRT type models the simulation results did not depend strongly on the choice of the diffuse prior.

4 Performance of the Model

4.1 Simulations

We undertook a simulation study to assess the accuracy of the MCMC estimation method. We simulated response profile matrices $\boldsymbol{X}= (X_{ik})_{NxM}$ from the model as described in Axioms 1-6 with the density given in Eq. (7). We considered both the case of no response bias ($b_i = 0.5$ in Axiom 6) in one simulation, and in a second simulation we introduced the possibility of bias. The simulated data matrices consist of $N = 4$ respondents and $M = 40$ items, and the true parameters were spaced uniformly in their corresponding ranges. We conducted 150 simulations for data with bias and 150 simulations for data without bias. In each simulation, we obtained a 4x40 response profile matrix, $\mathbf{X}=(x_{ik})_{4x40}$, and for each we used the Metropolis-Hastings algorithm to make 20,000 draws from the posterior distribution discarding the initial 5,000 draws. We calculated parameter estimates from the marginal posterior distributions in several different ways, and we were satisfied that the MCMC algorithm was performing well in estimating all the parameters of the model. In particular the recovery of the z_k was much better than a simple, descriptive arithmetic mean of the X_{ik} for each item, and the estimation scheme was able to pick up differential competence and bias parameters among the four stat-informants.

4.2 Real Data

To study the performance of the model with real data, we analyzed the data reported in a cross-cultural study (see [16], [17]). The purpose of the study was to compare the perceptions of the similarity of emotion terms between English and Japanese monolingual speakers. The dataset consisted of N= 65 respondents and M= 105 items. The respondents fell into two groups: 33 monolingual English speakers interviewed in the United States and 32 monolingual Japanese speakers interviewed in Japan. The

items consisted of all pairs from 15 selected terms to describe emotions, e.g. bored (tsumaranai), fear (osoroshii), and shame (hazukashii) (see [16] for a complete list). The English terms were translated into Japanese and then back translated by experts to assure correspondence. Each respondent was required to rate each of the 105 pairs of emotion terms for perceived similarity on a five point likert scale, with 1 most dissimilar and 5 most similar.

We estimated the model for the English and Japanese respondents separately using the model with bias and error distribution described in Axioms 1-6. The model, as discussed above, assumes continuous responses, but the data in this dataset, as in many other datasets, are ordinal. Thus, before estimating the parameters, we had to adopt an MCMC estimation procedure to handle ordinal responses. We tried several approaches including simply entering the actual scores rescaled to the unit interval. Based on simulation studies we were satisfied that in the current case the simple approach was a suitable approximation to using continuous scores.

Table 1. Summary statistics of point estimates under MCMC

Parameter	Culture	Mean	Median	Std.
D_i	English	9.2558	9.2364	3.5471
	Japanese	9.3338	9.2779	3.7627
b_i	English	0.2769	0.2385	0.1656
	Japanese	0.3441	0.2771	0.205

Note: Each point estimate of a parameter is calculated as a descriptive statistic from the corresponding marginal posterior obtained from the MCMC algorithm.

Table 1 provides some summary statistics of the posterior distribution of the competence and bias parameters for the English and the Japanese samples. Both cultures had about equal mean competence, though for the Japanese, there was more variance in competence. Most bias parameters in both cultures were less than 0.5, indicating a prevailing bias toward using the extremes of the scale. In addition the estimates of the z_k in both cultures were highly correlated.

As mentioned, the main purpose of the original study [16] was to compare the perceptions of similarity among the emotion terms between the English and Japanese samples. The methodology in [16] was to subject the similarity ratings to a three-way correspondence analysis as developed in [18] and [19]. The approach allows each respondent and each item to be located in a best-fitting multi-dimensional Euclidean space. In the original study, the mean placement of each of the 15 emotion terms was calculated separately for each culture. Then the closeness of these placements was compared, term-by-term, to see how much the two cultures differed. The authors of [16] and [17] concluded that the two cultures had quite similar representations of the emotion terms. The similarity between the cultures was also indicated by the fact that the relative configurations of the 15 terms were very similar, with the same clusters showing up in both sets of data. Of possible interest, the word “shame” was the term that was perceived most differently than the other terms by the two cultures.

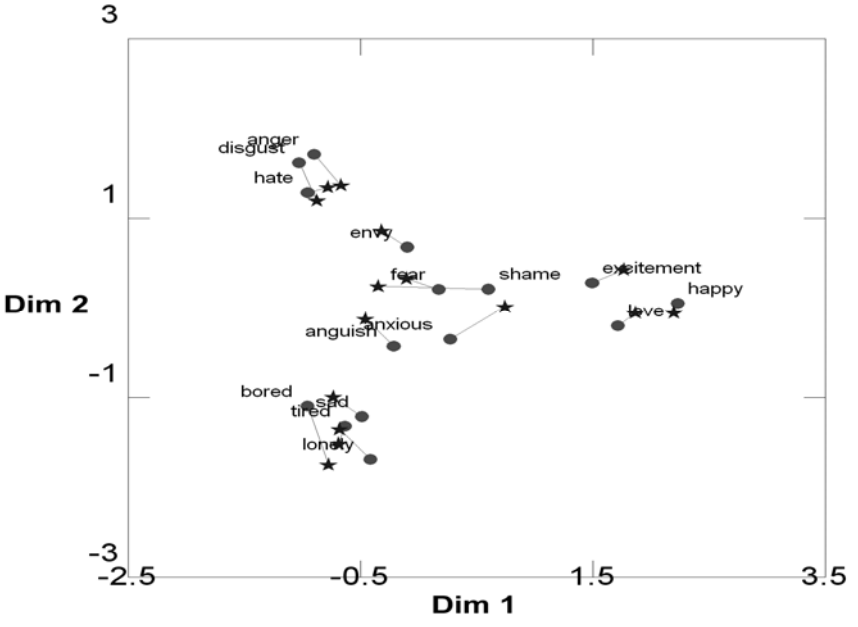


Fig. 2. Three A representation of the mean placements of the 15 emotion terms for the English (star) and Japanese (circle) groups in the first two dimensions based on principle component metric multidimensional scaling of the means of the 105 consensus answer key parameters

In the correspondence analysis methodology for cross-cultural comparison, there was no differential weighting of respondents for competency nor was there any explicit modeling of possible response biases, both of which are features of our model. In order to compare our results with those of the original study, we subjected the 105 estimated answer key parameters from both groups of respondents to separate metric multidimensional scaling analyses based on principle components factoring of the metric similarities. Each group was scaled separately, and then the two multidimensional spaces were merged in the first three dimensions just as in [16] using correspondence analysis. The results of the new analysis are displayed in Fig. 2 for the first two dimensions scaled to allow comparison with the results of the metric scaling of the consensus answers from both cultures.

Several things emerge immediately from the figure. First, just as in the original study [16], there is a great deal of similarity between the semantic representations of the two cultures. The line lengths comparing the two cultures on each term are much shorter than would occur with random placements of two points in the plot in Fig. 2, and there is a fairly close correspondence in the clusters that emerge, in the first two dimensions as revealed in the plot in Fig. 2 compared to the clusters found in [16]. In addition "shame" is the longest line in the plot. Finally similar comparisons with the original study occur for the first and third and second and third dimensions as well. There is one important difference that emerges in comparing the two analyses. The line lengths in the analysis based on the scaling of the CCT parameters were shorter than for the original correspondence analysis in 80% of the cases. This may be due to

the fact that the model gives more weight to informants with more cultural competence. Perhaps even more important, the incorporation of bias parameters in the models may have taken out some of the noise in the data produced when different respondents use the ordinal scale differently.

Based on our simulations and analysis of real data, we recommend that future cross- cultural studies of semantic systems use the CCT model first and then scale the consensus answer key parameters to compare two cultures. Having a cognitive model rather than using multidimensional scaling display techniques offers a deeper understanding of the data. Future work should focus on ways to collect cross-cultural data using a continuous response scale rather than the ordinal Likert scale used in the current study. This will allow the full capacity of the model to be used.

Acknowledgment. Work on this paper was supported by US Air Force Office of Scientific Research Award Number: FA9550-09-0510 to the first author.

References

1. Ariely, D., Au, W.T., Bender, R.H., Budescu, D.V., Diez, C.B., Gu, H., Wallsten, T.S., Zauberman, G.: The Effects of Averaging Subjective Probability Estimates Between and Within Judges. J. Exp. Psychol-Appl. 6, 130–147 (2000)
2. Wallsten, T.S., Diederich, A.: Understanding Pooled Subjective Probability Estimates. Math. Soc. Sci. 41, 1–18 (2001)
3. Johnson, V.E., Albert, J.H.: Ordinal Data Analysis. Springer, New York (1999)
4. Patz, R.J., Junker, B.W., Johnson, M.S., Mariano, L.T.: The Hierarchical Rater Model for Rated Test Items and its Application to Large-scale Educational Assessment Data. J. Educ. Behav. Stat. 27, 341–384 (2002)
5. Batchelder, W.H.: Cultural Consensus Theory: Aggregating Expert Judgments About Ties in a Social Network. In: Liu, H., Salarno, J., Young, M.J. (eds.) Social Computing, Behavioral Modeling, and Prediction, pp. 24–32. Springer, New York (2009)
6. Romney, A.K., Batchelder, W.H.: Cultural Consensus Theory. In: Wilson, R., Keil, F. (eds.) The MIT Encyclopedia of the Cognitive Sciences, pp. 208–209. The MIT Press, Cambridge (1999)
7. Shepard, R.N.: Psychological Relations and Psychological Scales: On the Status of "Direct" Psychological Measurement. J. Math. Psychol. 24, 21–57 (1981)
8. Lord, F.M., Novick, M.R.: Statistical Theory of Mental Test Scores. Addison-Wesley, Reading (1968)
9. de Casteljau, P.: Courbes et Surfaces a Poles. Technical report, A. Citroen, Paris (1963)
10. Bezier, P.: Essay de Definition Numerique des Courbes et des Surfaces Experimentales. Ph.D. Thesis, University of Paris VI (1977)
11. Hogg, R.V., Craig, A.T.: Introduction to Mathematical Statistics. Macmillan, New York (1978)
12. Karabatsos, G., Batchelder, W.H.: Markov Chain Estimation Theory Methods for Test Theory Without an Answer Key. Psychometrika 68, 373–389 (2003)
13. Carlin, B.P., Lewis, T.A.: Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC, Boca Raton (1998)
14. Patz, R.J., Junker, B.W.: Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing data, and Rated Responses. J. Educ. Behav. Stat. 24, 342–366 (1999a)

15. Patz, R.J., Junker, B.W.: A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models. J. Educ. Behav. Stat. 24, 146–178 (1999b)
16. Romney, A.K., Moore, C.C., Rusch, C.D.: Cultural Universals: Measuring the Semantic Structure of Emotion Terms in English and Japanese. P. Natl. A. Sci. USA. 94, 5489–5494 (1997)
17. Romney, A.K., Moore, C.C., Batchelder, W.H., Hsia, T.: Statistical Methods for Characterizing Similarities and Differences Between Semantic Structures. P. Natl. A. Sci. USA. 97, 518–523 (2000)
18. Romney, A.K., Moore, C.C., Brazill, T.J.: Correspondence Analysis as a Multidimensional Scaling Technique for Non-frequency Similarity Matrices. In: Blasius, J., Greenacre, M.J. (eds.) Visualization of Categorical Data, pp. 529–546. Academic Press, New York (1998)
19. Kumbasar, E., Romney, A.K., Batchelder, W.H.: Systematic biases in social perception. Am. J. Sociol. 100, 477–505 (1994)

Information Overload and Viral Marketing: Countermeasures and Strategies

Jiesi Cheng[1], Aaron Sun[1], and Daniel Zeng[1,2]

[1] Department of Management Information Systems, University of Arizona, Tucson, Arizona
[2] The Key Lab of Complex Systems and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, China

Abstract. Studying information diffusion through social networks has become an active research topic with important implications in viral marketing applications. One of the fundamental algorithmic problems related to viral marketing is the Influence Maximization (IM) problem: given an social network, which set of nodes should be considered by the viral marketer as the initial targets, in order to maximize the influence of the advertising message. In this work, we study the IM problem in an information-overloaded online social network. Information overload occurs when individuals receive more information than they can process, which can cause negative impacts on the overall marketing effectiveness. Many practical countermeasures have been proposed for alleviating the load of information on recipients. However, how these approaches can benefit viral marketers is not well understood. In our work, we have adapted the classic Information Cascade Model to incorporate information overload and study its countermeasures. Our results suggest that effective control of information overload has the potential to improve marketing effectiveness, but the targeting strategy should be re-designed in response to these countermeasures.

1 Introduction

The recent explosively-growing popularity of social communities such as Digg.com and Twitter.com has attracted remarkable attention and hilighted the Internet as a vital medium for marketing and advertising. These Web sites provide a rich set of social networking features to encourage information sharing among users. Viral marketing is also emerging as an important means to stimulate the awareness and adoption of products or services. However, despite the increasing emphasis on viral marketing, it is also widely recognized that as the costs of generating and transmitting information are almost neglectable, these online communities become increasingly information-saturated for running viral campaigns. As a result, information overload occurs as messages are arriving in numbers larger than what users can process, which generates undesired results on both marketer and user sides. Profit-seeking marketers broadcast messages,

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 108–117, 2010.

but only a small proportion of these messages can catch the attention of users - with a vast majority of messages "wasted". On the other hand, due to the limited information-processing capability, it becomes increasingly difficult for users to discover relevant messages, let alone recommending them to other users.

To effectively address the information overload issue arising in this online social context, many countermeasures have been proposed in the literature. Some representative methods include applying filters to screen out irrelevant contents, or manipulating the communication cost to regulate the volume of information. These methods, although differing in operational details, share the same objective to balance the online information supply and demand. Numerous studies have assessed these countermeasures from an information-recipient perspective, and demonstrated their effectiveness in mitigating the information load on users. However, less attention is given to understanding how the alleviation of information overload can in turn affect online marketers, particularly viral marketers. In this paper, we model the information diffusion pattern within a major online social community - Digg.com. A probabilistic model adapted from a well-known diffusion model - Independent Cascade Model (ICM) - is developed to simulate the effect of information overload as well as its countermeasures. The proposed model is then used for re-examining the key viral marketing question concerning initial nodes to be targeted for maximizing the influence of viral campaigns. Our results show that larger influence can be achieved after the implementation of either countermeasures. However, targeting strategies should be adjusted in response to these changes.

The rest of this paper is organized as follows. We start with reviewing relevant viral marketing and information overload literature in Section 2. In Section 3, we briefly introduce the functionalities of Digg.com and our dataset components. We then present our diffusion model and apply it to model the social network structure and diffusion patterns identified from the dataset in Section 4. Modeling results and evaluations are also discussed in this section. Finally, conclusions and future directions are discussed in Section 5.

2 Literature Review

2.1 Viral Marketing and Influence Maximization

Viral marketing originates from the word-of-mouth (WOM) advertising widely studied in the last few decades. WOM advertising refers to the informal communication between two or more persons concerning a product or service on a non-commercial basis [1]. The advent and popularity of various Internet-based communication tools lead to the proliferation of viral marketing as a new interpretation of WOM advertising in the Internet era [2]. Viral marketing utilizes (mostly) the Internet to transmit and spread viral messages among individuals, which has been found to be of several significant benefits. One important benefit is that viral marketing is relatively inexpensive in comparison to other traditional mass media communications. With a considerably low cost, viral marketing can enable the advertising messages to reach wide audiences within a short period

of time [3]. Such rapid diffusions can substantially boost the speed of adoption of the promoted product or service. Viral communication also provides a message delivery medium that is more intimate and personalized, and advertising messages may be viewed more favorably by the recipient, thereby increasing the likelihood of accessing to "hard-to-get" audience members [4].

The success of viral marketing campaigns critically depends on a variety of factors. One of them is the initial targeting (seeding) [5]. In this procedure, the marketer is supposed to estimate for the extent to which individuals influence one another within the particular community where the campaign is launched. Following the estimation process, a group of influential "seed" members are initially targeted with the hope that they can trigger a larger cascade of influence subsequently. In [6], Domingos and Richardson formulated the viral seeding process as a combinatorial optimization problem. In this so-called *Influence Maximization* (IM) problem, the *influence* of a set of individuals A is defined as the expected number of individuals who finally adopt the promoted product or service, given that A is chosen as the initial seeds of the viral campaign. The IM problem then asks, considering the resource constraints, how the influence can be maximized when only no more than k individuals can be chosen for seeding.

The IM problem offers valuable insights into the cost/benefit aspects of a viral campaign. The problem can be formulated and solved over a number of viral marketing models, with the Independent Cascade Model (ICM) being one of the simplest and most popular models [5]. The ICM investigates the information diffusion processes underlying the viral campaign on a directed graph G. Each individual node is either in the state of active (an adopter of the product) or inactive. Nodes can switch from being inactive to being active, but not reversely. The diffusion process is modeled as a progressive case in discrete steps: when node u first becomes active in step t, u can activate each currently inactive neighbor v - with a successful probability p_{uv}. If u succeeds, then u will become active in step $t + 1$. Note that whether or not v succeeds, it cannot make any further attempts to activate u subsequently. The entire process of diffusion starts with an initial set of active nodes, and repeats until no more activations are possible. The objective of the IM problem is then to determine the optimal set of initial targets, to maximize the total number of activations.

2.2 Information Overload and Its Countermeasures

Most of the prior WOM and viral marketing studies have viewed that individuals interact and communicate on a face-to-face or one-to-one basis [7]. In reality, the Internet has significantly lowered the communication cost, and expanded the reach and availability of individuals. This in turn results in an enormous increase in the volume of messages one can send and receive, which can potentially produce adverse information overload effects. Information overload refers to the situation that the amount of information supply exceeds the information-processing capacity of an individual [8]. For example, in the application of email marketing, "email overload" [9] was commonly recognized as one of the main reasons for the failure of many email marketing campaigns. When the volume of

incoming emails exceeds one's processing capacity, the person tends to become highly selective and ignore those marketing emails. Similar observations were made in SMS marketing applications as well.

To alleviate the information overload syndrome, considerable research has been conducted for proposing effective countermeasures. Two major research directions can be distinguished, namely, a filter-based [10] approach and a cost-based [11] approach . The former is the most common method to combat heavy load of information, which adopts sophisticated filtering techniques to identify and block unwanted messages based on user customization. The cost-based approach, on the contrary, shifts the task of screening message from recipients to senders. The key insight of this control mechanism is to impose a fixed or a market-determined cost on each sent message. As such, advertisers are expected to be more selective in choosing recipients and send messages more rationally.

Both approaches have been extensively studied and evaluated under different settings. Nevertheless, prior studies have primarily focused on individual user experiences with these countermeasures. When the amount of wasteful information exposed to each user is successfully reduced, how these countermeasures can in turn affect online marketers remains to be unclear, especially in a viral marketing context. In the following sections, we built a diffusion model based on the standard ICM to illustrate the information diffusion patterns identified in a popular online social community - Digg.com. We then incorporate the information overload countermeasures into the model to quantitatively examine their impacts on viral marketing planning and outcomes.

3 Data Set

3.1 Introduction to Digg.com

Digg.com is arguably the most popular social news aggregator, which allows its users to share/discover news stories. The basic functionality of Digg can be described as follows: users submit links to stories they discover from the Internet. A new submission immediately appears on a repository Web page called "Upcoming Stories", where other users can vote on – digg on – the submitted story if they like it. If a submission manages to earn a critical mass of diggs in a short period of time, it becomes popular and is promoted to the front page. Digg is considered to be social because a user can designate any other user as a friend, and in the future track these friends' submissions and digg activities. In this study, for user v who adds user u as a friend, we call u as one of v's friends, and v as one of u's fans. Digg then provides v a *friends history page* displaying **all** the digg activities from v's subscribed friends in a reverse-chronological order. Note that the connection itself is directed, which means that u cannot track v's activities, unless a reverse friend link is set up.

Such a network-based information-sharing architecture can potentially lead to the occurrence of information overload, as was demonstrated in previous studies [12]. In Digg and other social communities alike, users tend to track their friends' activities in a "batch" mode. It frequently happens that when one user checks her

friends history page, multiple newly-updates are simultaneously listed (especially for those who have many friends.) Constrained by a limited attention span, this user can hardly pay equal attention to each news story, and it is likely for a specific relevant message to be buried in a flood of irrelevant ones. As a consequence of this batching effect, the spreading of any single story in the network cannot be treated as an isolated event.

3.2 Data Elements

We have collected data from Digg in the category "apple." Digg has predefined 50 different categories with each exclusively devoted to a specific topic. The category of "apple" covers news stories that are related to Apple products. We have sampled 53 popular stories that have been digged more than 1,000 times. Altogether, these 53 URLs included 117,105 total diggs from 13,082 unique users, which comprised our first dataset "URL." We continued to construct our second dataset "people" which consisted of users' individual histories. A complete user history includes this user's friends and fans, and digg histories such as digged URLs and date of digg. Both the "URL" and "people" datasets constitute a basis for the estimation of model parameters.

4 Modeling Information Overload and Its Countermeasures

4.1 General ICM

We now formally present our model called the General ICM - GICM, which has its origins in the ICM. GICM also investigates diffusion processes in a directed network $G = (V, E)$. Unlike the ICM which is limited to modeling one single message, the GICM allows multiple messages to simultaneously propagate through the network G. This extension represents the most prominent difference between the two models, which enables the GICM to model the information-rich nature of online environment as well as the effect of information overload. Obviously, the fact that a number of messages co-exist on the social network causes that each individual node can be activated by different messages. In case there exist m distinct messages under study, we use an m dimensional vector to represent each node u's ($u \in V$) state $\overrightarrow{s_u} = (s_u^1, s_u^2, ..., s_u^m)$. Each node maintains a default *inactive* state $\overrightarrow{s_u} = 0$. When u adopts the content conveyed by message i, the i^{th} element of the state (s_v^i) switches to 1 indicating that the node has been *activated* by message i. For simplicity, we consider such an activation process as unidirectional and irreversible. Each directed edge $e = (u, v)$ in the edge set E indicates the direction of message flow from u to v. When a message is sent from u to v through $e = (u, v)$, u is called the sender and v is the recipient. The complete set of recipients of u is denoted as $r(u)$, and $s(v)$ denotes the complete set of senders who can send messages to v.

The information diffusion process can now be modeled on the GICM in discrete steps. First of all, for each node u in the network, we use a system parameter

$\pi_u \geqslant 1$ to indicate u's message sending capacity, namely, π_u equals the maximum number of messages u can generate in one time step. We assume that in any time step, u would operate at its full capacity to send π_u messages to each of its receiver. Now, in step t, node v can receive a sequence of incoming messages $in_v(t)$, where $|in_v(t)| = \sum_u \pi_u$, $u \in s(v)$, $\forall t$. This sequence of messages are presented in an arbitrary order and they sequentially attempt to influence and activate v, with a certain probability. The probability for v to be activated by any message $msg_{in} \in in_v(t)$ depends upon msg_{in}'s position in the sequence. We assume that the activation probabilities for any two messages in separate positions are mutually independent. We use $act_v(t)$ to denote the set of messages that have activated v in step t, where $act_v(t) \subseteq in_v(t)$ and $|act_v(t)| \leqslant \pi_v$. The reason to limit the size of $act_v(t)$ is that in the next $t+1$ step, v will be given a single chance to activate all v's inactive neighbors $w \in r(v)$, with respect to each of the message in the set of $act_v(t)$. Therefore, the size of $act_v(t)$ should not exceed v's sending capacity π_v. On the contrary, if the size of $act_v(t)$ is less than π_v, an extra $\pi_v - |act_v(t)|$ number of new messages will also be sent by v to reach the full capacity. This assumption can be understood that in Digg, users not only digg messages received from their friends, but also submit new messages themselves.

In a viral marketing context, tracking all available messages on the network is unnecessary since a majority of messages are irrelevant to the viral campaign. We further introduce the distinction between the *marketing* message and the *regular* message, with a marketing message carrying advertising information that could possibly trigger the adoption of the advertised product or service, and the regular message carrying no commercial information. As such, our modeling objective can be simplified by only monitoring the activation of marketing messages.

4.2 Modeling Information Overload

From the decision making perspective, when information overload occurs, the relationship between the amount of information provided and the amount of information utilized by decision maker can be represented as an inverted U-curve [8] (See Figure 1a). As is illustrated in this figure, increasing the volume of incoming information (x-axis) can increase the amount of information being utilized (y-axis) up to a certain point. However, beyond this particular point, the utilization level starts to decline due to the limited information-processing capacity, or other dysfunctional effects, such as anxiety and confusion.

In Digg and other social communities alike, a similar inverted U-curve pattern can describe users' behavior in processing incoming messages. In [12], the authors discussed how the positioning of a specific message can affect its likelihood of activation. When incoming messages are ordered sequentially by their arrival time, the authors claimed that a predominant portion of activations can be attributed to messages listed in the first several positions. Based on such observations, the probability of activation can be defined as follows. Given a sequence of incoming messages and a particular message ranked at position s (the smaller s, the higher the rank,) and use len to denote the total number of messages in the sequence.

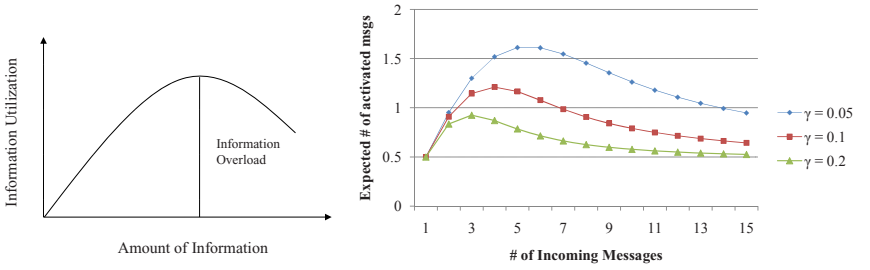


Fig. 1. a. Inverted U-curve Pattern b. Simulated Inverted U-curve Pattern ($\bar{p}$=0.5)

We further set a random variable $X_s = 1$ if this message successfully activates the target, and 0 otherwise. Then the probability of message s being activated, denoted as $p(X_s = 1)$, satisfies (a) $p(X_s = 1)$ is inversely proportional to the length of the entire sequence, $\forall s$, (b) $p(X_s = 1)$ drops exponentially when s increases. To be more specific, $p(X_s = 1) = \bar{p} \cdot \exp(-\gamma \cdot len \cdot s)\,(s \geqslant 1)$. In this expression, $\bar{p}$ equals the probability of activation in the absence of information overload. The latter part represents the load of information, whose intensity is controlled by a scaling parameter $\gamma > 0$. In addition, since all the messages enter the sequence randomly, we can assume that the chance for a message to be located in any position is $1/len$. Consequently, the probability for a message to be activated is $\bar{p} \cdot \dfrac{1}{len} \cdot \sum_s \exp(-\gamma \cdot len \cdot s)\,(s \geqslant 1)$, which is shown as a clear inverted U-curve with increased length of message sequence (Figure 1b).

4.3 Filter-Based Countermeasure

In this sub-section, we first integrate the filtering function into the basic GICM to explore the effects of user filters on viral marketing. To start with, we assume that each user under study has installed a content-based filter whose output can be adjusted based on user preference. Next, we further assume that marketing messages are of users' interest and thus will not be filtered. This can be achieved by carefully selecting the user community where the campaign is launched (e.g. spreading Apple-related ads in the community of Apple fans.) As a result, all of the regular messages received will be captured by the filter, and only a certain proportion of them can pass the screening process and be presented to the recipient. Although such content-based filters are currently not available in Digg, users' digg histories provide trustworthy indications of user preferences that can help approximate the outcomes of filtering. There exist 50 different system-defined topic categories in Digg, and each news story falls into exactly one of these categories. By tracking one user's digg histories, if a topic category receives enough diggs from the user (i.e. more than a given threshold value,) we then label this category as one of this user's favorites, and so do stories under this category. It is then fair to assume that, user u only diggs stories that belong to u's favorite categories fav_u. In addition, if a content-based filter had been

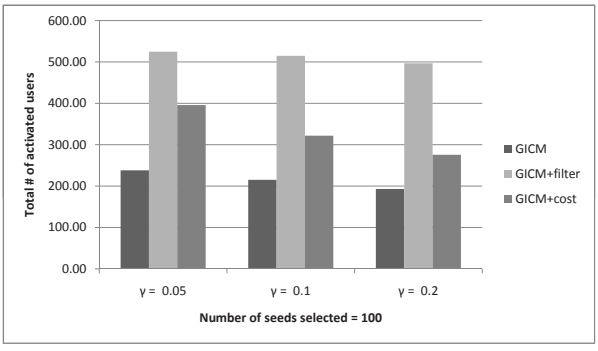


Fig. 2. Total # of Activated Users by the Three Simulation Algorithms

installed, for each user $v \in r(u)$, only stories that belong to fav_v would have passed the filter, with the rest being blocked.

Following the above discussion, we analyzed each user u's digged categories ctg_u, as well as the number of diggs fell into each category c, denoted by $digg_c$. The largest connected component of the user community - a network consisting of 6,656 unique users was extracted from the people dataset. The threshold for classifying favorites and non-favorites is defined as follows. If $digg_c$ is larger than the mean number of diggs that all categories receive, namely, $digg_c \geqslant \frac{1}{|ctg_u|} \sum_{c \in ctg_u} digg_c$, then category c is one of the favorites. We also defined another two thresholds by adding/subtracting one standard deviation of the mean to/from the first threshold value.

We then adopted a greedy hill-climbing algorithm proposed in [5] to identify the k-node initial active set. Each selected node in the set sends a marketing message at the first step, and no further marketing messages are introduced into the system thereafter. The algorithm starts with the empty set, and repeatedly adds a node that gives the maximum influence gain. This greedy algorithm has been proved in [5] to be an effective heuristic solution that is provably within a factor of at least 63% of the optimal. The results illustrated in Figure 2 indicate that, the existence of content-based filters screens recipients' incoming messages and results in larger activation probabilities and a higher influence when the same number of seeds are selected. In addition, we compare the targeting strategies with/without implementing the filtering mechanism. In the former situation, users who have a relatively large number of non-overloaded recipients are more favored by the marketer. While in considering the filtering effect, users who have a relative large number of narrow-interest fans become better targets.

4.4 Cost-Based Countermeasure

Digg and other social communities alike provide an environment in which the marginal cost of sending one message is almost zero. Appropriate pricing strategies can be applied to impose a mandatory per-message cost on senders. Due to

the increased communication cost, message senders are forced to behave more rationally by sharing messages with those potentially interested recipients only. Much of the previous literature has treated the price of the message as a monetary cost, such as charging postage for email communications. However, we have not observed the real-world use of such a pricing scheme in major online social communities thus far, partly because of its complexity in implementation [13] and the difficulty in convincing people to transit from from free to paid services [11]. In reality, the price of sending a message is also determined by other non-monetary factors, such as time and effort needed. In Digg.com, there exists an supplementary information-sharing function through which users can only designate message recipients by typing names manually. The operations of "send to all" and "copy&paste" are not provided so that senders can rarely afford the time and effort required to reach every possible recipient.

In this subsection, we incorporate such a pricing policy into the basic GICM. Since the recipients are selected through manual inputs, we assume that users are more likely to type names of their frequently-contacted acquaintances. We define the frequent-contact of each user u through tracing u digg history. For any news story s digged by u, if user $v \in r(u)$ diggs the same story after u, the frequency of contacts between u and v, denoted by f_{uv} then increases by one. If f_{uv} is larger than a given threshold value, v is then considered as one of u's frequent contacts. As before, the threshold is defined by averaging f_{uv} over all u's fans v. In the meantime, additional adjustments are made to each user's sending capacity. We assume that it incurs one unit of effort to send one message to one single recipient. In the previous setting, sending one message to multiple recipients does not add any extra effort. However, when one message is sent to $r(\bar{u})$ manually selected contacts, the effort increases by $r(\bar{u})$ times than before. Given that the total available time and other constraints remain the same, the sending capacity thus decreases by the same factor.

We adopted the same algorithm described in Section 4.3 to identify the initial targeting set. The results illustrated in Figure 2 indicate that, by imposing non-trivial costs on sending messages, messages are essentially circulated in a smaller network than before but generate larger influences. In such a case, our targeting strategy shows that active users (i.e. users who can afford the time and effort to send more messages out) become favorable candidates.

5 Conclusions and Future Directions

In this study, we proposed a novel diffusion model that considers information overload and its countermeasures. By removing the "one-message-at-a-time" constraint, our model can provide more accurate and flexible characterization of the information diffusion processes. We then conducted simulation studies to address the Influence Maximization problem which is critical to the success of viral marketing campaigns. Our study has direct implications for the management of viral campaigns. Policy-makers should recognize the importance of mitigating excessive amount of information and viral marketers should accordingly adjust their

marketing strategies for maximizing influences. Future work along this direction could examine the effectiveness of the proposed targeting strategies in real-world settings, and improve the current simplistic model by relaxing assumptions such as sending capacity and attention span.

Acknowledgement

Research reported in this paper is partly supported by the Chinese Academy of Sciences (#2F07C01, #2F08N03) and the NNSFC (#90924302 and #60621001).

References

1. Arndt, J.: Role of product-related conversations in the diffusion of a new product. Journal of Marketing Research 4(3), 291–295 (1967)
2. Subramani, M.R., Rajagopalan, B.: Knowledge-sharing and influence in online social networks via viral marketing. Communications of the ACM 46(12), 300–307 (2003); Social information processing theory
3. Leskovec, J., Adamic, L.A., Hubermans, B.A.: The dynamics of viral marketing. ACM Transactions on the Web 1(1), Article 5 (2007)
4. Bampo, M., Ewing, M.T., Mather, D.R., Stewart, D., Wallace, M.: The effects of the social structure of digital networks on viral marketing performance. Information Systems Research 19(3), 273–290 (2008)
5. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (2003)
6. Domingos, P., Richardson, M.: Mining the network value of customers. In: Seventh International Conference on Knowledge Discovery and Data Mining (2001)
7. Bruyn, A.D., Lilien, G.L.: A multi-stage model of word-of-mouth influence through viral marketing. International Journal of Research in Marketing 25, 151–163 (2008)
8. Eppler, M.J., Mengis, J.: The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. The Information Society 20(5), 325–344 (2004)
9. Whittaker, S., Sidner, C.: Email overload: exploring personal information management of email. In: Proceedings of the SIGCHI conference on Human factors in computing systems (1996)
10. Reshef, E., Solan, E.: The effect of filters on spam mail. tech. rep., School of Mathematical Sciences, Tel Aviv University (2005)
11. Kraut, R.E., Sunder, S., Morris, J., Telang, R., Filer, D., Cronin, M.: Markets for attention: Will postage for email help? In: CSCW (2002)
12. Sun, A.R., Zeng, D.D.: Maximizing influence through online social networks. In: Proceedings of the 18th Workshop on Information Technologies and Systems (2008)
13. Zandt, T.V.: Information overload in a network of targeted communication. RAND Journal of Economics 35(3), 542–560 (2004)

Using Model Replication to Improve the Reliability of Agent-Based Models

Wei Zhong[1] and Yushim Kim[2]

[1] Arizona State University, Phoenix, AZ, USA
wzhong1@asu.edu
[2] Arizona State University, Phoenix, AZ, USA
ykim@asu.edu

Abstract. The basic presupposition of model replication activities for a computational model such as an agent-based model (ABM) is that, as a robust and reliable tool, it must be replicable in other computing settings. This assumption has recently gained attention in the community of artificial society and simulation due to the challenges of model verification and validation. Illustrating the replication of an ABM representing fraudulent behavior in a public service delivery system originally developed in the Java-based MASON toolkit for NetLogo by a different author, this paper exemplifies how model replication exercises provide unique opportunities for model verification and validation process. At the same time, it helps accumulate best practices and patterns of model replication and contributes to the agenda of developing a standard methodological protocol for agent-based social simulation.

Keywords: Agent-based model, verification, validation, model replication.

1 Introduction

Agent-based models (ABM) have become increasingly popular as a modeling approach in social sciences. While social scientists commit themselves to developing new models to understand the complex social phenomenon, they are persistently confronting the challenge of how to verify and validate their models. As a response, the ABM community identifies model replication as an important and unique venue to approach the challenge. Motivated by the need to improve the reliability of ABM as a standard tool in social science, this paper attempts to replicate an ABM that was developed to understand a complex social problem. Besides developing a case of model replication which may contribute to the agenda of establishing general principles for social simulation, we demonstrate how the verification and validation process can be supported by model replication. In the next section, we briefly explain key concepts in modeling and discuss why model replication might be beneficial, particularly for ABM verification and validation. We then illustrate an example, replicating an ABM for a public service delivery program and evaluate the replication effort. Finally, we conclude

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 118–127, 2010.

the paper with a brief discussion of challenges of ABM replication and how the ABM community can benefit from model replication.

2 Key Concepts in Modeling

An objective of agent-based social simulation is to represent social processes, and the purpose is to better understand such complex processes [19]. Simulation helps social scientists because there are some commonalities between reality and the simulation model, although such similarities are rarely direct [4,6]. What mediates the two is a conceptual model, reflecting how modelers frame the social process, phenomenon, or system [5]. Figure 1 presents the relationships among key concepts in modeling that are particularly relevant in social simulations.

Wilensky and Rand [23] clarified what is a model, conceptual model, and implementation model: a model is defined as "a simplified representation of a real-world process or object", a conceptual model refers to "some description, often textual, of a real-world process or object that is not executable and thus has some ambiguities in regards to how to map inputs to outputs of the target system", and an implementation model is a formalization of the conceptual model "into a computational format so that the model can be given input and generate output". The implementation model is equivalent to a simulation model that runs in a particular computer setting using specific codes. Agent-based models are one type of implementation model based on different programming orientations.

Since model building is a process of representing a complex reality in a simplified manner, modelers aim to build a model right (verification) or build the right model (validation) [16]. In the modeling literature, calibration refers to the process of tuning the implementation model by adjusting a set of less well-known

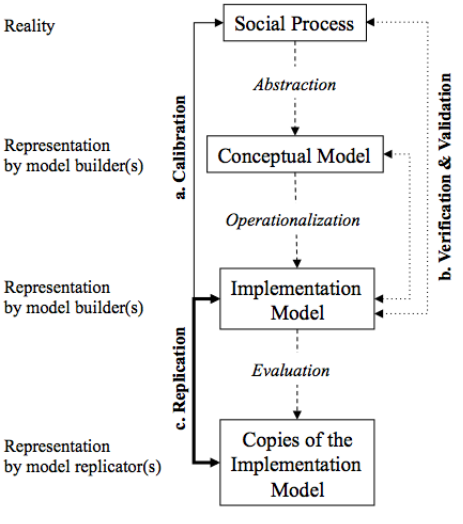


Fig. 1. Concepts and key terms in modeling

parameters. The set of parameters is often obtained by matching the simulated distribution with the observed distribution in a particular context [8,22]. Verification is the process of making sure that the implementation model correctly implements its conceptual model [17]. If the correspondence between simulation output and the textual or graphical description in a conceptual model could be established, one could say that the operationalization of the conceptual model is not incorrect. On the other hand, verification does not necessarily imply validation, which refers to the process of determining whether the implementation model represents the social phenomenon in the real world [2]. Replication is an activity to create another implementation model of the same conceptual model in different computing environments by a different modeler [23].

2.1 Introducing Model Replication into ABM Verification and Validation

It is well-recognized that the verification and validation of ABM for complex social phenomena presents special challenges because social processes are not physical processes for which the model can be evaluated [14,19,20]. Even if there is a physical system observable, numerical models can only be evaluated in relative terms; verification and validation are not easily achieved [15]. Agent-based social simulation models usually involve a complex dynamic social system where counterintuitive properties could emerge. Furthermore, a simulated process itself is complex. Researchers "discover" the "output" after simulation [6]. Without thorough examination, they rarely know the causes for those unexpected outputs from modeling. It could be a logic error, a misrepresentation between the conceptual and implementation models, or simply a consequence of the model itself.

While verification and validation have gained attention in modeling communities, an interest in replication is more recent. In the natural sciences, replication is necessary to conduct scientific research, while less replication of reported results is found in other areas, such as marketing and health-related fields [13]. In the community of artificial society and simulation, replication is sometimes considered as a student learning activity [18]. Only recently have researchers become aware of the importance of model replication and attempted to develop a conversation on how to do it [12,14].

Despite growing efforts toward model replication, not enough has been done [1,13]. Replication is consuming in terms of both time and resources, and sometimes it is not even practical and feasible [6,20]. Even if replication is successful, the results provide nothing new, and the opportunity to publish a paper is much smaller than when building one's own model [12,24]. Other than the incentive issue, there are no standards for the model replication process [23], and there is a "lack of a simulation integrated environment that supports the whole research process from conceptual modeling to simulation implementation and analysis" [19:245]. To facilitate the replication practices, Wilensky and Rand argue for building "... up a body of cases of replication" and extracting "... general principals regarding the replication process" [23:2]. We contribute this paper to the recent

efforts on agent-based model replication, especially how to make ABM through replication a standard and reliable tool for social scientists.

3 An Example: Replicating Vendor Fraud Levels in a Public Service Delivery System

The operation of a public service delivery program can be described as follows: (1) Participants visit a local agency for program enrollment and receive food vouchers, (2) Participants visit stores that have a contract with the program and exchange the vouchers for authorized foods, (3) Stores collect vouchers and submit them to the state agency for payment every month, and (4) State agency pays stores for the vouchers and also monitors overall sales activities in the stores.

In such a public service delivery system, there are at least two systemic reasons for why fraud occurs. First, the program issues vouchers with maximum face values for authorized foods, but participants do not always use 100% of the face value for some reasons (i.e. they may not want to buy all foods listed in the voucher or there can be price fluctuations). At the moment of voucher exchange, the store records how much of the voucher's face value is redeemed by a participant, acquiring the signature of the participant to confirm it. Second, the public agency cannot examine and validate every voucher transaction due to technical, administrative, and economic constraints. With information management systems, voucher transactions are stored for examination, but there are still limitations in processing the large volume of vouchers for analysis and taking corrective actions immediately. The characteristics of the system provide an opportunity for the fraudulent use of vouchers by some players in such a program.

3.1 Agent-Based Modeling

Conceptual Model

The routine business mechanism of the Ohio Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) was used as the conceptual framework for FraudSim [9]. Crime opportunity theory served as a theoretical rationale for modeling fraud as an opportunistic behavior of some players. In the theory, crime depends upon opportunities presented by the routine activities of everyday life, where motivated offenders and suitable targets without guardianship converge in time and space [3,7]. Fraud is defined as the 'misrepresentation of asset values' [21] and is specifically committed by some stores in FraudSim.

Implementation Models

Individual Decision-making: Two different interaction rules were designed to specify how participants choose a store and how decisions related to fraud are made during the voucher exchange process. Participant agents select a store that maximizes their utility based on proximity and store size (store choice rule). Once

they choose a store, participant and store agents negotiate a voucher exchange. The negotiation outcomes depend upon their risk propensities and chance. The higher the risk propensities, the higher the chance of agreeing on the fraudulent exchange of a voucher (fraud negotiation rule). For store agents, the size of the store was also considered. It is more likely for stores of smaller size to engage in fraudulent behavior. When the negotiation is successful for honest or dishonest exchanges, the participant continues to visit the store. Otherwise, the participant leaves for a subsequent choice in the preference list based on the store choice rule. When a participant is not involved in fraud but has a relatively high risk propensity, a random choice is introduced in the simulation process to seek out fraud opportunity. If agents continue to be involved in, or avoid, fraudulent exchanges, their risk propensity is positively or negatively influenced by the agent's decisions, respectively.

Modeling Fraudulent Behavior: Fraud decisions are simultaneously manifested in the behavioral outcomes of store agents. When a voucher is collected from a normal benefit exchange, the store records the amount of the voucher used. For example, in the simulation, the average usage of a voucher is approximately 75% of the face value of the voucher itself. If a participant received a $40 voucher and used 70%, then the actual redemption of the voucher is $28. With a normal exchange, the store records that $28 out of a $40 voucher was used by a participant. When a voucher is collected from a fraudulent benefit exchange, the store agent misrepresents the amount of the voucher, acting as if 100% of the voucher was used. If the store actively engages in this behavior with a greater number of participants, their sales amounts reported to the public agency agent become larger than the true sales amounts. Thus, such cumulated hidden activities can signal store's abnormal business activities to the public agency agent.

Examining Fraud Levels: For monitoring purposes, the public agency agent examines the store's monthly sale activities using three different indicators: a ratio of the actual sales amount to the sum of the vouchers' face values collected from the store (redemption ratio), sales amounts per check-out lane, and food costs per participant at the store. Each indicator separately identifies stores who belong to greater than 90th percentile, and the level of risk of each vendor is decided by summing the result of the indicators. Stores with two or more risk indicators are considered to be high-risk. If only one risk indicator is met, the store is considered to be at-risk.

Kim and Xiao published the agent-based model called "FraudSim" in a book chapter using Java language in conjunction with MASON[10]. The simulation consisted of three agents (10,000 participants, 200 stores, and 1 public agency) that included several stochastic functions. This model was calibrated by matching the level of fraudulent stores in the simulated system with the empirical data available from a county of Ohio in April 2004. The second implementation (replication) model was built by a different modeler using NetLogo (Figure 2).

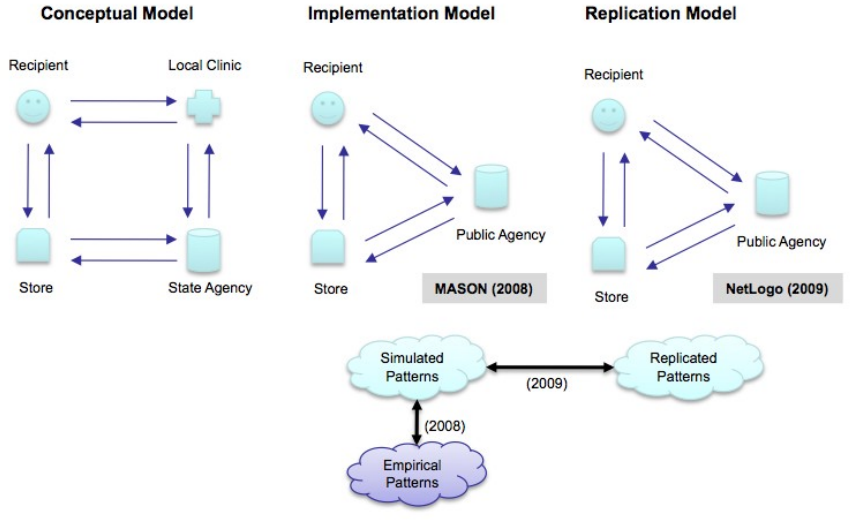


Fig. 2. Conceptual, implementation, and replication models of FraudSim

Table 1. Results from empirical data, original simulation, and replication models

Risk Level	Empirical (2004)	Kim & Xiao (2008)	Zhong (2009)
	N = 188	N = 200	N = 200
0	81.4	75.0	79.0
1	13.3	19.5	13.5
2	4.8	4.5	6.0
3	0.5	1.0	1.5

3.2 Results of the Model Replication

We compare three sets of data for key measurements of the simulated system (distributions of sales activities by three risk indicators and percentage of stores at different risk levels): empirical patterns collected from Ohio WIC in 2004, simulation results published by Kim and Xiao in 2008, and simulation results replicated by Zhong in 2009. Since the replicator had an access only to the simulation results published, the outputs of the replicated model were reported and compared in the same formats. For an illustration purpose, here we reported the distribution of one indicator, stores' monthly redemption ratio in Figure 3.

Using the three risk indicators mentioned, stores were categorized by different levels of risk. The risk level is an additive sum of the indicators met. Table 1 shows that approximately 5−6 percent of stores was identified as being high-risk, meeting two or three indicators at the same time in the three examinations (empirical, original simulation, and replication).

3.3 Evaluating the ABM Replication

Although the replication process shares the same conceptual model with the original simulation process, those two models should differ in some way. Based

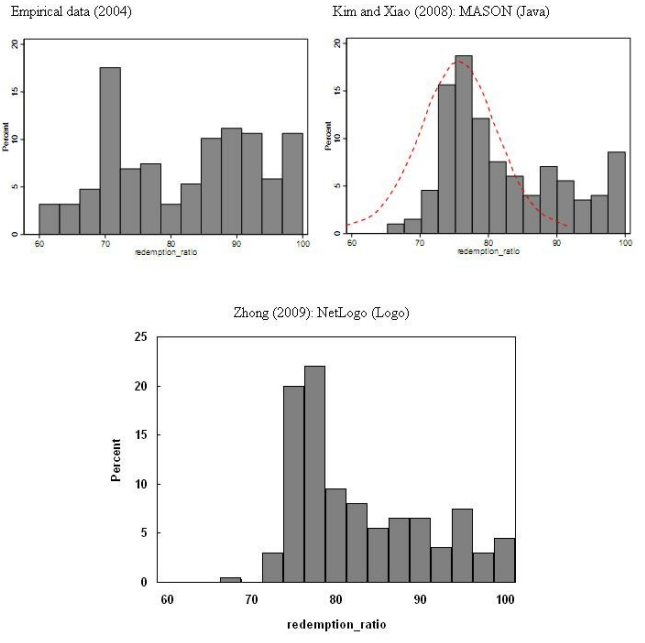


Fig. 3. Vendors' monthly redemption ratio (%)

Table 2. Comparison of implementation models

	Original Model	Replication Model
Time(Published)	2006(2008)	2009
Hardware	Sony Computer	Dell Computer
Language	Java	Netlogo
Toolkits	MASON	Netlogo
Algorithms	Store choice; Fraud negotiation	Store choice; Fraud negotiation
Authors(Published)	Kim (Kim & Xiao)	Zhong

on how likely the replication and original models produce different results, six dimensions were identified to distinguish different implementation efforts: time, hardware, languages, toolkits, algorithms, and authors [23]. Among the six dimensions, the authors dimension is considered to be a strong test to help verify the model. Table 2 compares the implementation models we illustrated with the original implementation model based on the six dimensions.

The simulation outcome was evaluated by comparing distributional equivalence for the standard category. The focal measure was aggregated patterns at a certain point of time. There were occasional rich discussions and personal meetings during the replication process. The replicator had some understanding on Java and MASON before the NetLogo replication model was built, but no experience of building a model using the language or toolkit. If there were

specific questions on programming, these questions were asked to the original modeler. However, the replicator was not exposed to the original code before the replication model was built. At the current replication exercise, the replicator examined results only from the original paper without examining other areas of the parameter space.

4 Discussion and Conclusion

Major distinctions between the original simulation and replication models are time, language and toolkit, and author. First, time can influence the outcome of model replication by making variations in the simulation output and increasing the replicator's understanding of the conceptual model. Technically, time plays a role in creating discrepancies in simulation output relying upon the generation of random seeds. The replication did not get the exactly same numerical outputs, but the general pattern and distribution were equivalent to the ones reported in the previous publication. Time also plays a role in the cognitive understanding of the replicator. Social systems are complex due to the interconnectedness of, and the relationships among, the stakeholders and the problem context. Over time, trial and error enhances the understanding of both the context and the model. Therefore, if a replication model produces different outputs than the original study, the source of the difference can be checked from three different directions: whether it is due to programming errors, random seeds, or lack of understanding in terms of the context that the model attempted to represent.

Both Java and NetLogo are targeted as desirable languages for agent-based modeling, but the difference between them imposes unique challenges. In contrast to Java, for example, some attribute values in NetLogo are not able to be updated immediately after performing a certain function. To resolve this issue, a "link" was created after each negotiation between a participant and a store in NetLogo. The attributes of the link save and store those values for which the values of that store's attributes need to be updated. For the replicator, the challenges of model replication from MASON to NetLogo helped understand which agent should do what, as well as how and when the behavior happens. In Java-based MASON, each agent behaves based on his or her own schedule. While there is a flow of agents' behaviors, these behaviors are separately scheduled for each agent and are not necessarily procedure-oriented as in NetLogo.

Two different authors were involved in two different implementation models. The original simulation model was built by an author who is skilled only in Java-based MASON with personal experience in the public service delivery system. The replication model was built by a different author who simultaneously learned both languages and toolkits (a Java-based MASON and NetLogo with a Logo language) without knowledge or experience of the context. This could lead to a cognitive bias regarding the conceptual model, which may weaken the purpose of model verification and validation. However, the NetLogo replication model was not influenced by the first author due to the different languages used by the authors.

While model reliability is essential for models to be utilized as a supportive tool for social science research, verification and validation of social simulation models are not easy or always feasible. Agent-based models, given the high disaggregation and bottom-to-top logic, hold more assumptions than traditional models [11]. Such characteristics further burden the verification and validation issue. Model replication helps verification and validation of social simulation models in three distinct ways: (1) replication forces replicators to re-examine the conceptual model and assumptions in details, (2) replication engages replicators in the validation process of social systems, and (3) replication helps model builders and replicators share a context for further discussion regarding to the system of interest. This process improves the reliability of agent-based models as both a research approach and a tool. Model replication should be an integral part of the model building process, as well as gain particular attention in the community due to its unique issues and benefits.

References

1. Axelord, R.M.: The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration. Princeton University Press, Princeton (1997)
2. Deffuant, G., Moss, S., Jager, W.: Dialogues concerning a (possibly) new science. Journal of Artificial Societies and Social Simulation 9(1) (2006), http://jasss.soc.surrey.ac.uk/9/1/1.html
3. Eck, J.E.: Examining routine activity theory: A review of two books. Justice Quarterly 12(4), 783–797 (1995)
4. Edmonds, B.: The use of models: Making mabs actually work. In: Moss, S., Davidsson, P. (eds.) MABS 2000. LNCS (LNAI), vol. 1979, pp. 15–32. Springer, Heidelberg (2001)
5. Edmonds, B.: Simulation and complexity: How they can relate. In: Feldmann, V., Mhlfeld, K. (eds.) Virtual Worlds of Precision: Computer-Based Simulations in the Sciences and Social Sciences, pp. 5–32. Lit Verlag, Mnster (2005)
6. Edmonds, B., Hales, D.: Replication, replication and replication: Some hard lessons from model alignment. Journal of Artificial Societies and Social Simulation 6(4) (2003), http://jasss.soc.surrey.ac.uk/6/4/11.html
7. Felson, M.: Crime and Everyday Life: Impact and Implications for Society. Pine Forge Press, Thousand Oaks (1994)
8. Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke, H., Weiner, J., Wiegand, T., DeAngelis, D.L.: Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. Science 310(5750), 987–991 (2005)
9. Kim, Y.: Analysis for Adaptive Complex Public Enterprises. PhD thesis, The Ohio State University, Columbus (2006)
10. Kim, Y., Xiao, N.: Fraudsim: Simulating fraud in a public delivery program. In: Liu, L., Eck, J. (eds.) Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems, pp. 319–338. IGI Global, Hershey (2008)
11. Macal, C.M., North, M.J.: Validation of an agent-based model of deregulated electric power markets. In: North American Association for Computational Social and Organization Sciences Conference, Notre Dame, Indiana, USA, June 22-23 (2006)

12. Macy, M., Sato, Y.: Reply to will and hegselmann. Journal of Artificial Societies and Social Simulation 11(4) (2008), http://jasss.soc.surrey.ac.uk/11/4/11.html
13. Martins, A.C.R.: Replication in the deception and convergence of opinions problem. Journal of Artificial Societies and Social Simulation 11(4) (2008), http://jasss.soc.surrey.ac.uk/11/4/8.html
14. Merlone, U., Sonnessa, M., Terna, P.: Horizontal and vertical multiple implementations in a model of industrial districts. Journal of Artificial Societies and Social Simulation 11(2) (2008), http://jasss.soc.surrey.ac.uk/11/2/5.html
15. Oreskes, N., Shrader-Frechette, K., Belitz, K.: Verification, validation, and confirmation of numerical models in the earth sciences. Science 263(5147), 641–646 (1994)
16. Pace, D.K.: Modeling and simulation verification and validation challenges. Johns Hopkins APL Technical Digest 25(2), 163–172 (2004)
17. Rand, W., Wilensky, U.: Verification and validation through replication: A case study using axelrod and hammond's ethnocetnrism model. In: North American Association for Computational Social and Organization Sciences Conference., Notre Dame, Indiana, USA, June 22-23 (2006)
18. Rouchier, J., Cioffi-Revill, C., Polhill, J.G., Takadama, K.: Progress in model-to-model analysis. Journal of Artificial Societies and Social Simulation 11(2) (2008), http://jasss.soc.surrey.ac.uk/11/2/8.html
19. Sansores, C., Pavon, J.: Agent-based simulation replication: A model driven architecture approach. In: Gelbukh, A., de Albornoz, Á., Terashima-Marín, H. (eds.) MICAI 2005. LNCS (LNAI), vol. 3789, pp. 244–253. Springer, Heidelberg (2005)
20. Smith, M.J., Goodchild, M.F., Longley, P.A.: Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools. Troubador Publishing Ltd., Leicester (2007)
21. Sutherland, E.H.: White-collar criminality. American Sociological Review 5(1), 1–12 (1940)
22. Trucanoa, T., Swilera, L., Igusab, T., Oberkampfc, W., Pilch, M.: Calibration, validation, and sensitivity analysis: What's what. Reliability Engineering and System Safety 91(10-11), 1331–1357 (2006)
23. Wilensky, U., Rand, W.: Making models match: Replicating an agent-based model. Journal of Artificial Societies and Social Simulation 10(4) (2007), http://jasss.soc.surrey.ac.uk/10/4/2.html
24. Will, O., Hegselmann, R.: A replication that failed. Journal of Artificial Societies and Social Simulation 11(3) (2008), http://jasss.soc.surrey.ac.uk/11/3/3.html

Multiscale Comparison of Three-Dimensional Trajectories Based on the Curvature Maxima and Its Application to Medicine

Shoji Hirano and Shusaku Tsumoto

Department of Medical Informatics, Shimane University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
hirano@ieee.org, tsumoto@computer.org

Abstract. This paper presents a novel multiscale comparison method for three-dimensional trajectories. We propose to use the maxima of curvature instead of curvature zero-crossings (inflection points) for splitting a trajectory into subtrajectories (segments), so that a segment-based multiscale matching scheme can be applied for three-dimensional trajectories whose curvature is by nature sign-less. We demonstrate on the synthetic data and medical data that our method could successfully capture the structural similarity of three-dimensional trajectories.

1 Introduction

Advances in sensors and information technologies over the past few decades made it possible to automatically collect huge amount of multivariate time-series data in many domains such as medicine, meteorology, and social sciences. These multivariate time-series data can be viewed as trajectories representing temporal transitions of the status of subjects; therefore, through the cross-subject analysis, i.e. clustering of the trajectories, it may be able to obtain interesting knowledge such as temporal relationships between the variables, common courses of temporal changes, and characteristics of exceptional cases. However, it is still difficult to perform such a large-scale analysis due to the following problems: (1) both short-term and long-term changes may coexist in time series; therefore multiscale observation scheme is required. (2) in order to recognize implicit correlation among variables that may reflect some related phenomena in the subjects, comparison of multidimensional trajectory is needed.

Multiscale comparison of time series or planar curves has been widely studied since 80's mainly in the area of pattern recognition. Based on the Witkin's framework of scale space filtering [1], many methods have been proposed [2]. Ueda et al. [3] enabled the use of discrete scales and comparison of largely distorted curves by introducing a segment-based matching scheme, where a segment corresponds to a subsequence between two adjacent inflection points. Based on these methods, we have developed multiscale comparison and clustering methods for one-dimensional time-series and two-dimensional trajectories of medical data [4,5]. However, comparison of trajectories grater than three-dimension is still a challenging problem because the zero-crossing of curvature (inflection point), that plays an important role in recognizing segment hierarchy, is difficult to be determined for space curves as curvature is always positive.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 128–137, 2010.

In this paper, we propose a multiscale comparison method for three-dimensional trajectories based on the maxima on a curvature scale space. We define a segment as a partial trajectory between two adjacent maxima where curvature becomes locally maximal. Then we trace the place of maxima across the scales in order to obtain the hierarchy of segments. By applying segment-based matching technique, we obtain the best correspondences between partial trajectories. We demonstrate on the synthetic data and real medical data that our method could successfully capture structural similarity of three-dimensional trajectories.

The remainder of this paper is organized as follows. Section 2 briefly introduces the conventional segment-based matching method for two-dimensional trajectories. Section 3 describes the method for comparing three-dimensional trajectories. Section 4 shows the results of matching experiments, and Section 5 is a conclusion of this paper.

2 Multiscale Comparison of Two-Dimensional Trajectories

2.1 Multiscale Representation

Let us denote by $c(t) = \{x(t), y(t)\}$ two-dimensional trajectories composed of two time series $x(t)$ and $y(t)$. Also let us denote by σ an observation scale of the trajectory. Time series $x(t)$ at scale σ, $X(t, \sigma)$, is then derived by the discrete convolution of $x(t)$ and smoothing kernel $I_n(\sigma)$ as follows [6].

$$X(t, \sigma) = \sum_{n=-\infty}^{\infty} e^{-\sigma} I_n(\sigma) x(t - n)$$

where $I_n(\sigma)$ denotes the modified Bessel function of order n, which has better properties for dealing with discrete scales than a sampled Gaussian kernel [6]. By applying this convolution independently to $x(t)$ and $y(t)$, we obtain the trajectory at scale σ as $C(t, \sigma) = \{X(t, \sigma), Y(t, \sigma)\}$. By changing σ, we can represent the trajectory at various observation scales. Figure 1 shows an example of multiscale representation of two-dimensional trajectories. An increase of σ causes an increase of weights for temporally distant points, together with the decrease of weights around the neighbors. Therefore it produces a more smoothed trajectories with less inflection points.

2.2 Hierarchy of Inflection Point

For each trajectory we locate the curvature zero-crossings (inflection points) and represent the trajectory as a set of convex/concave segments. A segment is defined as a partial trajectory between adjacent inflection points. Next, we chase the cross-scale correspondence of inflection points successively from top scale to bottom scale. It defines the hierarchy of segments and guarantees the connectivity of segments across scales. Details of the algorithm for checking segment hierarchy is available in ref. [3]. In order to apply the algorithm to an open trajectory, we modified it to allow the replacement of odd number of segments at start and end, since cyclic property of a set of inflection points can be lost. In Figure 1, segments are represented by $\{a_i^{(k)} \mid i = 1, 2, \cdots, n\}$, where k and n denotes the scale and the number of segments at k, respectively.

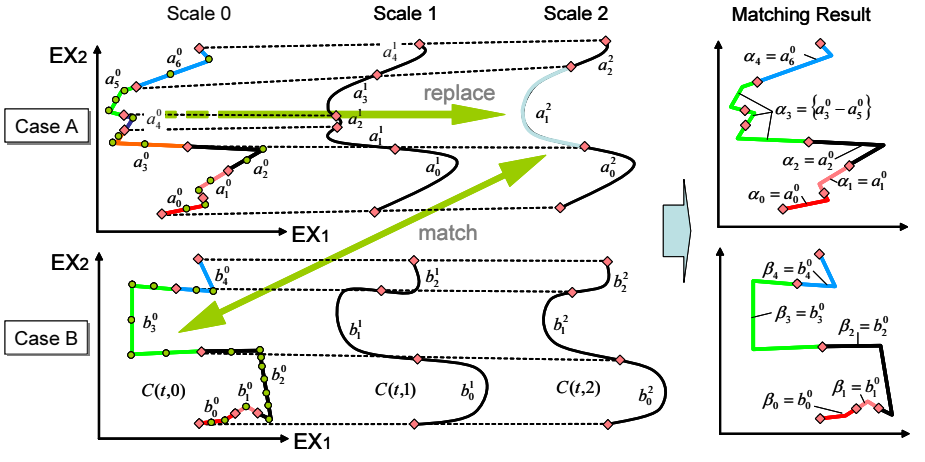


Fig. 1. An illustrative example of multiscale comparison for 2D trajectories

2.3 Matching

The main procedure of multiscale matching is to search the best set of segment pairs that satisfies both of the following conditions: (1) Complete match: By concatenating all segments, the original trajectory must be completely formed without any gaps or overlaps. (2) Minimal difference: The sum of segment dissimilarities over all segment pairs should be minimized.

The search is performed throughout all scales. For example, in Figure 1, three contiguous segments $a_3^{(0)} - a_5^{(0)}$ at the lowest scale of case A can be integrated into one segment $a_1^{(2)}$ at upper scale 2, and the replaced segment well matches to one segment $b_3^{(0)}$ of case B at the lowest scale. Thus the set of the three segments $a_3^{(0)} - a_5^{(0)}$ and one segment $b_3^{(0)}$ will be considered as a candidate for corresponding segments. On the other hand, segments such as $a_6^{(0)}$ and $b_4^{(0)}$ are similar even at the bottom scale without any replacement. Therefore they will be also a candidate for corresponding segments. In this way, if segments exhibit short-term similarity, they are matched at a lower scale, and if they present long-term similarity, they are matched at a higher scale.

2.4 Segment Dissimilarity

The dissimilarity between segments can be defined alternatively. In [5], we used three shape parameters: (1) Gradient at starting point $g(a_m^{(k)})$, (2) Rotation angle $\theta(a_m^{(k)})$, and (3) Velocity $v(a_m^{(k)})$, and defined the local dissimilarity $d(a_m^{(k)}, b_n^{(h)})$ between two segments $a_m^{(k)}$ at scale k and $b_n^{(h)}$ at scale h as

$$d(a_m^{(k)}, b_n^{(h)}) = \sqrt{\left(g(a_m^{(k)}) - g(b_n^{(h)})\right)^2 + \left(\theta(a_m^{(k)}) - \theta(b_n^{(h)})\right)^2} + \left|v(a_m^{(k)}) - v(b_n^{(h)})\right| + \gamma\left\{cost(a_m^{(k)}) + cost(b_n^{(h)})\right\}$$

where $cost()$ denotes a cost function for suppressing excessive replacement of segments [3], and γ is the weight of costs.

3 Multiscale Comparison of Three-Dimensional Trajectories

Curvature zero-crossings is widely used in the applications of scale-space filtering [1,7], because it constitutes a fundamental feature of a planar curve [8] and preserves the monotonicity against the change of scale [9]. However, approaches based on the curvature zero-crossings may not be directly applied to three-dimensional trajectories. For space curves, curvature takes only positive value; therefore it is difficult to determine inflection points. Mokhatarian et al. [12] focused on torsion, which is another major property of the space curve, and proposed multiscale comparison of three-dimensional object shapes using torsion scale space. But it involves a problem that the zero-crossings of torsion may not necessarily satisfy the monotonicity; hence it is difficult to trace the hierarchy of partial trajectories across scales.

In this work, we focus on the maxima of curvature that also satisfies the monotonicity against the change of scale [6], and propose a multiscale comparison method that utilizes *maxima on curvature scale space* for splitting partial trajectories (segments) and recognizing their hierarchy. Its matching procedure is basically similar to the two-dimensional case, but different in following points:

1. A segment is defined as a partial trajectory not between adjacent inflection points but between adjacent maxima.
2. Polarity of a segment (the sign of curvature) is no longer taken into account when matching two segments because every segment has positive sign.
3. Not only odd number of segments, but also even number of segments can be replaced into one segment when scale increases.

In the followings we describe the way of constructing multiscale representation of three-dimensional trajectories, making segments and tracing segment hierarchy based on the maxima of curvature, and defining dissimilarity between segments. By incorporating these procedure with the matching algorithm described in Section 2.3, we finally obtain the best correspondence between segments.

3.1 Multiscale Representation of Three-Dimensional Trajectories

Let us denote by $c(t) = \{x(t), y(t), z(t)\}$ a three-dimensional trajectory constituted of three time series $x(t)$, $y(t)$ and $z(t)$. Similarly to the two-dimensional case, the trajectory $C(t, \sigma)$ at scale σ is derived by the discrete convolution of each time series and the modified Bessel smoothing kernel $I_n(\sigma)$ as $C(t, \sigma) = \{X(t, \sigma), Y(t, \sigma), Z(t, \sigma)\}$.

3.2 Derivation of Curvature Maxima

Next, for each trajectory we compute the curvature of each point and locate their local maxima. Curvature $\kappa(t, \sigma)$ of $C(t, \sigma)$ is defined by

$$\kappa(t, \sigma) = \frac{\sqrt{(Z''Y' - Y''Z')^2 + (X''Z' - Z''X')^2 + (Y''X' - X''Y')^2}}{(X'^2 + Y'^2 + Z'^2)^{3/2}}$$

where X' and X'' respectively denote the first- and second-order derivatives of $X(t, \sigma)$ about t. Similar treatment applies to $Y(t, \sigma)$ and $Z(t, \sigma)$.

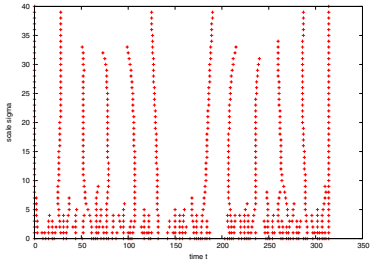


Fig. 2. Maxima scale space for $T2n_1$

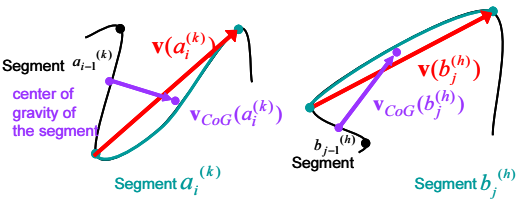


Fig. 3. Vector representation of segments

3.3 Construction of the Maxima Scale Space and Trace of Segment Hierarchy

Curvature maxima are then plotted on a two-dimensional plane of time and scale. We call this plane *curvature maxima scale space*. Figure 2 shows an example of the maxima scale space for three-dimensional trajectory $T2n_1$ used in our experiments. The horizontal axis denotes time t, the vertical axis denotes scale σ, and '+' denotes a point of maximal curvature. For each scale, a segment is defined as a set of data points separated by two adjacent curvature maxima. The number of maxima will decrease when scale increases because of the smoothing, and segments at a fine scale will be merged into one segment at a coarse scale. Maxima are successively linked from the top scale to the bottom scale based on the minimal distance criterion for forming segment hierarchy.

3.4 Matching

After constructing segments and recognizing their hierarchy across scales, segment matching is performed in the same way with the two-dimensional case. Since structural feature points are changed from zero-crossings (inflection points) to curvature maxima, the shape of a segment also changes from convex/concave shape to 's' shape. We model this shape as a straight vector connecting both ends of the segment and define the dissimilarity between segments as follows. First, let us denote by $a_m^{(k)}$ and $b_n^{(h)}$ two segments to be compared as shown in Figure 3. Secondly, let us denote by $\mathbf{v}(a_m^{(k)})$ a three-dimensional vector connecting both ends of $a_m^{(k)}$, and similarly denote $\mathbf{v}(b_n^{(h)})$ for $b_n^{(h)}$. Thirdly, let us denote by $\mathbf{v}_{CoG}(a_m^{(k)})$ a vector connecting the center of gravities of segments $a_m^{(k)}$ and $a_{m-1}^{(k)}$. Then we define the dissimilarity between $a_m^{(k)}$ and $b_n^{(h)}$ by

$$
\begin{aligned}
d(a_m^{(k)}, b_n^{(h)}) &= \mathrm{vdiff}(a_m^{(k)}, b_n^{(h)}) \cdot \mathrm{cost}(a_m^{(k)}, b_n^{(h)}) \\
\mathrm{vdiff}(a_m^{(k)}, b_n^{(h)}) &= \|\mathbf{v}(a_m^{(k)}) - \mathbf{v}(b_n^{(h)})\| + \|\mathbf{v}_{CoG}(a_m^{(k)}) - \mathbf{v}_{CoG}(b_n^{(h)})\| \\
\mathrm{cost}(a_m^{(k)}, b_n^{(h)}) &= \left(\frac{n_A^{(0)}}{n_A^{(k)}} \cdot \frac{n_B^{(0)}}{n_B^{(h)}} \right)^{\lambda}
\end{aligned}
$$

where $n_A^{(0)}$ and $n_B^{(0)}$ denote the number of segments at scale 0 on trajectory A and B respectively. The first term in vdiff quantifies the difference of segment shapes (directions) and the second term quantifies the difference of location changes of segments.

The replacement cost becomes large proportional to the ratio of the number of segments at scale 0 to that of the current scale k. The term λ represents weight for cost.

4 Experimental Results

4.1 Synthetic Data

We have firstly conducted a preliminary matching experiment using synthetic data for checking the basic functionality of the proposed method. We generated two simple three-dimensional trajectories T1 and T2 by using triangular functions as follows.

$T1$	$T2$
$x = \sin(t) + \frac{1}{3}\sin(3t) + \frac{1}{5}\sin(5t)$	$x = \sin(t) + \frac{1}{3}\sin(3t) + \frac{1}{5}\sin(5t)$
$y = \sin(\frac{t}{2})$	$y = \sin(\frac{t}{2}) + \frac{1}{3}\sin(\frac{3}{2}t)$
$z = \cos(\frac{4}{5}t)$	$z = \cos(t)$

The parameters used for matching were: starting scale = 1.0, max stage = 40.0, minimal scale interval = 1.0, weight for replacement cost = 1.0.

Figure 4 shows the shapes of trajectories T1 and T2 and their multiscale representation. They were divided into segments by the curvature maxima denoted by '+'. Figure 5 shows the matching result. Matched segments are represented in the same color. The original T1 and T2 were different around $z = 0$ in terms of their y-direction changes, but they were successfully matched according to the structural similarity of the entire shapes. We also conducted a matching experiment for noisy trajectories. We generated two noisy trajectories denoted by T2n1 and T2n2 by adding Gaussian noise to T2. Figure 6 shows multiscale representations of their shapes and Figure 7 shows their matching results. We could confirm that their structural similarity was successfully captured in a global scale if there existed local differences at fine scales caused by noise.

4.2 Medical Data

Next, we applied our method to the chronic hepatitis dataset which was a common dataset in ECML/PKDD discovery challenge 2002-2004 [11]. The dataset contained time series laboratory examinations data collected from 771 patients of chronic hepatitis B and C. In this study, we focused on analyzing the temporal relationships between platelet count (PLT), albumin (ALB) and cholinesterase (CHE), that were often used to examine the status of liver function. Our goals were set to: (1) find groups of trajectories that exhibit interesting patterns, and (2) analyze the relationships between these patterns and the stage of liver fibrosis. We here chose cases of type C hepatitis without interferon (IFN) treatment as the subjects of analysis because they could be considered as the natural courses of type C hepatitis. We excluded the cases that did not contain valid examination results for all of PLT, ALB, CHE and liver biopsy, and consequently, we obtained a total of 99 cases. Experiments were conducted as follows.

1. Select a pair of cases (patients) and calculate the dissimilarity by using the proposed method. Apply this procedure for all pairs of cases, and construct a dissimilarity matrix.

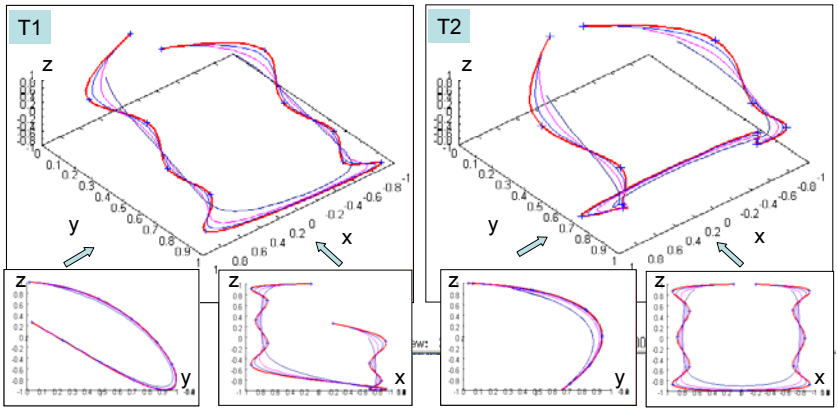


Fig. 4. Multiscale representations of test trajectories T1 (left) and T2 (right). The red curves represent the original shapes. Maxima points are denoted by '+'.

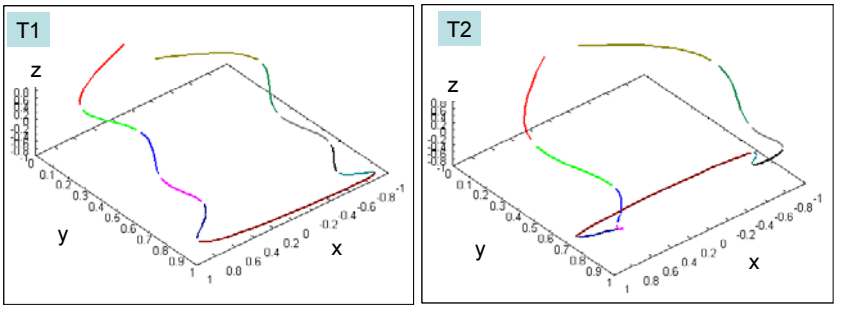


Fig. 5. Matching results. Matched segments are represented in the same color.

2. Create a dendrogram by using conventional hierarchical clustering [10] and the dissimilarity matrix. Then perform cluster analysis.

Parameters for multiscale matching were empirically determined as follows: starting scale = 1.0, max scale = 800, minimal scale interval = 0.2, weight for segment replacement cost = 1.0. We used group average as a linkage criterion for hierarchical clustering.

Figure 8 shows the dendrogram. By manually inspecting cluster constitutions, we selected 9 cluster solution shown as the horisontal line on the dendrogram. Table 1 shows constitution of clusters stratified by fibrotic stage. F0 represents no fibrosis, and F4 represents severe fibrosis. We could observe interesting feature in their distribution. For example, cluster 4 contained a large number of progressed cases whereas clusters such as no.9 contained mostly un-progressed cases. These results implied that the shapes of trajectories might have some relationships to the fibrotic stages.

Figure 9 shows an example of matching between MID 170 (F4) and 602 (F4) grouped into the same cluster (no.4). In detail these trajectories were largely different, however, as one can recognize visibly, they were partly and globally similar. As shown in the

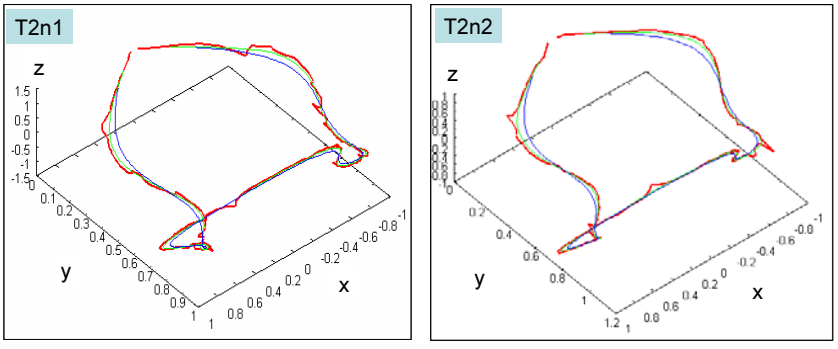


Fig. 6. Multiscale representations of noisy test trajectories T2n1 (left) and T2n2 (right). The red curve represents their original shapes.

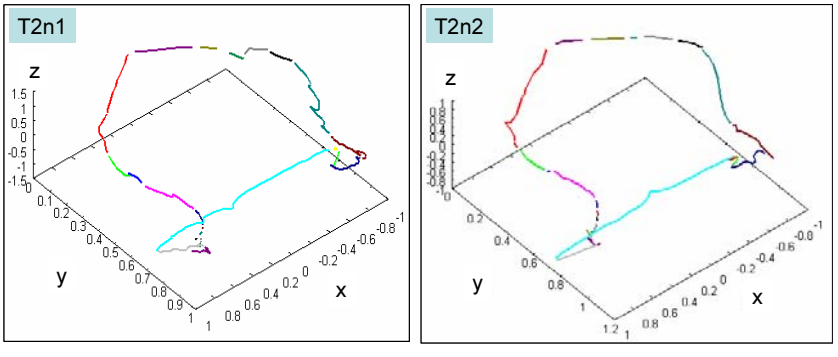


Fig. 7. Matching results. Matched segments are represented in the same color.

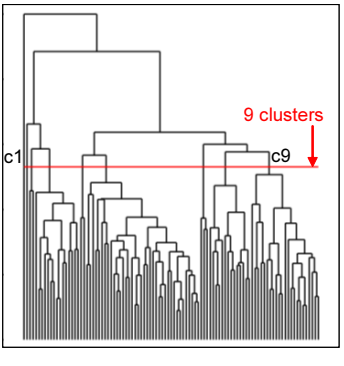


Fig. 8. Dendrogram for ALB-CHE-PLT trajectories in Type C without IFN dataset

Table 1. Cluster constitutions of ALB-CHE-PLT trajectories, stratified by fibrotic stages

Cluster	# of Cases / Fibrotic stage				Total
	F0,F1	F2	F3	F4	
1	1	0	0	0	1
2	0	1	0	0	1
3	0	0	1	1	2
4	1	2	2	10	15
5	0	0	1	1	2
6	22	5	7	4	38
7	2	1	0	0	3
8	13	1	0	0	14
9	19	2	2	0	23
total	58	12	13	16	99

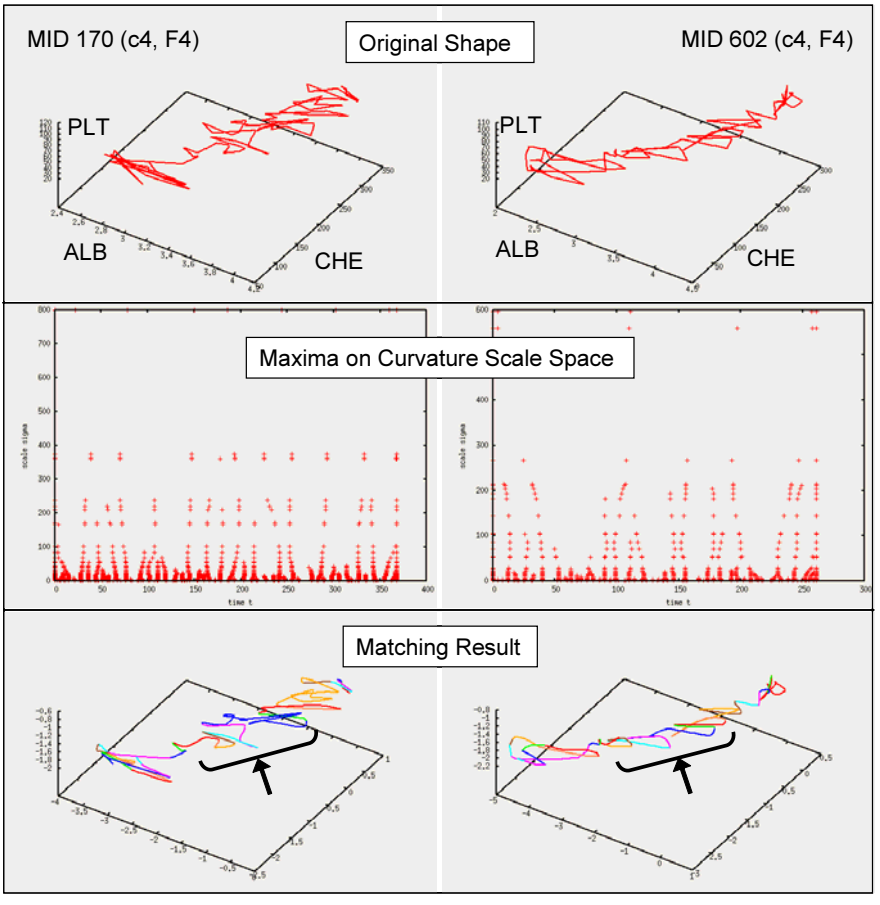


Fig. 9. Matching result of two trajectories. Left: MID 170 (F4), Right MID 602 (F4). Top: original shapes. Middle: maxima on curvature scale space. a '+' represents curvature maximum. Bottom: matching results. Matched segments are represented with same color.

arrowed region on the matching results, these similarity were successfully captured by our method.

We implemented the proposed method on a Linux server (Intel Xeon E5430 2.6GHz x2, 4GB memory). It took about 15 minutes to perform comparison on 99 cases.

5 Conclusions

In this paper we have presented a multiscale comparison method for three-dimensional trajectories. In order to deal with the problem that zero-crossings of curvature cannot be determined for space curve, we focused on the maxima of curvature. The hierarchy of partial trajectories was recognized by tracing the positions of maxima across scales. Then we performed segment-by-segment matching across the scales, and obtained the

best correspondence of segments. In the experiments we demonstrated that reasonable correspondences were obtained on the simple but noisy trajectories. We also demonstrated using real medical data that the method could generate interesting clusters that might reflect distribution of fibrotic stages.

This work is still at an early stage and there are lots of work to be done. We will continue to tackle the following issues: (1) investigation of the characteristics of maxima on the curvature scale space, (2) refinement of the segment dissimilarity, (3) quantitative evaluation of the performance.

Acknowledgment

This work is supported in part by the Grant-in-Aid for Young Scientists (B) (#20700140) by MEXT, Japan.

References

1. Witkin, A.P.: Scale-space filtering. In: Proc. of the 8th International Joint Conference on Artificial Intelligence, pp. 1019–1022 (1983)
2. Mokhtarian, F., Bober, M.: Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization. Springer, Heidelberg (2003)
3. Ueda, N., Suzuki, S.: A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Transactions on Information and Systems J73-D-II(7), 992–1000 (1990)
4. Hirano, S., Tsumoto, S.: Clustering Time-series Medical Databases based on the Improved Multiscale Matching. In: Hacid, M.-S., Murray, N.V., Raś, Z.W., Tsumoto, S. (eds.) ISMIS 2005. LNCS (LNAI), vol. 3488, pp. 612–621. Springer, Heidelberg (2005)
5. Hirano, S., Tsumoto, S.: Cluster analysis of trajectory data on hospital laboratory examinations. In: Proc. of AMIA Annual Symp., vol. 11, pp. 324–328 (2007)
6. Lindeberg, T.: Scale-Space for Discrete Signals. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(3), 234–254 (1990)
7. Mokhtarian, F., Mackworth, A.K.: Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI 8(1), 24–43 (1986)
8. Dudek, G., Tostsos, J.K.: Shape Representation and Recognition from Multiscale Curvature. Comp. Vis. Img Understanding 68(2), 170–189 (1997)
9. Babaud, J., Witkin, A.P., Baudin, M., Duda, O.: Uniqueness of the Gaussian kernel for scale-space filtering. IEEE Transactions on Pattern Analysis and Machine Intelligence 8(1), 26–33 (1986)
10. Everitt, B.S., Landau, S., Leese, M.: Cluster Analysis, 4th edn. Arnold Publishers (2001)
11. http://lisp.vse.cz/challenge/
12. Mokhtarian, F.: Multi-scale description of space curves and three-dimensional objects. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 298–303 (1988)

A Knowledge Collaboration Network Model across Disciplines

Anna Nagurney[1] and Qiang Qiang[2]

[1] Isenberg School of Management
University of Massachusetts
Amherst, Massachusetts 01003
nagurney@gbfin.umass.edu
[2] Management Division
Pennsylvania State University
Great Valley School of Graduate Professional Studies
Malvern, Pennsylvania 19355

Abstract. We propose a theoretical framework for the optimal collaboration among researchers in a knowledge network in which researchers are not limited to a single discipline and in which multiple modes of communication, including communication via the Internet, are available. We introduce a novel concept of distance to measure not only the communication distance but also the distance between disciplines. We formulate the knowledge network collaboration model as a variational inequality problem whose solution yields the optimal allocation of effort/time of the researchers as well as the associated opportunity costs.

Keywords: Knowledge networks, Scientific collaboration.

1 Introduction

Knowledge and the production of knowledge are driving forces in the modern economy. The study of knowledge production and the formation and evolution of knowledge networks, have been addressed by economists, sociologists, as well as management and organizational theorists (see, e.g., [2], [3]). Beckmann significantly advanced the modeling, analysis, and understanding of researchers' behavior in scientific collaboration including collaborations in which the system had an optimally efficient time allocation from an economics perspective. According to his models, which assumed face-to-face communication, the likelihood of two researchers collaborating decreases as the physical distance between them increases. However, by using today's highly developed communication technologies, researchers can now exchange ideas with one another with virtually no time or monetary costs. Moreover, given the availability of current advanced technologies, notably, the Internet, researchers have significantly more options when it comes to selecting modes of collaboration. In today's society, two collaborators may not even have to meet face-to-face. In addition, the effects of advances in research productivity due to advances in technology have also been profound.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 138–148, 2010.

Gibbons et al. [5] and Hudson [6] emphasized that the emergence of critical technologies greatly impacts researchers' output. Rosenblat and Mobius [10] found that the number of co-authored papers in economics has increased thirty percent since the rise of the Internet.

Another assumption in the original knowledge network collaboration model that merits relaxation is that researchers collaborate exclusively within a field or discipline. Given today's complex world, we believe researchers will have to collaborate interdisciplinarily to solve problems more effectively. Börner et al. [4] indicated that there are many bibliometric studies that address the trends of co-authorship and interdisciplinary research in the literature of collaboration networks. These studies, however, are mainly descriptive in nature in that they focus on longitudinal publication data of certain journals. Although almost all of these studies reach similar conclusions regarding the rising trend of interdisciplinary co-authorship among more spatially separated co-authors, there is no theoretical model to support the empirical findings, which could also illuminate the economic decision-making behavior of the researchers.

2 The Knowledge Collaboration Network Model

Assume that there are N researchers, with a typical researcher denoted by i, j, etc. A researcher is an individual and may represent a particular discipline based on his education, research experiences, etc. Assume that pairs of researchers can collaborate with one another via O modes of communication (such as face-to-face, email, telephone, fax, etc.) with a typical mode of communication denoted by k. The following fundamental assumptions of the model are adopted from [2]:

1. researchers only collaborate in pairs, and
2. a pair of researchers has to mutually agree to collaborate with one another.

However, unlike Beckmann's assumption that the physical distance between two researchers determines the collaboration time, we assume that there are two factors affecting the collaboration time, namely, *virtual distance* and *communication distance* (both are measured in a time scale). According to the most commonly used definition, *interdisciplinarity* is used to refer to increasing levels of interaction among disciplines. The virtual distance is then used as a proxy to represent the time spent by two researchers on understanding each other's discipline. Bibliometric studies of interdisciplinarity can help in identifying a good measure of virtual distance. [8] classified journal publications by using subject categories in Science Citation Index, Social Sciences Citation Index, and the Arts & Humanities Citation Index from the Institute for Scientific Information. The bibliometric data was then employed to establish links between these categories. The strength of the links indicates how close disciplines are to one another. The study results were shown to agree with data in journal publications. We also use the strength of links as a proxy for the measure of virtual distance. In Table 1, the model primary notation appears.

The knowledge collaboration network problem is now formulated as a system-optimization problem. Nodes correspond to researchers and links to different

Table 1. Parameters in the Knowledge Collaboration Network Model

Notation	Definition
a_{ijk}	coefficient denoting how often researchers i and j communicate via k
r_{1ij}	virtual distance between researchers i and j
r_{2ijk}	communication distance between i and j communicating via k
r_{ijk}	$=r_{1ij}+r_{2ijk}$ – total distance between researchers i and j via k
t_{ijk}	$=1+a_{ijk}r_{ijk}$ – actual time spent in mode k to achieve one time unit of effective collaboration between researchers i and j
T_i	time budget for researcher i
λ_i	opportunity cost for researcher i
λ	N-dimensional vector formed by grouping the λ_i over the i
x_{ijk}	effort of researcher i communicating with j via k in time units
x_{jik}	effort of researcher j communicating with i via k in time units
x	NNO-dimensional vector of the efforts x_{ijk} grouped over i, j, and k
$u_{ijk}(x_{ijk}, x_{jik})$	utility of researcher i collaborating with j via k
$u_{jik}(x_{jik}, x_{ijk})$	utility of researcher j collaborating with i via k

modes of communication. In particular, given a knowledge network system in which there are N researchers, together with their time budget constraints, we wish to determine a time (effort) allocation plan that maximizes the total utility of the knowledge collaboration network as represented by the sum of the utilities of the individual researchers involved in the collaboration network. For example, such an optimization problem may be faced by the manager of R&D in a knowledge organization or company, a director of a research organization, an academic dean of a school of science in a university, or a principal investigator of a major research project. The optimization problem can, hence, be expressed as:

$$\text{Maximize} \quad \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{O} u_{ijk}(x_{ijk}, x_{jik}) \tag{1}$$

subject to:

$$\sum_{j=1}^{N}\sum_{k=1}^{O} t_{ijk}x_{ijk} \leq T_i, \qquad i = 1, \ldots, N, \tag{2}$$

$$x \in R_{+}^{N\times N\times O}. \tag{3}$$

We assume that the utility functions are continuously differentiable, concave, and strictly monotonically increasing. Since the time budget constraints are linear, according to the Karush-Kuhn-Tucker conditions (cf. [1]), an optimal solution x^* to (1), subject to (2) and (3), is guaranteed, with the optimality conditions given as follows: for all i, j, k:

$$\frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} + \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{ijk}} \begin{cases} = t_{ijk}\lambda^*_i, & \text{if} \quad x^*_{ijk} > 0, \\ \leq t_{ijk}\lambda^*_i, & \text{if} \quad x^*_{ijk} = 0, \end{cases} \tag{4}$$

$$T_i - \sum_{j=1}^{N}\sum_{k=1}^{O} t_{ijk}x^*_{ijk} \begin{cases} = 0, & \text{if } \lambda^*_i > 0, \\ \geq 0, & \text{if } \lambda^*_i = 0. \end{cases} \tag{5}$$

According to the definition of t_{ijk} and λ_i, t_{ijk} is the amount of actual time spent to achieve one unit of effective collaboration time while λ_i is the shadow price for each time unit. Hence, $t_{ijk}\lambda_i$ can be interpreted as the cost of time that researcher i is willing to spend on collaborating with researcher j in mode k. The optimality condition (4) can be interpreted as: researcher i will collaborate with researcher j via mode k given that the total marginal utility of the pair of collaborators i and j with respect to i's marginal contribution is equal to the cost of time that i is willing to spend on such collaboration. Researcher i will not collaborate with researcher j via k if the total marginal utility of the pair from i's marginal contribution cannot "cover" the cost of time that i is willing to spend on such collaboration. Since the model is a system-optimization model, the above interpretation is quite intuitive. The condition (5) states that a researcher has positive opportunity cost only if he uses up his time resources.

Theorem 1: Optimal Opportunity Costs
If in the knowledge collaboration network model presented in (1) – (3), the utility functions u_{ijk}, $\forall i, j, k$, are continuously differentiable and strictly monotonically increasing, then the optimal opportunity costs λ^*_i, $\forall i$, are positive.

Proof: Since u_{ijk} is assumed to be continuously differentiable and strictly monotonically increasing, its first-order derivative is positive. Hence, using also (4):

$$t_{ijk}\lambda^*_i \geq \frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} + \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{ijk}} > 0. \tag{6}$$

Since $t_{ijk} > 0$ the conclusion follows. □

The variational inequality formulation (cf. [9]) of the optimality conditions (4) and (5) is given in the following theorem.

Theorem 2: Variational Inequality Formulation
A solution to the knowledge network collaboration model is an optimal solution if and only if it satisfies the variational inequality problem: determine $(\lambda^*, x^*) \in \mathcal{K}$, where $\mathcal{K} \equiv \{(\lambda, x) \mid (\lambda, x) \in R_+^{N+N\times N\times O}\}$, such that

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{O}(t_{ijk}\lambda^*_i - \frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} - \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{ijk}}) \times (x_{ijk} - x^*_{ijk})$$

$$+ \sum_{i=1}^{N}(T_i - \sum_{j=1}^{N}\sum_{k=1}^{O} t_{ijk}x^*_{ijk}) \times (\lambda_i - \lambda^*_i) \geq 0, \quad \forall (\lambda, x) \in \mathcal{K}. \tag{7}$$

Proof: See [9].

For easy reference in the subsequent sections, variational inequality problem (7) can be rewritten in standard form (cf. [9]): determine $X^* \in \mathcal{K}$ satisfying:

$$\langle F(X^*)^T, X - X^* \rangle \geq 0, \quad \forall X \in \mathcal{K} \equiv R_+^{N+N\times N\times O}, \tag{8}$$

where $X \equiv (\lambda, x)$, $F(X) \equiv (F_{ijk}, F_i)_{i=1,\ldots,N;j=1,\ldots,N;k=1,\ldots,O}$, with the specific components of F given by the functional terms preceding the multiplication signs in (7), respectively. Here $\langle \cdot, \cdot \rangle$ denotes the inner product in M–dimensional Euclidean space where $M = N + NNO$.

We now impose the following assumptions on the above model: every researcher has a utility function characterized by the property that a pair of collaborators shares the utility (which may also be interpreted as the research credit associated with co-authorship) evenly, that is, $u_{ijk}(x_{ijk}, x_{jik})$ is equal to $u_{jik}(x_{jik}, x_{ijk})$. This assumption is also a fundamental assumption in [2] but in the case of a single mode of collaboration (and single discipline).

With such a "symmetric" utility function for each pair of collaborators, $u_{ijk}(x_{ijk}, x_{jik})$ and $u_{jik}(x_{jik}, x_{ijk})$ are equal everywhere in the feasible set. Therefore, we have that $\frac{\partial u_{ijk}(x_{ijk}, x_{jik})}{\partial x_{ijk}} = \frac{\partial u_{jik}(x_{jik}, x_{ijk})}{\partial x_{ijk}}$.

The following expressions can, hence, be obtained for the symmetric utility function case from (4):

$$\text{if } x^*_{ijk} > 0, \text{ then } \frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} = \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{ijk}} = \frac{1}{2} t_{ijk} \lambda^*_i; \tag{9}$$

$$\text{if } x^*_{jik} > 0, \text{ then } \frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{jik}} = \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{jik}} = \frac{1}{2} t_{jik} \lambda^*_j. \tag{10}$$

Consequently, the optimality condition (4) can now be interpreted as follows: in the case of symmetric utility functions, researcher i; respectively, j, will collaborate with researcher j; respectively, i, via k only if i's marginal utility is equal to half of the total time cost that he is willing to pay. He will not collaborate if his marginal utility cannot "cover" the time cost he is willing to pay.

Further analysis of the optimality conditions (9) and (10) is also very interesting and worthy of interpretation. First, we discuss the relevant results for a particular collaboration mode. Let's assume that there is a pair of researchers i and j in the knowledge network. We know from Theorem 1 that the opportunity cost in the optimal solution is positive under the assumptions on the utility functions specified before. Therefore, the following results for researcher i can be obtained:

$$\text{if } x^*_{iik} > 0, \text{ then } \frac{\partial u_{iik}(x^*_{iik}, x^*_{iik})}{\partial x_{iik}} = \frac{1}{2} \lambda^*_i; \tag{11}$$

$$\text{if } x^*_{ijk} > 0, \text{ then } \frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} = \frac{1}{2} t_{ijk} \lambda^*_i. \tag{12}$$

The analogous results hold for researcher j.

We now discuss three distinct cases for researcher i collaborating with researcher j: **Case i).** i works independently; **Case ii).** i works with j who shares the same discipline with i; **Case iii).** i works with j who is not in the same discipline as i.

We denote the t_{ijk}s in the above cases as: t^1_{iik}, t^2_{ijk}, and t^3_{ijk}, respectively. According to the definition of t_{ijk}, we have that $t^3_{ijk} > t^2_{ijk} > t^1_{iik} = 1$. We also let $u^1_{iik}(x_{iik}, x_{iik})$, $u^2_{ijk}(x_{ijk}, x_{jik})$, and $u^3_{ijk}(x_{ijk}, x_{jik})$ denote researcher i's utility functions corresponding, respectively, to the above three cases.

From (11), (12), and that $t^3_{ijk} > t^2_{ijk} > t^1_{iik} = 1$, we conclude:

$$\frac{\partial u^3_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} > \frac{\partial u^2_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} > \frac{\partial u^1_{iik}(x^*_{iik}, x^*_{iik})}{\partial x_{iik}}. \tag{13}$$

The same relationships as in (13) also hold for researcher j.

From the above discussion, we can see that interdisciplinary collaboration will only occur if the marginal utility of such a collaboration is higher than that of the intradisciplinary collaboration while intradisciplinary collaboration will occur only if the associated marginal utility is higher than that of a researcher working independently. Hence, as we conjectured earlier in this paper, researchers will not collaborate with one another across disciplines unless such collaboration brings them higher "benefit." This phenomenon has been witnessed in many empirical studies.

Moreover, for a pair of researchers i and j, if there exists a collaboration between them via communication mode k, that is, $x^*_{ijk} > 0$ and $x^*_{jik} > 0$, from (12) and the corresponding condition for researcher j, we have that: according to the definitions of t_{ijk} and t_{jik}: $\frac{1}{2}t_{ijk} = \frac{1}{2}t_{jik} = \frac{\frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}}}{\lambda^*_i} = \frac{\frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{jik}}}{\lambda^*_j}$, and, therefore,

$$\frac{\frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}}}{\frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{jik}}} = \frac{\lambda^*_i}{\lambda^*_j}. \tag{14}$$

From (14), we can see that in order to achieve optimality, a researcher with higher opportunity cost must have higher marginal utility and his collaborator who has lower opportunity cost must have a lower marginal utility. Furthermore, given the definition of productivity (the amount of output based on the unit input), if two researchers share the credit evenly, the researcher with lower productivity has to contribute more time to the collaboration in order to compensate for his counterpart's effort. This finding coincides with Beckmann's [2] conclusion.

Having discussed comunication/collaboration via a particular mode, we now analyze the collaboration of a certain pair of researchers across different communication modes. Let's assume that in the optimal solution, researcher i has a collaboration with researcher j and that there are two communication modes available to them, namely, f and h. Let's further assume that the communication distance r_{2ijf} is larger than r_{2ijh}. According to the definition of t_{ijk}, we know

that $t_{ijf} > t_{ijh}$. We now discuss the following three possibilities regarding x^*_{ijf} and x^*_{ijh}:

Case iv). $x^*_{ijf} > 0$ and $x^*_{ijh} > 0$. According to (12), we have that

$$\frac{\frac{\partial u_{ijf}(x^*_{ijf}, x^*_{jif})}{\partial x_{ijf}}}{t_{ijf}} = \frac{\frac{\partial u_{ijh}(x^*_{ijh}, x^*_{jih})}{\partial x_{ijh}}}{t_{ijh}} = \frac{1}{2}\lambda^*_i. \tag{15}$$

By a small perturbation of (15), we obtain

$$\frac{\frac{\partial u_{ijf}(x^*_{ijf}, x^*_{jif})}{\partial x_{ijf}}}{\frac{\partial u_{ijh}(x^*_{ijh}, x^*_{jih})}{\partial x_{ijh}}} = \frac{t_{ijf}}{t_{ijh}}. \tag{16}$$

We have a "constant elasticity of substitution" between the two communication modes, that is, mode f which consumes more communication time yields a higher marginal utility while mode h which consumes less communication time yields a lower marginal utility. However, a researcher does not care which mode to use in order to collaborate in this case.

Case v). $x^*_{ijf} > 0$ and $x^*_{ijh} = 0$. Via a similar derivation to that constructed for Case iv, we obtain

$$\frac{\frac{\partial u_{ijf}(x^*_{ijf}, x^*_{jif})}{\partial x_{ijf}}}{\frac{\partial u_{ijh}(x^*_{ijh}, x^*_{jih})}{\partial x_{ijh}}} \geq \frac{t_{ijf}}{t_{ijh}}. \tag{17}$$

In this scenario, a pair of researchers selects mode f for collaboration. Although mode f consumes more communication time, it yields a larger marginal utility that compensates for the additional time. This result is interesting and intuitive. For instance, some collaboration work cannot be completed without collaborators having a meeting in person although other time-efficient communication modes are available. Such cases occur often in physics, for example, when an important experiment has to be conducted and analyzed by both of the collaborators. Although traveling is time-consuming, it may be necessary.

Case vi). $x^*_{ijf} = 0$ and $x^*_{ijh} > 0$. Similarly, we have that, in this case

$$\frac{\frac{\partial u_{ijf}(x^*_{ijf}, x^*_{jif})}{\partial x_{ijf}}}{\frac{\partial u_{ijh}(x^*_{ijh}, x^*_{jih})}{\partial x_{ijh}}} \leq \frac{t_{ijf}}{t_{ijh}}. \tag{18}$$

Here, the marginal utility associated with mode f cannot compensate for the additional time. Hence, mode f will not be used as a method of communication.

From the above three cases, it is clear that for a communication mode that consumes more communication time to be selected, it must yield a sufficiently large marginal utility to compensate for its additional communication time. Through the inclusion of communication and virtual distance, the model provides insightful interpretations into the manner in which researchers make their decisions in terms of optimal collaboration.

3 Qualitative Properties

Note that the collaboration times, x_{ijk} and x_{jik}, are bounded by each researcher's time budget. The opportunity costs, however, do not lie in a compact set. Hence, according to the standard theory of variational inequalities (cf. [9]), one can impose either a coercivity condition on the vector function $F(X)$ (see (8)) or a boundedness condition in order to guarantee the existence of the opportunity costs. Here, we apply a more direct approach. We note that, from the standard theory of variational inequality theory, optimization problem (1), subject to (2) and (3), under the assumption that the utility functions are concave and continuously differentiable, can also be formulated as the variational inequality problem: determine $x^* \in \mathcal{K}^2$, where $\mathcal{K}^2 \equiv \{x | x \in R_+^{N \times N \times O}$ and satisfies (2), (3)$\}$ such that

$$-\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{O}\left(\frac{\partial u_{ijk}(x^*_{ijk}, x^*_{jik})}{\partial x_{ijk}} + \frac{\partial u_{jik}(x^*_{jik}, x^*_{ijk})}{\partial x_{ijk}}\right) \times (x_{ijk} - x^*_{ijk}) \geq 0, \quad \forall x \in \mathcal{K}^2. \tag{19}$$

Theorem 3: Existence
Existence of a solution $x^* \in \mathcal{K}^2$ to variational inequality (19) is guaranteed.

Proof: Follows from the standard theory of variational inequalities since the marginal utility functions are continuous and the feasible set $\mathcal{K}^2$ is compact.

Theorem 4: Uniqueness
Assume that the utility functions u_{ijk}, $\forall i, j, k$, are strictly concave functions. Then the optimal effort (time) allocation pattern satisfying variational inequality (19) is unique.

Proof: Under the assumption of strictly concave utility functions, the vector function F^2 with components: F^2_{ijk} given by $-\frac{\partial u_{ijk}(\cdot)}{\partial x_{ijk}} - \frac{\partial u_{jik}(\cdot)}{\partial x_{ijk}}$ is strictly monotone and the conclusion follows (cf. [9]). □

Clearly, under the same assumptions as in Theorem 4, variational inequality (19) also admits a unique optimal effort (time) allocation pattern, from which one can then by using (4) and (5) recover the unique opportunity cost pattern.

4 Computational Procedure and Numerical Examples

We recall the modified projection method [7], which converges if the function F that enters the variational inequality is monotone and Lipschitz continuous and a solution exists, which all hold under our assumptions.

Step 0: Initialization: Set $X^0 \in \mathcal{K}$. Let $\mathcal{T} = 1$ and set α so that $0 < \alpha \leq \frac{1}{L}$, where L is the Lipschitz continuity constant.

Step 1: Computation: Compute $\bar{X}^{\mathcal{T}}$ by solving the variational inequality subproblem:

$$\langle (\bar{X}^{\mathcal{T}} + \alpha F(X^{\mathcal{T}-1}) - X^{\mathcal{T}-1})^T, X - \bar{X}^{\mathcal{T}} \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{20}$$

Step 2: Adaptation: Compute $X^{\mathcal{T}}$ by solving the variational inequality subproblem:

$$\langle (X^{\mathcal{T}} + \alpha F(\bar{X}^{\mathcal{T}}) - X^{\mathcal{T}-1})^T, X - X^{\mathcal{T}} \rangle \geq 0, \quad \forall X \in \mathcal{K}. \tag{21}$$

Step 3: Convergence Verification: If $\max |X_l^{\mathcal{T}} - X_l^{\mathcal{T}-1}| \leq \epsilon$, for all l, with $\epsilon > 0$, a prespecified tolerance, then stop; else, set $\mathcal{T} := \mathcal{T} + 1$, and go to Step 1.

The above computational procedure for the solution of the knowledge collaboration network model yields closed form expressions for (20) and (21) for the collaboration times and the opportunity costs.

The modified projection method described above was implemented in FORTRAN and the computer used was a Sun at the University of Massachusetts Amherst. The convergence criterion was as above with ϵ set to .0001 and we set $\alpha = .01$. We assume that the utilities are directly proportional to the efforts spent and we consider Cobb-Douglas production-type functions, which were also utilized in [2], of the form: $u_{ijk}(x_{ijk}, x_{jik}) = b_{ijk}(x_{ijk})^{1/2}(x_{jik})^{1/2}, \quad \forall i, j, k$. We have that $b_{ijk} = b_{jik}$, which is adopted from Beckmann's notation. We identify the optimal effort allocation plan (and associated opportunity costs) for the knowledge collaboration network model with symmetric utility functions.

In the examples, there are two computer scientists, one operations researcher, and one economist working in a knowledge organization. There are two communication modes with the communication mode represented by the Internet denoted by mode 1 and with the face-to-face communication mode denoted by mode 2. We indexed the four researchers as: 1, 2, 3, and 4, where researchers 1 and 2 are the computer scientists; researcher 3 is the operations researcher, and researcher 4 is the economist. Hence, in the numerical examples, $N = 4$, $O = 2$, and $M = 36$. There are 32 elements in the time/effort allocation vector x and 4 elements in the opportunity cost vector λ. There are 36 variables that we need to compute to determine the optimal solution. In Table 2 we list the data for the relevant parameters: All other t_{ijk} and b_{ijk} terms not reported in Table 2 can be identified from the relationships: $t_{jik} = t_{ijk}$ and $b_{jik} = b_{ijk}$. In addition, recall that $t_{iik} = 1$ for all i, k and we set $b_{iik} = 1$ for all i, i, k.

Example 1: In Example 1, the data were as given above with the time budgets for each researcher $i = 1, \ldots, 4$ being set equal to 100. The modified projection method yielded the following optimal solution: the optimal efforts were given by: $x^*_{121} = x^*_{211} = 100.00$, $x^*_{331} = x^*_{332} = 50.00$, and $x^*_{441} = x^*_{442} = 50$, with all other x_{ijk}s =0.00; the optimal opportunity costs are: $\lambda^*_1 = \lambda^*_2 = 2.00$ and $\lambda^*_3 = \lambda^*_4 = 1.00$. In this example, the computer scientists collaborated only with one another and both the operations researcher and the economist worked alone. Note that all scientists in this knowledge collaboration network used up all the time available in their time budgets. Hence, in this example, there was no

Table 2. Data for Example 1

Collaborators i, j	**Mode** k	$\mathbf{r_{1ij}}$	$\mathbf{r_{2ijk}}$	$\mathbf{r_{ijk}}$	$\mathbf{a_{ijk}}$	$\mathbf{t_{ijk}}$	$\mathbf{b_{ijk}}$
1, 2	1	0	0	0	5	1	2
1, 3	1	3	0	3	5	16	5
1, 4	1	5	0	5	5	26	10
2, 3	1	3	0	3	5	16	5
2, 4	1	5	0	5	5	26	10
3, 4	1	5	0	5	5	26	9
1, 2	2	0	1	1	5	6	1
1, 3	2	3	1	4	5	21	4
1, 4	2	5	1	6	5	31	8
2, 3	2	3	1	4	5	21	6
2, 4	2	5	1	6	5	31	8
3, 4	2	5	1	6	5	31	12

interdisciplinary collaboration. The computer scientists collaborated exclusively using communication mode 1.

Example 2: Example 2 was constructed from Example 1 as follows: the data were identical to that in Example 1, except that the time budgets were increased for all researchers so that now we had that: $T_i = 120$, for $i = 1, \ldots, 4$. The modified projection method yielded the new optimal solution: the optimal efforts were now given by: $x^*_{121} = x^*_{211} = 120.00$, $x^*_{331} = x^*_{332} = 60.00$, and $x^*_{441} = x^*_{442} = 60$, with all other x_{ijk}s =0.00; the optimal opportunity costs were: $\lambda^*_1 = \lambda^*_2 = 2.00$ and $\lambda^*_3 = \lambda^*_4 = 1.00$. Qualitatively, we have the same result as in Example 1, in that an increase in the time budgets still resulted in the computer scientists collaborating with one another, and both the operations researcher and the economist working individually.

Numerous sensitivity analysis exercises and simulations are possible with the above framework as well as game theoretic extensions to capture competition between researchers.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear programming: theory and algorithms. John Wiley & Sons, New York (1993)
2. Beckmann, M.J.: Economic Models of Knowledge Networks. In: Batten, D., Casti, J., Thord, R. (eds.) Networks in Action, pp. 159–174. Springer, Berlin (1995)
3. Beckmann, M.J., Johansson, B., Snickars, F., Thord, R. (eds.): Knowledge and Networks in a Dynamic Economy. Springer, Heidelberg (1998)
4. Börner, K., Maru, J.T., Goldstone, R.L.: The Simultaneous Evolution of Author and Paper Networks. PNAS 101, 5266–5273 (2004)
5. Gibbons, M., Limoges, G., Nowotny, H., Schwartzman, S., Scott, P., Trow, M.: The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies. SAGE Publications, London (1994)

6. Hudson, J.: Trends in Multi-authored Papers in Economics. Journal of Economics Perspectives 10, 153–158 (1996)
7. Korpelevich, G.M.: The Extragradient Method for Finding Saddle Points and Other Problems. Matekon 13, 35–49 (1977)
8. Morillo, F., Bordons, M., Gómez, I.: Interdisciplinarity in Science: A Tentative Typology of Disciplines and Research Areas. Journal of the American Society for Information Science and Technology 54, 1237–1249 (2003)
9. Nagurney, A.: Network Economics: A Variational Inequality Approach, 2nd and revised edn. Kluwer Academic Publishers, Dordrecht (1999)
10. Rosenblat, T.S., Mobius, M.M.: Getting Closer or Drifting Apart. Quarterly Journal of Economics 3, 971–1009 (2004)

Behavioral Analyses of Information Diffusion Models by Observed Data of Social Network

Kazumi Saito[1], Masahiro Kimura[2], Kouzou Ohara[3], and Hiroshi Motoda[4]

[1] School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp
[2] Department of Electronics and Informatics, Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp
[3] Department of Integrated Information Technology, Aoyama Gakuin University
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We investigate how well different information diffusion models explain observation data by learning their parameters and performing behavioral analyses. We use two models (CTIC, CTLT) that incorporate continuous time delay and are extension of well known Independent Cascade (IC) and Linear Threshold (LT) models. We first focus on parameter learning of CTLT model that is not known so far, and apply it to two kinds of tasks: ranking influential nodes and behavioral analysis of topic propagation, and compare the results with CTIC model together with conventional heuristics that do not consider diffusion phenomena. We show that it is important to use models and the ranking accuracy is highly sensitive to the model used but the propagation speed of topics that are derived from the learned parameter values is rather insensitive to the model used.

1 Introduction

The growth of Internet has enabled to form various kinds of large-scale social networks, through which a variety of information including innovation, hot topics and even malicious rumors can be propagated in the form of so-called "word-of-mouth" communications. Social networks are now recognized as an important medium for the spread of information, and a considerable number of studies have been made [1,2,3,4,5]. Widely used information diffusion models in these studies are the *independent cascade (IC)* [6,7,8] and the *linear threshold (LT)* [9,10]. They have been used to solve such problems as the *influence maximization problem* [7,11].

These two models focus on different information diffusion aspects. The IC model is sender-centered and an active node influences its inactive neighbors *independently* with diffusion probabilities assigned to links. On the other hand, the LT model is receiver-centered and a node is influenced by its active neighbors if the sum of their weights

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 149–158, 2010.

exceeds the threshold for the node. Which model is more appropriate depends on the situation and selecting appropriate model is not easy. In order to study this problem, first of all, we need to know how different model behaves differently and how well or badly explain the observation data. Both models have parameters that need be specified in advance: diffusion probabilities for the IC model, and weights for the LT model. However, their true values are not known in practice. This poses yet another problem of estimating them from a set of information diffusion results that are observed as time-sequences of influenced (activated) nodes. To the best of our knowledge, there are only a few methods that can estimate the parameter values for the IC models and its variant that incorporates continuous time delay (referred to as the CTIC model) [3,12,13], but none for the LT model.

With this background, we first propose a novel method of learning the parameter values of a variant of the LT model that incorporates continuous time delay, similar to the CTIC model. We refer to this model as the CTLT model. It is indispensable to be able to cope with continuous time delay to do realistic analyses of information diffusion because, in the real world, information propagates along the continuous time axis, and time-delays can occur during the propagation. Thus, the proposed method has to estimate not only the weight parameters but also the time-delay parameters from the observed data. Incorporating time-delay makes the time-sequence observation data structural. In order to exploit this structure, we introduce an objective function that rigorously represents the likelihood of obtaining such observed data sequences under the CTLT model on a given network, and obtain parameter values that maximize this function by deriving parameter update EM algorithm. Next, we experimentally analyze how different models affect the information diffusion results differently by applying the proposed method to two tasks and comparing the results with the method which we already developed with the CTIC model [13]. The first task is ranking influential nodes in a social network, and we show that ranking is highly sensitive to the model used. We also show that the proposed method works well and can extract influential nodes more accurately than the well studied conventional four heuristic methods that do not take diffusion phenomena explicitly. The second task is the behavioral analysis of topic propagation on a real world blog data. We show that both model well capture the propagation phenomena on different topics at this level of abstract characterization.

2 Proposed Method

2.1 Information Diffusion Model

For a given directed network (or equivalently graph) $G = (V, E)$, let V be a set of nodes (or vertices) and E a set of links (or edges), where we denote each link by $e = (v, w) \in E$ and $v \neq w$, meaning there exists a directed link from a node v to a node w. For each node v in the network G, we denote $F(v)$ as a set of child nodes of v as follows: $F(v) = \{w; (v, w) \in E\}$. Similarly, we denote $B(v)$ as a set of parent nodes of v as follows: $B(v) = \{u; (u, v) \in E\}$. We define the LT model. In this model, for every node $v \in V$, we specify a *weight* ($\omega_{u,v} > 0$) from its parent node u in advance such that $\sum_{u \in B(v)} \omega_{u,v} \leq 1$. The diffusion process from a given initial active set S proceeds according to the following randomized rule. First, for any node $v \in V$, a *threshold* θ_v is

chosen uniformly at random from the interval [0, 1]. At time-step t, an inactive node v is influenced by each of its active parent nodes, u, according to weight $\omega_{u,v}$. If the total weight from active parent nodes of v is at least threshold θ_v, that is, $\sum_{u \in B_t(v)} \omega_{u,v} \geq \theta_v$, then v will become active at time-step $t+1$. Here, $B_t(v)$ stands for the set of all the parent nodes of v that are active at time-step t. The process terminates if no more activations are possible. Next, we extend the LT model so as to allow continuous-time delays, and refer to the extended model as the *continuous-time linear threshold (CTLT) model*. In the CTLT model, in addition to the weight set $\{\omega_{u,v}\}$, we specify real values r_v with $r_v > 0$ in advance for each node $v \in V$. We refer to r_v as the *time-delay parameter* on node v. Note that r_v depends only on v, which means that it is the node v's decision when to receive the information once the activation condition has been satisfied. The diffusion process unfolds in continuous-time t, and proceeds from a given initial active set S in the following way. Suppose that the total weight from active parent nodes of v became at least threshold θ_v at time t for the first time. Then, v will become active at time $t + \delta$, where we choose a delay-time δ from the exponential distribution with parameter r_v. Further, note that even though some other non-active parent nodes of v become active during the time period between t and $t + \delta$, the activation time of v, $t + \delta$, still remains the same. The other diffusion mechanisms are the same as the LT model.

For an initial active node v, let $\varphi(v)$ denote the number of active nodes at the end of the random process for the CTLT model. Note that $\varphi(v)$ is a random variable. Let $\sigma(v)$ denote the expected value of $\varphi(v)$. We call $\sigma(v)$ the *influence degree* of v for the CTLT model.

2.2 Learning Problem

For the sake of technical convenience, we introduce a slack weight $\omega_{v,v}$ for each node $v \in V$ so as to be $\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$. Here note that such a slack weight $\omega_{v,v}$ never contributes to the activation of v. We define the parameter vectors $\boldsymbol{r}$ and $\boldsymbol{\omega}$ by $\boldsymbol{r} = (r_v)_{v \in V}$ and $\boldsymbol{\omega} = (\omega_{u,v})_{(u,v) \in E}$. In practice, their true values are not available. Thus, we must estimate them from past information diffusion histories.

We consider an observed data set of M independent information diffusion results, $\mathcal{D}_M = \{D_m;\ m = 1, \cdots, M\}$. Here, each D_m is a time-sequence of active nodes in the mth information diffusion result (called mth result, hereafter for simplicity),

$$D_m = \langle D_m(t);\ t \in \mathcal{T}_m \rangle, \qquad \mathcal{T}_m = \langle t_m, \cdots, T_m \rangle,$$

where $D_m(t)$ is the set of all the nodes that have first become active at time t, and $\mathcal{T}_m$ is the observation-time list; t_m is the initial observed time and T_m is the final observed time. We assume that for any active node v in the mth result, there exits some $t \in \mathcal{T}_m$ such that $v \in D_m(t)$. Let $t_{m,v}$ denote the time at which node v has become active in the mth result, i.e., $v \in D_m(t_{m,v})$. For any $t \in \mathcal{T}_m$, we set

$$C_m(t) = \bigcup_{\tau \in \mathcal{T}_m \cap \{\tau;\, \tau < t\}} D_m(\tau)$$

Note that $C_m(t)$ is the set of nodes that had become active before time t in the mth result. We also interpret D_m as referring to the set of all the active nodes in the mth result for convenience sake. The problem is to estimate the values of $\boldsymbol{r}$ and $\boldsymbol{\omega}$ from $\mathcal{D}_M$.

2.3 Likelihood Function

For the learning problem described above, we derive the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ in a rigorous way to use as our objective function. Here note that for each node v, since a threshold θ_v is chosen uniformly at random from the interval $[0, 1]$, we can regard each weight $\omega_{*,v}$ as a multinomial probability, namely, $\omega_{v,v} + \sum_{u \in B(v)} \omega_{u,v} = 1$.

Suppose that a node v became active at time $t_{m,v}$ for the mth result. Then, we know that the total weight from active parent nodes of v became at least threshold θ_v at the time when one of these active parent nodes, $u \in B(v) \cap C_m(t_{m,v})$, became first active. However, in case of $|B(v) \cap C_m(t_{m,v})| > 1$, there is no way of exactly knowing the actual node due to the continuous time-delay. Suppose that a node v was actually activated when a node $\zeta \in B(v) \cap C_m(t_{m,v})$ became activated. Then θ_v is between $\sum_{u \in B(v) \cap C_m(t_{m,\zeta})} \omega_{u,v}$ and $\omega_{\zeta,v} + \sum_{u \in B(v) \cap C_m(t_{m,\zeta})} \omega_{u,v}$. Namely, the probability that θ_v is chosen from this range is $\omega_{\zeta,v}$. Here note that such events with respect to different active parent nodes are mutually disjoint. Thus, the probability density that the node v is activated at time $t_{m,v}$, denoted by $h_{m,v}$, can be expressed as

$$h_{m,v} = \sum_{u \in B(v) \cap C_m(t_{m,v})} \omega_{u,v} r_v \exp(-r_v(t_{m,v} - t_{m,u})). \tag{1}$$

Next, we consider any node $w \in V$ belonging to $\partial D_m = \{w; (v, w) \in E \wedge v \in C_m(T_m) \wedge w \notin D_m\}$ for the mth result. Let $g_{m,w}$ denote the probability that the node w is not activated by the node v within the observed time period $[t_m, T_m]$. Here we can naturally assume that each information diffusion process finished sufficiently earlier than the observed final time, i.e., $T_m \gg \max\{t; D_m(t) \neq \emptyset\}$. Thus, as $T_m \to \infty$, we obtain

$$g_{m,w} = 1 - \sum_{v \in B(w) \cap C_m(T_m)} \omega_{v,w}. \tag{2}$$

Therefore, by using Equations (1) and (2), and the independence properties, we can define the likelihood function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M)$ with respect to $\boldsymbol{r}$ and $\boldsymbol{\omega}$ by

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\omega}; \mathcal{D}_M) = \prod_{m=1}^{M} \left(\prod_{t \in \mathcal{T}_m} \prod_{v \in D_m(t)} h_{m,v} \prod_{w \in \partial D_m} g_{m,w} \right). \tag{3}$$

Thus, our problem is to obtain the time-delay parameter vector $\boldsymbol{r}$ and the diffusion parameter vector $\boldsymbol{\omega}$, which maximizes Equation (3). For this estimation problem, we can derive an estimation method based on the Expectation-Maximization algorithm in order to stably obtain its solutions, although we skip its derivation due to a space limitation.

2.4 Behavioral Analysis

Thus far, we assumed that the time-delay and diffusion parameters can vary with respect to nodes and links but independent of the topic of information diffused. However, they may be sensitive to the topic.

Our method can cope with this by assigning a different m to a different topic, and placing a constraint that the parameters depends only on topics but not on nodes and links throughout the network G, that is $r_{m,v} = r_m$ and $\omega_{m,u,v} = q_m|B(v)|^{-1}$ for any node $v \in V$ or link $(u, v) \in E$. Here note that $0 < q_m < 1$ and $\omega_{v,v} = 1 - q_m$. This constraint is required because, without this, we have only one piece of observation for each (m, u, v) and there is no way to learn the parameters. Noting that we can naturally assume that people behave quite similarly for the same topic, this constraint should be acceptable. Under this setting, we can easily obtain the parameter update formulas. Using each pair of the estimated parameters, (r_m, q_m), we can analyze the behavior of people with respect to the topics of information, by simply plotting (r_m, q_m) as a point of 2-dimensional space (See Fig. 2 in Section 3.2).

3 Experiments

We applied the proposed learning method to two tasks to analyze how different models affect the information diffusion results differently and compared the results with the method which we already developed with the CTIC model [13]. First, we applied it to the problem of extracting influential nodes, and evaluated the performance of the CTLT model, i.e. parameter learning and influential node prediction, using the topologies of four large real network data. Next, we applied our method to behavioral analysis using a real world blog data based on the method described in section 2.4 and investigated how each topic spreads throughout the network.

3.1 Ranking Influential Nodes

Experimental Settings. We employed four datasets of large real networks, which are all bidirectional connected networks. The first one is a trackback network of Japanese blogs used in [14] and had 12,047 nodes and 79,920 directed links (the blog network). The second one is a network of people that was derived from the "list of people" within Japanese Wikipedia, also used in [14], and had 9,481 nodes and 245,044 directed links (the Wikipedia network). The third one is a network derived from the Enron Email Dataset [15] by extracting the senders and the recipients and linking those that had bidirectional communications and there were 4,254 nodes and 44,314 directed links (the Enron network). The fourth one is a co-authorship network used in [16] and had 12,357 nodes and 38,896 directed links (the coauthorship network).

Here, we assumed the simplest case where $\omega_{u,v} = q|B(v)|^{-1}$ and $r_v = r$ for any $u, v \in V$. One reason behind this assumption is that there is no need that the observation sequence data have to pass through every link at least once. This drastically reduces the amount of data necessary to learn the parameters. Then, our task is to estimate the values of q and r. The true value of q was decided to be set to 0.9 in order to achieve reasonably high influence degrees of nodes, and the true value of r was decided to be chosen from two values, one with a relatively high value $r = 2$ (a short time-delay case) and the other with a relatively low value $r = 1/2$ (a long time-delay case). The training data $\mathcal{D}_M$ in the learning stage was constructed by generating each D_m from a randomly selected initial active node $D_m(0)$ using the true CTLT model. We chose

Table 1. Parameter estimation accuracy by the proposed method

Blog network		
r^*	$\mathcal{E}_q$	$\mathcal{E}_r$
2	0.024	0.060
1/2	0.017	0.012

Wikipedia network		
r^*	$\mathcal{E}_q$	$\mathcal{E}_r$
2	0.015	0.028
1/2	0.016	0.007

Enron network		
r^*	$\mathcal{E}_q$	$\mathcal{E}_r$
2	0.013	0.031
1/2	0.011	0.004

Coauthorship network		
r^*	$\mathcal{E}_q$	$\mathcal{E}_r$
2	0.023	0.043
1/2	0.024	0.011

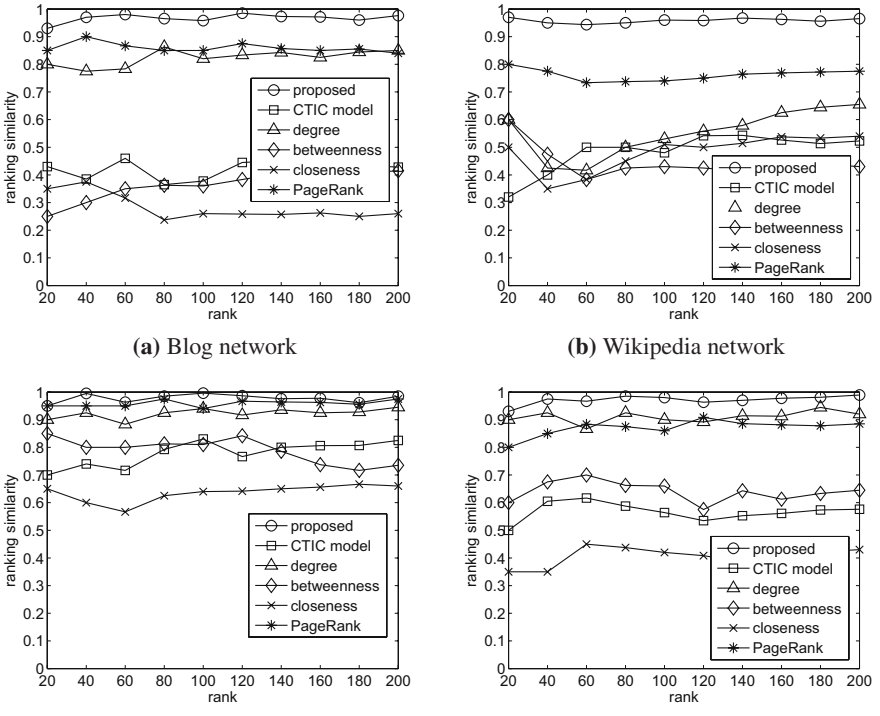
$T_m = \infty$ and used $M = 100$. We repeated the same experiment for each network five times independently.

We measure the influence of node v by the influence degree $\sigma(v)$ for the CTLT model that has generated $\mathcal{D}_M$. We compared the result of the high ranked influential nodes for the true CTLT model predicted by the proposed method with four heuristics widely used in social network analysis and the CTIC model based method [13]. The four heuristics are the same as those used in [13], "degree centrality", "closeness centrality", "betweenness centrality", and "authoritativeness" . The first three heuristics are commonly used as influence measure in sociology [17]. The authoritativeness is obtained by the "PageRank" method [18] which is a well known method for identifying authoritative or influential pages in a hyperlink network of web pages[1]. The CTIC model based method employs the CTIC model as the information diffusion model[13], where we learn the parameters of the CTIC model from the observed data $\mathcal{D}_M$, and rank nodes according to the influence degrees based on the learned model.

Experimental Results. First, we examined the performance of estimating parameters by the proposed method. Let q^* and r^* denote the true values of q and r, respectively. Let $\hat{q}$ and $\hat{r}$ be the values of q and r estimated by the proposed method, respectively. We evaluated the parameter estimation accuracy by the errors $\mathcal{E}_q = |q^* - \hat{q}|$ and $\mathcal{E}_r = |r^* - \hat{r}|$. Table 1 shows the average values of $\mathcal{E}_q$ and $\mathcal{E}_r$ of five trials. We observe that the estimated values were close to the true values. The results demonstrate the effectiveness of the proposed method.

Next, in terms of extracting influential nodes from the network $G = (V, E)$, we evaluated the performance of the ranking methods mentioned above by the *ranking similarity* $\mathcal{F}(k) = |L^*(k) \cap L(k)|/k$ within the rank $k(> 0)$, where $L^*(k)$ and $L(k)$ are the true set of top k nodes and the set of top k nodes for a given ranking method, respectively. We focused on the performance for high ranked nodes since we are interested in extracting influential nodes. Figure 1 shows the results in the case of $r^* = 2$ for the blog, the Wikipedia, the Enron, and the coauthorship networks, respectively. For the proposed and the CTIC model methods, we plotted the average value of $\mathcal{F}(k)$ at k for five experimental results stated earlier. The results in the case of $r^* = 1/2$ for the proposed and the CTIC model methods were very similar to those in the case of $r^* = 2$. We see that the proposed method gives better results than the other methods for these networks, demonstrating the effectiveness of our proposed learning method. We also observe that the CTIC model method does not work well for predicting the high ranked influential nodes for the CTLT model for the problem setting we employed.

[1] As for the jump parameter ε of PageRank, we used a typical setting of $\varepsilon = 0.15$.



(a) Blog network **(b)** Wikipedia network

(c) Enron network **(d)** Coauthorship network

Fig. 1. Performance comparison in extracting influential nodes in the case of $r^* = 2$

3.2 Behavioral Analysis of Real World Blog Data

Experimental Settings. To compare the result by the proposed method with that by the CTIC model based method [13], we used the same real blogroll network as [13], which was generated from the database of a blog-hosting service in Japan called *Doblog*[2]. In the network, bloggers are connected to each other and we assume that topics propagate from blogger x to another blogger y when there is a blogroll link from y to x because this means that y is a reader of the blog of x. In addition, according to [19], it is supposed that a topic is represented as a URL which can be tracked down from blog to blog. We used the same propagation sequences of 172 URLs as [13] for this analysis, each of which is longer than 10 time steps. Please refer to [13] for more detailed description of the network generation and URL sequences.

Experimental Results. We ran the experiments for each identified URL and obtained the corresponding parameters q and r. Figure 2 is a plot of the results for the major URLs. The horizontal axis is the diffusion parameter q and the vertical axis is the delay parameter r. The latter is normalized such that $r = 1$ corresponds to a delay of one day, meaning $r = 0.1$ corresponds delay of 10 days. In general, from this result, it can

[2] Doblog(`http://www.doblog.com/`), provided by NTT Data Corp. and Hotto Link, Inc.

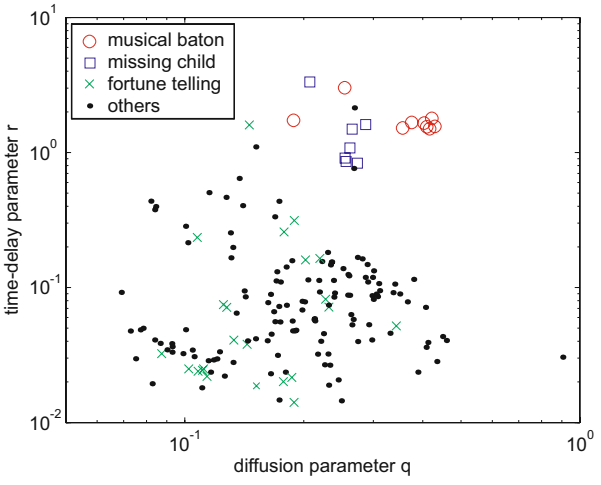


Fig. 2. Results for the Doblog database

be said that the proposed method can extract characteristic properties of certain topics reasonably well only from the observation data. We only explain three URLs that exhibit some interesting propagation properties. The circle is a URL that corresponds to the musical baton which is a kind of telephone game on the Internet. It is shown that this kind of message propagates quickly (less than one day on the average) with a good chance (one out of 25 to 100 persons responds). This is probably because people are easily interested in and influenced by this kind of message passing. The square is a URL that corresponds to articles about a missing child. This also propagates quickly with a meaningful probability (one out of 80 persons responds). This is understandable considering the urgency of the message. The cross is a URL that corresponds to articles about fortune telling. Peoples responses are diverse. Some responds quickly (less than one day) and some late (more than one month after), and they are more or less uniformly distributed. The diffusion probability is also nearly uniformly distributed. This reflects that each individual's interest is different on this topic. The dot is a URL that corresponds to one of the other topics (not necessarily the same).

4 Discussion

With the addition of the proposed method, we now have ways to compare the diffusion process with respect to two models (the CTIC model and the CTLT model) for the same observed dataset. Being able to learn the parameters of these models enable us to analyze the diffusion process more precisely. Comparing the results bring us deeper insights into the relation between models and information diffusion processes. Hence, we consider the contribution of the proposed method is significant.

Indeed, we obtained two interesting insights through the comparative experiments in the previous section. The first one comes from the results of ranking influential nodes, in which the ranking accuracy by the proposed method was better than those by the conventional heuristics, which was sort of expected, but the accuracy by the CTIC method

was not, which is rather surprising. This means that the ranking results that involve detailed probabilistic simulation is very sensitive to the underlying model assumed to generate the observed data. In fact, the similar results were obtained when the role of the two models are switched, i.e. data generated by CTIC and the model assumed to be CTLT (results not shown due to the space limitation). In other words, it is very important to select an appropriate model for the analysis of information diffusion from which the data has been generated. However, this is a very hard problem in reality. The second one comes from the results of the behavior analysis of topic propagation. The pattern shown in Fig.2 was very similar to that by the CTIC method shown in [13]. Regardless of the model used, in both results, the parameters for the topics that actually propagated quickly/slowly in observation converged to the values that enable them to propagate quickly/slowly on the model. Namely, we can say that the difference of models used has little influence on the relative difference of topic propagation property which indeed strongly depends on topic itself. Both models are well defined and can explain this property at this level of abstraction. However, we have to carefully choose a model at least when solving such problems as the influence maximization problem [7,11], a problem at a more detailed level.

5 Conclusion

We considered the problem of analyzing information diffusion process in a social network using two kinds of information diffusion models, incorporating continuous time delay, the CTIC model and the CTLT model, and investigated how the results differ according to the model used. To this end, we proposed a novel method of learning the parameters of the CTLT model from the observed data, and experimentally confirmed that it works well on real world datasets. We also obtained the following two important observations through the experiments for the two tasks. One is that in learning the information diffusion parameters of nodes and links, the learning results are highly sensitive to the model used. The other is that in analyzing the topic-oriented characteristics such as the propagation speed of each topic, using different models has little influence on the analysis results. These two contrasting observations may hold only for well-defined diffusion models such as the CTIC and CTLT models. These findings would help us consider whether we should select a model carefully, or not. In practice, as there are numerous factors that affects the information diffusion process, it is difficult to select an appropriate model in a more realistic setting. This model selection is our future work.

Acknowledgment

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research, U.S. Air Force Research Laboratory under Grant No. AOARD-08-4027, and JSPS Grant-in-Aid for Scientific Research (C) (No. 20500147).

References

1. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
3. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations 6, 43–52 (2004)
4. Domingos, P.: Mining social networks for viral marketing. IEEE Intelligent Systems 20, 80–82 (2005)
5. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. In: Proceedings of the 7th ACM Conference on Electronic Commerce (EC 2006), pp. 228–237 (2006)
6. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters 12, 211–223 (2001)
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2003), pp. 137–146 (2003)
8. Kimura, M., Saito, K., Motoda, H.: Blocking links to minimize contamination spread in a social network. ACM Transactions on Knowledge Discovery from Data 3, 9:1–9:23 (2009)
9. Watts, D.J.: A simple model of global cascades on random networks. Proceedings of National Academy of Science, USA 99, 5766–5771 (2002)
10. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)
11. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI 2007), pp. 1371–1376 (2007)
12. Saito, K., Kimura, M., Nakano, R., Motoda, H.: Finding influential nodes in a social network from information diffusion data. In: Proceedings of the International Workshop on Social Computing and Behavioral Modeling (SBP 2009)., pp. 138–145 (2009)
13. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Zhou, Z.-H., Washio, T. (eds.) ACML 2009. LNCS, vol. 5828, pp. 322–337. Springer, Heidelberg (2009)
14. Kimura, M., Saito, K., Motoda, H.: Minimizing the spread of contamination by blocking links in a network. In: Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI 2008), pp. 1175–1180 (2008)
15. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 217–226. Springer, Heidelberg (2004)
16. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818 (2005)
17. Wasserman, S., Faust, K.: Social network analysis. Cambridge University Press, Cambridge (1994)
18. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
19. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), pp. 207–214 (2005)

Developing Social Networks for Artificial Societies from Survey Data

Stephen Lieberman and Jonathan Alt

Modeling, Virtual Environments and Simulation (MOVES) Institute,
Naval Postgraduate School,
Monterey, California 93943
stlieber@nps.edu

Abstract. Authentically representing large social collectivities remains a preeminent challenge throughout the social computing, and modeling and simulation communities. We demonstrate here a simple technique that uses survey and polling data to embed agents with attributes and endogenously elicit an authentic and theory-driven simulation social structure for an artificial society. We furthermore show that a representation of social structure based on internal agent attributes allows for the continuous representation of social dynamics that affect agent cognition and association, and that social structures for artificial societies can be generated without any loss to the granularity of the underlying data or simulation output. We provide a case study using social survey data to demonstrate the method and effects, document the visualization of social structure for the population of Indonesia, discuss the implications and uses of survey data for social simulation, and suggest several paths forward for social and behavioral predictive modeling.

Keywords: Artificial Societies, Social Network Analysis, Social Computing.

1 Introduction

Over the past several decades, substantial work has gone in to faithfully and reliably representing the cognition, sense-making and decision-making processes of individual agents, and to a lesser extent, aspects of group decision-making [1]. While considerable progress has been made regarding the attributes and cognition of individual agents, putting these agents in a faithful and reliable social context has proven much more elusive [2]. Theory and experimental work from cognitive psychology, behavioral psychology, and cybernetics has proven to be extremely valuable in both understanding and generating the conceptual and cognitive features of simulated actors (agents). Interest in modeling and simulation approaches to large human groups, such as societies, has stimulated applied research in sociology and social psychology, although some researchers have lamented the lack of focus on theory building for social simulations and artificial societies [3]. In a companion paper, we discussed using a survey instrument to population the cognitive identities of agents in an artificial society. Here we present a well-documented social theory and an application of the theory to social simulation using a contemporary case study.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 159–168, 2010.

Populating artificial societies for use in analysis, training, and experimentation requires the identification and proper use of reliable data sources for authentically representing the social structure of the population being modeled. Survey and polling data present a compelling source of information from which to draw inferences about a population's social structure, and the impact of this structure on the boundaries and opportunities for agents to communicate information and carry out actions. These types of data present a traceable means of informing simulation effort from empirical data. The ability to gain insight into how individuals and groups in a population are likely to think and behave makes social simulation an attractive tool for decision makers across a very wide variety of disciplines.

2 Models and Method

We demonstrate here the process of instantiating an artificial society with data from an existing open source survey, the World Values Survey. In a simulation context, each artificial society must include management mechanisms for 1) internal agent attributes (such as each agent's beliefs and planned actions), and 2) inter-agent interactions (such as group affiliations and communications between agents). In a companion paper, we discussed instantiating the aspects of internal entity cognition. Here we focus specifically on instantiating social structure based on internal agent attributes, however defined. As a contemporary example, we describe this process using a social survey of 2015 respondents from the population of Indonesia.

2.1 Instantiating Internal Agent Attributes

Although the precise mechanisms for ensouling agents with attributes for entity cognition is beyond the scope of this article, a brief explanation of the essential methods is necessary for full description of the social network representation[1]. The implementation used in this example relies on the concept of the narrative paradigm which states that each human possess a unique identity based on their culture and their life experiences [4]. This "narrative identity" forms the lens through which an individual views the world and interprets events, and is implemented as a Bayesian belief network (BBN) for each agent within the model [5]. These foundational beliefs are treated as Bayesian priors [6]. The theory of planned behavior, which regulates the formation of intention to act within the model, describes the manner in which people develop intentions to carry out behaviors, and states that individuals form an intention to act based on three factors as they relate to the behavior in question: 1) attitude, 2) perception of the group norms, and 3) perceived level of behavioral control [7]. Bayesian networks are a general approach to understanding how the human mind executes induction about the world based on noisy observations provided by world experience [8].

The instantiation of internal agent attributes starts with the identification of the analysis-relevant issues. A population entity's stance on these issues becomes the response of interest to the simulation user. When developing these internal attributes

[1] Available in a companion article, same volume, are a full description, analysis, and case study regarding entity that uses the same World Values Survey dataset.

from survey data the population's initial state on the issue of interest is informed directly by the data. Treating the population's response on that each issue as the target, factor selection is conducted over the remaining survey item responses for each socio-demographic subtype of interest. The most relevant contributors are used to generate a Bayesian network for each socio-demographic subtype from the source data, the relevant survey items. These Bayesian networks control the agent's stance on each issue. Multiple techniques exist for conducting feature selection and for generating Bayesian networks from various types of source data [9][10][11].

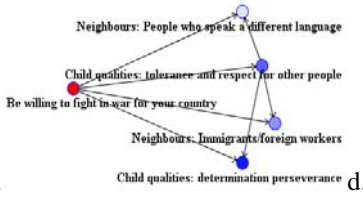
For example, analyzing contributors to responses to the Likert scale survey item, "I am willing to fight in war for my country" for a specific socio-demographic subtype (agent) identifies five relevant items (Figure 1a). Controlling for covariance and generating a BBN from this information yields a graphical representation of entity cognition (Figure 1b), and basis for instantiating agent cognition. Likewise, analyzing contributors and generating a BBN for a different subtype agent yields corresponding results and inputs for that agent (Figure 1c, d). In this way, the internal representation of cognition for each agent is taken directly from the original survey instrument.

a.

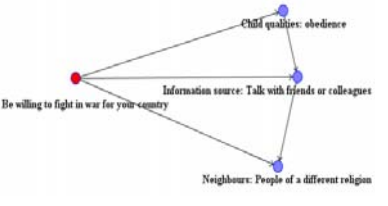
Item	Value
Child qualities: tolerance and respect for other people	1
Neighbours: Militant minority	0.982
Child qualities: determination perseverance	0.98
Neighbours: People who speak a different language	0.98
Neighbours: Immigrants/foreign workers	0.96

b.

Item	Value
Child qualities: determination perseverance	0.995
Information source: Talk with friends or colleagues	0.995
Moments of prayer, meditation...	0.992
Neighbours: People of a different religion	0.991
Child qualities: obedience	0.984



c.



d.

Fig. 1. Representation of internal agent cognition for two socio-demographic subtypes. Relevant contributors and Bayesian Belief Networks (BBNs) for subtype 1 (a, b) and subtype 2 (c, d).

2.2 From Internal Attributes to Social Structure

We define social structure as the distributions of social factors in the population under consideration. Social factors are those attributes we believe influence interaction, communication and affiliation among agents in the simulation, including sociocultural, sociodemogrpahic, and socioeconomic attributes such as age, ethnicity, religiosity, occupation, and socioeconomic status (SES). Our framework represents social structure in the form of a social network where agents are linked to one another on the basis of their similarity. The more closely two agents resemble one another in terms of their social factors, the more likely they are to interact and communicate within the simulation. While every agent in the simulation is linked, the strength of a link between two agents manages the frequency of their interaction such that agents will generally communicate about important issues with only those most similar to themselves.

Fig. 2. A sample 50-actor homophily network showing social clusters and intergroup connectivity

This realization of social structure follows on work from Blau, *et al.*[12][13], McPherson, *et al.*[14][14][15], and many others [16][17][18] that seeks to develop a framework for the understanding and investigation of society through the use of individual attributes in an ecology of identities. By generating a social structure based on similarity between attributes, we are employing in a simulation context one of the most empirically demonstrated findings in social life: the principle of homophily. Homophilous interactions are those that take place among actors that share (any number of) similar attributes. This is generally expressed as "birds of a feather", or "like associates with like".

Linking each agent on the basis of their social factor similarity results in what we call a *homophily network* of the population (Figure 2). Since the social factors of agents change during the course of the simulation, the homophily network is dynamic, changing shape as new strong ties emerge and weak ties drift on the basis of social similarity. As the simulation progresses, the underlying homophily network always characterizes the instantaneous likelihood of interaction between every pair of individuals in the population.

The simulation framework described here preserves the homophily of the original survey data in every way. Each agent in the simulation is ensouled with the complete range of social factors of a single survey or poll respondent. Thus, there is a one-to-one ratio of agents to original survey respondents preserving the granularity of the original data. Each survey respondent is a data point in the social structure and thus every agent in the simulation is a socially complete representation of a real individual. While a subset of the social factors input, the beliefs, values and interests (BVIs), are used in the entity cognition model, the social structure module can make use of the entirety of the survey data for the construction of the social network. Thus, the social structure of the population under consideration is a direct realization of the distribution of social factors from the survey input data.

The World Values Survey of Indonesia looks at over 250 questions regarding survey respondent's occupation, socioeconomic status (SES), BVIs. Many of these questions take the form of Likert scale responses such that similarity can be easily calculated between individual respondents. After each agent is ensouled with survey data (see above), the responses between every pair of agents for every survey item are compared and a final distance metric is calculated for every pair of agents. The result of these calculations is an exact representation of the original survey data where each individual is placed in "social space" and the weight of the link between two individuals indicates the level of similarity they share. Sharing a high level of

a.

ID	Age	Job	Class	Income
0	18	4	3	6
1	24	8	3	4
2	33	5	4	6
3	57	3	3	4
4	41	8	5	3

b. $s_{0,1} = |18-24|+|4-8|+|3-3|+|6-4|$

Fig. 3. Deriving simple social distance. Social factor input matrix (a), and calculation of simple social distance for agents 0 and 1 using input matrix (b).

similarity in this social space (i.e., a strong homophilous relationship) indicates a stable and strong tie with the corresponding strong likelihood of communication about important events.

The generation of a homophily network from survey data requires several steps. Following the selection of appropriate survey items (i.e., social factors), one must decide how to handle missing data. While the transformation appears robust against missing data, surveys of large populations generally have a large number of respondents, so we have chosen here to remove from the calculations any individual that did not respond to all of our selected items.

Homophily networks of populations tend to be quite large. The last WVS wave of Indonesia (from 2005 through 2007) has responses for 2015 individuals. The corresponding fully connected (undirected) network has over 2 million links. Following the description of instantiating internal attributes (above), we select four empirically robust dimensions for this demonstration: Age, Job, Class, and Income. The *Job* item here refers to the level of occupational prestige associated with a particular profession across 11 points in descending order of prestige from *manager of 10 or more employees* to *agricultural worker*. The *Class* item is a subjective measure of social class with 6 points from *Upper Class* to *Lower Class*, and the *Income* item is a subjective measure of relative income in 10 steps.

By removing individuals that did not respond to all four dimensions, our resulting population contains 1050 agents[2]. We index the individuals from 0 through n to provide docking for the internal attributes, and calculate the distance (delta) between each pair of agents (agent i and agent j) in *dimension* (e.g., difference in Likert responses for a specified survey item). Summing the distances in each dimension for a pair of agents yields the total social distance s_{ij} of the pair in terms of the input factors. Figure 3 shows a sample social factor input matrix and the straightforward corresponding calculation of *simple social distance* between agents 0 and 1.

Taking the square root of the sum of the squares of each term (survey item delta) yields the exact Euclidean distance of the agent-pair in an h-dimension hypercube where h is the number of survey items in the input matrix. Since the likelihood of a pair's interaction holds an inverse relationship to social distance, a final link weight w_{ij} must be calculated for each pair based on their social distance and the square root of h. Each of these calculations must be made dynamically during each time step of

[2] A complete undirected network of 1050 agents yields 550,725 links and just over 1 million network calculations per simulation time step.

the simulation to authentically represent the changing range of social factors in each dimension[3].

The final calculation of link weight connects every agent in the social network with a weight $0 < w_{ij} < 1$, yielding a complete (fully connected) weighted network that can immediately be used for network analysis, algorithmic and visual investigation of groups, isolates, key players, and so on. While this network is static—it is a "snapshot" of the homophily of the population at the time of the survey—it serves as the input to a dynamic simulation of social structure that changes in response to exogenous influences (such as simulation events) and endogenous influences (such as communication patterns between agents) as the simulation progresses. The homophily network of the population can be analyzed throughout the simulation as new groups and key players emerge, and others disappear.

3 Results and Discussion

3.1 Visualizing Societies

While any number of dimensions can be used for analysis, the final social distance and link weight calculations result in a *single* metric for each pair of agents, the results can be visualized (and analyzed) as a traditional 2-dimensional social network. Figure 4 shows two visualizations of the homophily network of Indonesia generated directly from the World Values Survey data. While all agents are connected in the simulation, for purposes of visualization we apply a trim to links below a certain threshold. Varying trim levels allows us to investigate different features of the society. We are currently undertaking work to characterize the semantics of various trim levels using social network survey information, such as the General Social Survey's Social Networks modules[4]. In general, for large social networks, we visualize only those *strongest* relationships between individuals thus reducing our rate of error and focusing our analysis on the most meaningful social connections.

Figure 4a visualizes the social structure of all of Indonesia with a trim of 90 (all links with a weight under 90 are not shown, and not used by the visualization algorithm). We clearly see two large groups of individuals with very high in-group connectivity and much sparser inter-group connectivity. We also see a few smaller groups and dozens of representative individuals on the social periphery. This representation immediately suggests approaches to understanding the culture and driving social forces of Indonesia in terms of two large and sparsely connected groups of people. Furthermore, it implies that social policy, marketing processes, and so on, must be geared towards, or at least cognizant of, two distinct socially-heterogeneous groups that comprise the vast majority of Indonesian society.

---

[3] Static normalization will save cycles and reduce run-time, but does not effectively model endogenous changes in social space volume, which is a central aspect to faithfully modeling changes in social structure.

[4] For a list of GSS social science modules, see `http://www.norc.org/GSS+Website/About+GSS/National+Data+Program+for+Social+Sciences/`

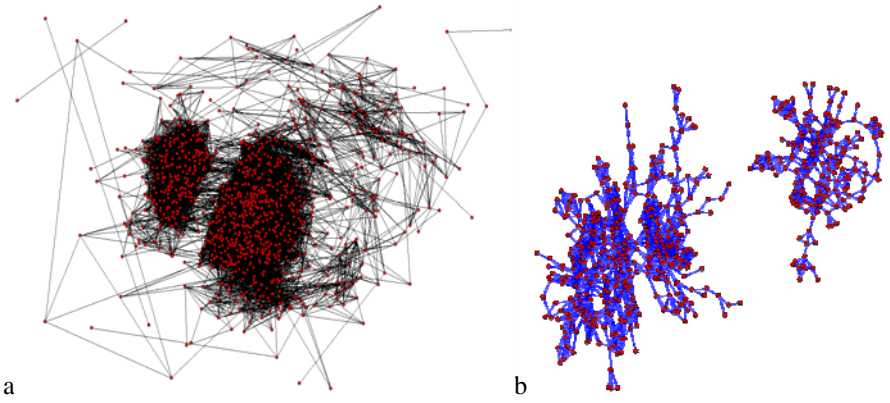


Fig. 4. Two visualizations of the population of Indonesia. Visualization of the entire population at a trim of 90 (a), and detailed visualization of the two major (largest) components of the population at a trim of 99 (b) [5].

Figure 4b visualizes the details of the social structure of the two largest components in the whole social network of Indonesia. At a trim of 99, only those most closely associated individuals are linked, note the similar (nearly identical) link lengths between all pairs of agents in the visualization. These two groups are very cohesive, with no apparent core-periphery structure, and dense interconnectivity and clustering (triadic closure) between individuals in each group. Most interestingly, while the in-group connectivity is extremely high, each component exhibits a high density, between-group connectivity is absent at a trim of 99. In Figure 2 above we can see that an alternate topology involves large dense groups linked through a small number of intermediaries that associate strongly with individuals at the periphery of two or more clusters. In fact, the larger group (on the left) appears to have two factions connected through intermediaries. The presence of factions can be determined analytically, but it is important to note that at a trim of 99, there is a strong homophilous tendency among individuals linked though short chains of agents. Since the structural barriers to connection are very low, the likelihood of actual association at this level is quite high.

3.2 Communication Networks

While the homophily network manages the likelihood of interaction, the *actual* interactions that take place in the simulation generate what we call the *communication network*. Communication networks can be analyzed for traceability and event-causal mapping. Where the homophily network is an instantaneous representation of the population, the communication network necessarily contains a representation of time, i.e., communications that have taken place 1) during a set period of time, 2) leading up to a specified event, or 3) taking place after a specified event (Figure 5).

[5] These networks were visualized using Pajek (`http://vlado.fmf.uni-lj.si/pub/networks/pajek/`).

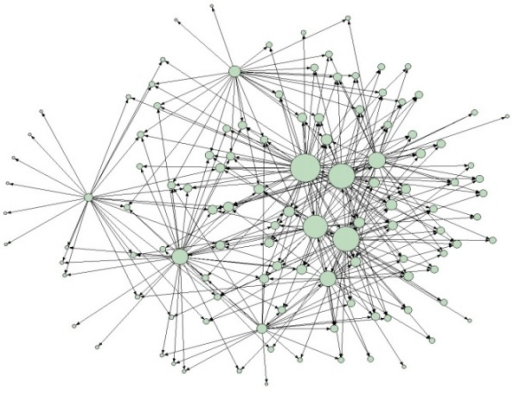


Fig. 5. Inter-agent communication network following an exogenous event. Node size reflects importance of each actor to the communication network[6].

Communication networks provide a fascinating and unique glimpse in to representative agent behavior that we could not obtain through noisy and incomplete observation using real populations. Homophily networks and communication networks can be analyzed together to reveal stunning facets of the simulated society. By analyzing both simultaneously, we can, for example, investigate questions about the control and spread of information that were before outside of empirical possibility. Future work in the social and behavioral modeling, and larger modeling and simulation communities will elucidate new analytic techniques and tools for these types of questions. It is imperative that these communities establish functional and timely practices to share modeling and simulation methods, and model output. Equally important is the necessity for multiple universities and research teams to coordinate work on common germane datasets, such as the World Values Survey described here, to share the implementations of applicable theory and simulation output.

3.3 Uses of Survey Data

The selection of criteria (survey items) for inclusion in the homophily network is dependent on survey familiarity and follows sociological theory with regards to social structure. Of the 360 survey questions in the WVS for Indonesia, approximately 230-250 are suitable for inclusion. Further research and sensitivity analyses are needed to determine how many dimensions are needed for accurate and reliable analysis, with the guiding theory that higher dimensionality results in a more authentic representation of social structure.

Given that the input data is survey responses, changes in the structure of the population network (and corresponding BVIs) can be seen as changes in survey responses over time. Many authors and researchers have used survey responses as a proxy for human behavior, and while in many situations it may be spurious to

[6] Visualized using UCINet (http://www.analytictech.com/downloaduc6.htm), nodes sized by eigenvector centrality.

casually link reported beliefs to actual behavioral tendencies, the methods described here *do* elucidate behavior in two specific ways. The tendency for the human behaviors of social association and group preference based on observed homophily is extremely well documented [12-21]. Furthermore, social simulations (including the one described here) can make use of sociological and anthropological ideas of association and behavior tendencies over time to elucidate changes in the likelihood that representative individuals will engage in specific dyadic and group associations, i.e., as an endogenous product of social structure. Likewise, simulations using a homophily network directly derived from survey data essentially track the behavior of changing survey responses over time. While there is a wealth of literature on survey interpretation, more research needs to be performed regarding changes in survey response behavior as related to social structure.

4 Conclusion

Authentically representing large social collectivities has remained a preeminent challenge throughout the history of artificial intelligence and the modern application of agent-based approaches to modeling and simulation [1]. We demonstrated here that a simple technique can transform survey and polling data to embed agents with attributes and represent large social structures, such as the population of Indonesia. We furthermore posit that these representations are scale-independent: functional social structure can be characterized for groups ranging from simple dyads to extremely large human collectivities including nation states and international collectivities without any loss to the granularity of the underlying data or simulation output.

Future work is underway to demonstrate how the use of these types of data can greatly enhance the feasibility of validation efforts for social and behavioral simulations, by authenticating simulation output against social survey waves. The above case study and analysis demonstrated these methods and tools, and the discussion suggested several paths forward for future social and behavioral predictive modeling research. Most notably, there is an urgent need within the social computer and behavioral modeling community to share models, simulations, and output, engage in open exchanges of source code, and share the results of analyzing common data sets (such as the World Values Survey) using a variety of established and novel techniques.

References

[1] Russel, S., Norvig, P.: Artificial intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs (2003)
[2] Liu, H., Salerno, J., Young, M. (eds.): Social Computing, Behavioral Modeling, and Prediction. Springer, Heidelberg (2008)
[3] National Research Council, Behavioral Modeling and Simulation: From Individuals to Societies. National Academies Press, Washington (2008)
[4] Fisher, W.: Clarifying the Narrative Paradigm. Communication Monographs 56, 55–58 (1989)

[5] Smith, K., Kalish, M.L., Griffiths, T.L., Lewandowsky, S.: Introduction. Cultural transmission and the evolution of human behaviour. Philosophical Transactions B 363, 3469 (2008)
[6] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182 (2003)
[7] Ajzen, I.: The theory of planned behavior. Organizational behavior and human decision processes 50, 179–211 (1991)
[8] Griffiths, T.L., Tenenbaum, J.B.: Randomness and coincidences: Reconciling intuition and probability theory. In: Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society. Human Communication Research Centre. University of Edinburgh, Edinburgh, August 1-4, p. 370 (2001)
[9] Pfeffer, A.: 1 The Design and Implementation of IBAL: A General-Purpose Probabilistic Language
[10] Tian, J., Kang, C., Pearl, J.: A characterization of interventional distributions in semi-Markovian causal models. In: Proceedings of The National Conference on Artificial Intelligence, p. 1239 (2006)
[11] Ullman, S., Sali, E., Vidal-Naquet, M.: A fragment-based approach to object representation and classification. In: Arcelli, C., Cordella, L.P., Sanniti di Baja, G. (eds.) IWVF 2001. LNCS, vol. 2059, pp. 85–102. Springer, Heidelberg (2001)
[12] Blau, P., Schwartz, J.: Crosscutting Social Circles: Testing A Macrostructural Theory of Intergroup Relations. Transaction Publishers (1997)
[13] Blau, P.: Structural Context of Opportunities. University of Chicago Press, Chicago (1994)
[14] McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology 27, 415–444 (2001)
[15] McPherson, M., Ranger-Moore, J.: Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space. Social Forces 70, 19–42 (1991)
[16] Smith-Lovin, L.: Self, identity, and interaction in an ecology of identities. Advances in identity theory and research, 167–178 (2003)
[17] Lewis, T.: Network Science: Theory and Applications. John Wiley and Sons, Chichester (2009)
[18] Jackson, M.: Social and Economic Networks. Princeton University Press, Princeton (2008)
[19] Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70, 66111 (2004)
[20] Edmonds, B.: How are physical and social spaces related? In: Agent-Based Computational Modelling, Springer (2006), http://cfpm.org/cpmrep127.html (Downloaded on March 10, 2008)
[21] Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103, 8577 (2006)

Understanding and Enabling Online Social Networks to Support Healthy Behaviors

Noshir Contractor

Director, Science of Networks in Communities (SONIC),
University of Illinois at Urbana-Champaign,
2145 Sheridan Road, TECH D241,
Evanston, Illinois 60208-3119 USA

Abstract. Recent advances in digital technologies invite consideration of social influence and social support as processes that are accomplished by global, flexible, adaptive, and ad hoc networks that can be created, maintained, dissolved, and reconstituted with remarkable alacrity. This presentation describes and empirically tests a multi-theoretical multilevel (MTML) model of the socio-technical motivations for creating, maintaining, dissolving, and reconstituting knowledge and social networks.

The presentation argues that MTML insights based on understanding the social motivations to create links in online social networks should be used to design more effective social networks to support healthy behaviors. Specifically, these insights should be used to implement algorithms that make network recommendations that can serve as interventions to enhance and target social support and social influence within online networks such as smoking cessation communities.

The insights we propose are derived from our study of large scale social networks within massively multi-player online role-playing games (MMORPGs). Enabled by advanced graphic and networking technologies, MMORPGs provide three-dimensional playgrounds for people to interact with one another. In this study, we analyzed activities in an MMORPG, Sony's EverQuest II. We analyzed the entire network of 3140 players who were on one server (Antonia Bayle) from Aug 25 to Aug 31 2006. Of these, 2998 were from the US, 142 were from Canada. 2447 were males. We examined whether their geographic distance offline and their demographic similarity (or homophily) influence the likelihood of four online interactions: partnering, instant messaging, trading, and mailing. The results show that geographical proximity of distance and temporal proximity of time zones have a strong impact in players' online behavior in creating relations. Individuals were 22.6 times more likely to link with others within 50 kms than from someone who was within 50 to 800 kms. In addition, homophily in age and game experience also had a strong impact on creating relations. However, there was no evidence of gender homophily in the virtual world.

These results indicate that online social networks to support healthy behaviors should implement network recommendation algorithms that underscore the importance of connecting individuals who are geographically proximate as well as exhibit homophily on a variety of personal traits.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, p. 169, 2010.

A Dynamical Systems Model for Understanding Behavioral Interventions for Weight Loss

J.-Emeterio Navarro-Barrientos[1], Daniel E. Rivera[1], and Linda M. Collins[2]

[1] Control Systems Engineering Laboratory,
School of Mechanical, Aerospace, Chemical and Materials Engineering,
Arizona State University, Tempe, AZ, USA
{jnavarrob, daniel.rivera}@asu.edu
[2] The Methodology Center and
Department of Human Development and Family Studies
Penn State University, State College, PA, USA
lmcollins@psu.edu

Abstract. We propose a dynamical systems model that captures the daily fluctuations of human weight change, incorporating both physiological and psychological factors. The model consists of an energy balance integrated with a mechanistic behavioral model inspired by the Theory of Planned Behavior (TPB); the latter describes how important variables in a behavioral intervention can influence healthy eating habits and increased physical activity over time. The model can be used to inform behavioral scientists in the design of optimized interventions for weight loss and body composition change.

Keywords: behavioral interventions, theory of planned behavior, dynamical systems, energy balance, weight loss.

1 Introduction

Obesity rates in the United States have increased substantially in the past few decades [13]. Because obesity represents a preventable cause of premature morbidity and death, much research activity has been devoted to understanding its causes, and a number of diverse solutions have been proposed [6,9,19]. Some of these solutions have major disadvantages, for instance, surgery and extreme diets usually lead to a regain of the weight lost within 1 to 5 years [2,11]. Solutions leading to permanent weight loss require sustained lifestyle changes in an individual; consequently, developing optimized behavioral interventions that promote healthy eating habits and increased physical activity represents a problem of both fundamental and practical importance.

The primary goal of this paper is to improve the understanding of behavioral weight change interventions by expressing these as dynamical systems. Dynamical systems modeling has been used in the analysis of novel behavioral interventions such as adaptive interventions [17]. Dynamical systems modeling considers how important system variables (e.g., proximal and distal outcomes, mediators)

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 170–179, 2010.

respond to changes in input variables (e.g., intervention dosages, external characteristics) over time. A dynamical systems model can be used to answer questions regarding what variables to measure, how often, and the speed and shape of the outcome responses as a result of decisions regarding the timing, spacing, and dosage levels of intervention components.

To achieve this goal, we develop in this paper a dynamical systems model for daily weight change incorporating both physiological and psychological considerations. For the physiological component, we rely on the concept of energy balance to obtain a model that describes the net effect of energy intake from food minus energy consumption, the latter which includes physical activity. For the psychological component, we present a model for the dynamics of diet and exercise behavior. This model explains how intentions, social norms, attitudes, and other system variables that are impacted by an intervention result in healthy eating habits and increased physical activity over time. A model based on the Theory of Planned Behavior (TPB) is used for this purpose. The dynamical systems model for weight change can be used to answer questions regarding how much to eat, what kinds of food to eat, how much physical activity to undertake, and how long it will take before desired weight loss goals are achieved.

The paper is organized as follows: Section 2 presents the energy balance model, while Section 3 gives a brief description of the Theory of Planned Behavior (TPB) and presents a mechanistic dynamical model for TPB based on fluid analogies. Section 4 describes a representative simulation from this model and discusses the role and importance of some of the parameters in the model. Finally, Section 5 summarizes our main conclusions and discusses areas of current and further study.

2 Energy Balance Model

In this section, we present the energy balance model used for our investigation. This model is based on a three-compartment model proposed in [10,11,12]. The normal daily energy balance $EB(t)$ is described as follows:

$$EB(t) = EI(t) - EE(t), \tag{1}$$

where $EI(t)$ is the energy intake and $EE(t)$ is the energy expenditure at time t, measured at daily intervals in this study. The energy intake EI, expressed in kilocalories (kcal), is modeled using the Atwater methods of energy calculation resulting from carbohydrate intake (CI), fat intake (FI), and protein intake (PI), all expressed in grams/day [14]:

$$EI(t) = a_1 CI(t) + a_2 FI(t) + a_3 PI(t) \tag{2}$$

Here $a_1 = 4$ kcal/gram, $a_2 = 9$ kcal/gram, and $a_3 = 4$ kcal/gram. The energy expenditure EE, expressed in kcal, is calculated as follows:

$$\begin{aligned} EE(t) = \beta EI(t) + \delta BM + K + \gamma_{LM} LM(t) + \gamma_{FM} FM(t) + \\ \eta_{FM} \frac{d\,FM}{dt} + \eta_{LM} \frac{d\,LM}{dt} \end{aligned} \tag{3}$$

The first term, $\beta EI(t)$, denotes the energy expended in processing food ($\beta = 0.24$), δ is the physical activity coefficient (expressed in kcal/kg), $\gamma_{LM} = 22$ kcal/kg/d, $\gamma_{FM} = 3.2$ kcal/kg/d, $\eta_{LM} = 230$ kcal/kg and $\eta_{FM} = 180$ kcal/kg are all coefficients for the calculation of the Resting Metabolic Rate (RMR) which depends on the lean mass LM and the fat mass FM. The constant K accounts for initial energy balance conditions and is determined by solving equation (3) assuming an initial steady-state at $t = 0$; the steady-state is denoted by a bar over any time-dependent variable ($\overline{EI} - \overline{EE} = 0$, with $d\overline{FM}/dt = d\overline{LM}/dt = 0$ by definition of steady-state):

$$K = -\gamma_{LM}\overline{LM} - \gamma_{FM}\overline{FM} - \delta\overline{BM} + \overline{EI}(1-\beta). \tag{4}$$

The three-compartment model for fat mass FM, lean mass LM and extra-cellular fluid volume ECF is summarized as follows:

$$\frac{d\,FM(t)}{dt} = \frac{(1-p(t))EB(t)}{\rho_{FM}} \tag{5}$$

$$\frac{d\,LM(t)}{dt} = \frac{p(t)EB(t)}{\rho_{LM}} \tag{6}$$

$$\frac{d\,ECF}{dt} = \frac{\Delta Na_{diet} - \xi_{Na}(ECF - ECF_{init}) - \xi_{CI}(1 - CI/CI_b)}{[\mathrm{Na}]\,\tau_{Na}}, \tag{7}$$

where $\rho_{FM} = 9400$ kcal/kg, $\rho_{LM} = 1800$ kcal/kg and p is given by the Forbes formula [8]:

$$p = \frac{C}{(C + FM)}; \qquad C = 10.4\frac{\rho_{LM}}{\rho_{FM}}. \tag{8}$$

For the extracellular fluid volume (in ml), ΔNa_{diet} is the change on sodium in mg/d, CI_b is the baseline carbohydrate intake, $[Na] = 3.22$ mg/ml, $\xi_{Na} = 3$ mg/ml/d, $\xi_{CI} = 4000$ mg/d, ECF_{init} is the initial ECF volume and $\tau_{Na} = 2$, which corresponds to a time constant of two days. Finally, the body mass is given by the sum of fat mass FM, lean mass LM and extracellular fluid volume ECF:

$$BM(t) = FM(t) + LM(t) + ECF(t). \tag{9}$$

In summary, the mechanistic energy balance leads to a dynamical model with EI (composed of CI, FI, and PI), δ, and ΔNa_{diet} as inputs, and FM, LM, and ECF as outputs; the outputs add up to total body mass (BM).

3 Behavioral Model

The Theory of Planned Behavior (TPB) [1] is an accepted and broadly used paradigm for describing the relationship between behaviors and intentions, attitudes, norms, and perceived control in behavioral science. Many studies have relied on TPB and its forerunner the Theory of Reasoned Action (TRA) [7] to describe behavioral changes for healthy eating [3] and exercising [5]. Behavior is the observable response in a given situation with respect to a given target, while intention is an indication of the readiness of a person to perform a given behavior. According to TPB, intention is influenced by the following components:

Attitude Toward the Behavior: This is the degree to which performing the behavior is positively or negatively valued. It is determined by the *strength of beliefs about the outcome* and the *evaluation of the outcome*.

Subjective Norm: This is the perceived social pressure to engage or not engage in a behavior. It is determined by the strength of the beliefs what people want the person to do, also called *normative beliefs*, and the desire to please people, also called *motivation to comply*.

Perceived Behavioral Control: This reflects the perception of the ability to perform a given behavior, i.e. the beliefs about the presence of factors that may facilitate or impede performance of the behavior. It is determined by the *strength of each control belief* and the *perceived power of the control factor*.

A standard mathematical representation for TPB relies on Structural Equation Modeling (SEM) [4]. The field of SEM is substantial, but in this work we limit ourselves to a special case of SEM called path analysis. The main characteristics of path analysis models is that they do not contain latent variables, i.e., all problem variables are observed, and the independent variables are assumed to have no measurement error [16]. The TPB represented as a path analysis model with a vector η of endogenous variables and a vector ξ of exogenous variables is expressed as follows:

$$\eta = \mathbf{B}\,\eta + \mathbf{\Gamma}\,\xi + \zeta \tag{10}$$

$$\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \eta_5\end{bmatrix} = \begin{bmatrix}0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ \beta_{41} & \beta_{42} & \beta_{43} & 0 & 0\\ 0 & 0 & \beta_{53} & \beta_{54} & 0\end{bmatrix}\begin{bmatrix}\eta_1\\ \eta_2\\ \eta_3\\ \eta_4\\ \eta_5\end{bmatrix} + \begin{bmatrix}\gamma_{11} & 0 & 0\\ 0 & \gamma_{22} & 0\\ 0 & 0 & \gamma_{33}\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\xi_1\\ \xi_2\\ \xi_3\end{bmatrix} + \begin{bmatrix}\zeta_1\\ \zeta_2\\ \zeta_3\\ \zeta_4\\ \zeta_5\end{bmatrix} \tag{11}$$

where $\mathbf{B}$ and $\mathbf{\Gamma}$ are matrices of β_{ij} and γ_{ij} regression weights, respectively, and ζ is a vector of disturbance variables. Figure 1 shows the intention-behavior TPB path analysis model for equation (11). Typically, the principles of TPB assume that the attitude toward the behavior ξ_1, the subjective norms ξ_2 and the perceived behavioral control ξ_3 are estimated using the expectancy-value model, which considers the sum over the person's behavioral beliefs, normative beliefs and control beliefs, respectively, that are accessible at the time. However, for simplicity of presentation and without loss of generality, we consider only one exogenous variable per compartment in this paper. Thus,

$$\xi_1 = b_1 \times e_1 \tag{12}$$

$$\xi_2 = n_1 \times m_1 \tag{13}$$

$$\xi_3 = c_1 \times p_1, \tag{14}$$

where b_1 is the behavioral belief, e_1 the evaluation of the outcome, n_1 the normative belief, m_1 the motivation to comply, c_1 the control belief and p_1 the power of the control belief.

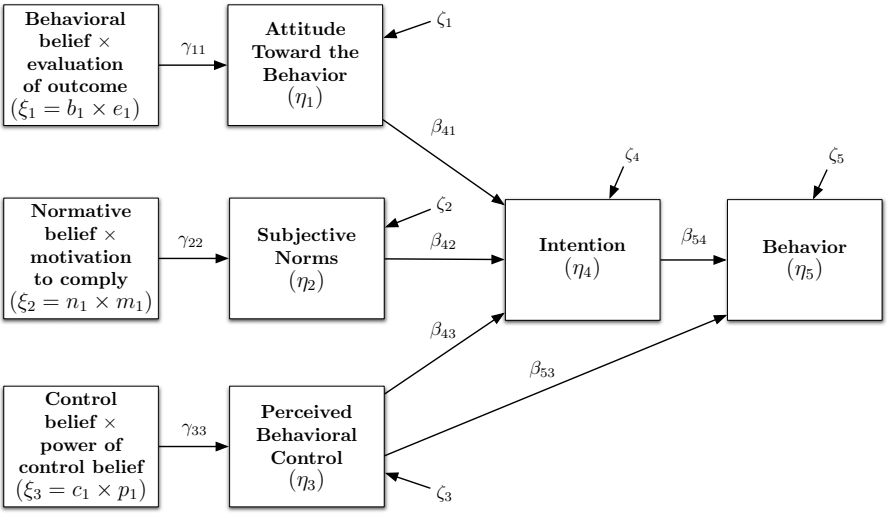


Fig. 1. Path analysis diagram for the Theory of Planned Behavior (TPB) with three exogenous variables ξ_i, five endogenous variables η_i, regression weights γ_{ij} and β_{ij} and disturbances ζ_i

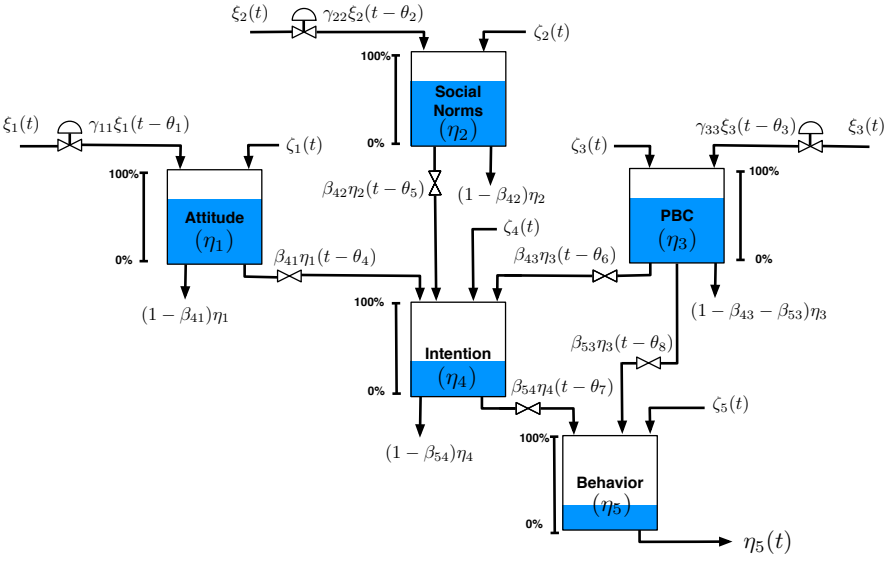


Fig. 2. TPB as a dynamical system, representation based on a fluid inventory control system. Time delays are modeled by $\theta_1, \cdots, \theta_8$. Additional parameters as in Figure 1.

3.1 Dynamic Fluid Analogy for TPB

The classical TPB model as expressed in equation (10) represents a static (i.e., steady-state) system that does not capture any changing behavior over time. In order to expand the TPB model to include dynamic effects, we propose the use of a fluid analogy which parallels the problem of inventory management in supply chains [18]. This analogy is expressed diagrammatically in Figure 2. We consider a dynamic fluid analogy of TPB with five *inventories*: attitude η_1, social norms η_2, perceived behavioral control η_3, intention η_4 and behavior η_5. Each inventory is replenished by inflow streams and depleted by outflow streams. The path diagram model coefficients $\gamma_{11}, \cdots, \gamma_{33}$ are the inflow resistances and $\beta_{41}, \cdots, \beta_{54}$ are the outflow resistances, which can be physically interpreted as those fractions of the inventories of the system that serve as inflows to the subsequent layer in the path analysis model.

To generate the dynamical system description we apply the principle of conservation of mass to each inventory, where accumulation corresponds to the net difference between mass inflows and outflows:

$$\text{Accumulation} = \text{Inflow} - \text{Outflow} \tag{15}$$

Relying on the rate form for equation (15) leads to a system of differential equations according to:

$$\tau_1 \frac{d\eta_1}{dt} = \gamma_{11}\xi_1(t-\theta_1) - \eta_1(t) + \zeta_1(t) \tag{16}$$

$$\tau_2 \frac{d\eta_2}{dt} = \gamma_{22}\xi_2(t-\theta_2) - \eta_2(t) + \zeta_2(t) \tag{17}$$

$$\tau_3 \frac{d\eta_3}{dt} = \gamma_{33}\xi_3(t-\theta_3) - \eta_3(t) + \zeta_3(t) \tag{18}$$

$$\tau_4 \frac{d\eta_4}{dt} = \beta_{41}\eta_1(t-\theta_4) + \beta_{42}\eta_2(t-\theta_5) + \beta_{43}\eta_3(t-\theta_6) - \eta_4(t) + \zeta_4(t) \tag{19}$$

$$\tau_5 \frac{d\eta_5}{dt} = \beta_{54}\eta_4(t-\theta_7) + \beta_{53}\eta_3(t-\theta_8) - \eta_5(t) + \zeta_5(t), \tag{20}$$

where, following equations 12-14, $\xi_1(t) = b_1(t)e_1(t)$, $\xi_2(t) = n_1(t)m_1(t)$, $\xi_3(t) = c_1(t)p_1(t)$, and $\zeta_1, \cdots, \zeta_5$ are zero-mean stochastic signals. The dynamical system representation according to equations 16 through 20 includes all the path analysis model parameters and is enhanced by the presence of *time delays* $\theta_1, \cdots, \theta_8$ (which model the lag in the inflow/outflow process) and *time constants* $\tau_1, \cdots, \tau_5$ (which model the capacities of the inventories) for each inventory in the system. These parameters can be used to determine the speed at which an individual or population can transition between values for $\eta_1, \cdots, \eta_5$ as a result of changes in the variables ξ_i. A number of important points of interest are summarized below:

1. At steady-state, i.e., when $\frac{d\eta_i}{dt} = 0$, equations 16 through 20 reduce to the path analysis model in equation (11) *without approximation.*

2. The path analysis model coefficients γ_{ij} and β_{ij} correspond directly to *gains* in the dynamical system.
3. The outflow resistances from the inventory PBC are subject to the constraint: $\beta_{53} + \beta_{54} \leq 1$.
4. The dynamical model representation is not limited to describing single subjects. Typically, path analysis models are naturally estimated cross-sectionally from data obtained from multiple participants. Dynamical system model parameters can similarly be estimated over a group or cohort, but doing so will require availability of repeated measurements of system variables over time.

4 Simulation Study

The overall dynamical systems model for the behavioral intervention integrates the model described in Section 2 and the TPB dynamical model described in Section 3 by having the outputs of the TPB model serve as inputs to the mechanistic energy balance model. This enables the impact of the intervention to be observed in both psychological and physical outcome variables over time.

The simulation study consists of examining the effects over time of an intervention promoting healthy eating habits and increased physical activity for a representative male participant at the following initial conditions: $BM = 100$ kg, $FM = 30$ kg, $LM = 45$ kg and $ECF = 25$ liters. Figure 3 (top) shows the responses of the intervention on TPB models for energy intake behavior (EI-TPB) and physical activity behavior (PA-TPB). Figure 3 (bottom) shows the changes in the body compartments corresponding to these interventions. We consider a scenario that as a result of the intervention the intensity of beliefs about healthy eating habits increases from $b_1 = 7$ to $b_1 = 10$. This change leads to an increase on the exogenous variable ξ_1 in the EI-TPB system. In the same manner, we assume that as a result of the intervention there is a change in the beliefs about proper exercising from $b_1 = 1$ to $b_1 = 3$, which also leads to an increase on the variable ξ_1 but in the PA-TPB system. No outflow from the inventory PBC to the inventory behavior is considered for both behavioral models, i.e. $\beta_{53} = 0$. For this simulation study we consider the following three sub-scenarios:

i) The participant completely assimilates the intervention and immediately starts changing eating habits and exercising. This means a rapid time constant in the attitude inventory ($\tau_1 = 0.1$), no depletion in the intention inventory ($\beta_{41} = 1$) and no delay in the behavior inventory ($\theta_7 = 0$).

ii) The participant partially and slowly assimilates the intervention. This means a slow time constant in the attitude ($\tau_1 = 20$ days), depletion on intention of $\beta_{41} = 0.5$ (only 50% of the outflow makes the next inventory) and a delay in behavior of $\theta_7 = 15$ days.

iii) The same scenario as for case ii) but with disturbances on the attitude towards healthy eating and exercising. These disturbances are represented as white noise signals on energy intake EI and physical activity δ with means and variances $N(0, 20)$ and $N(0, 50)$, respectively.

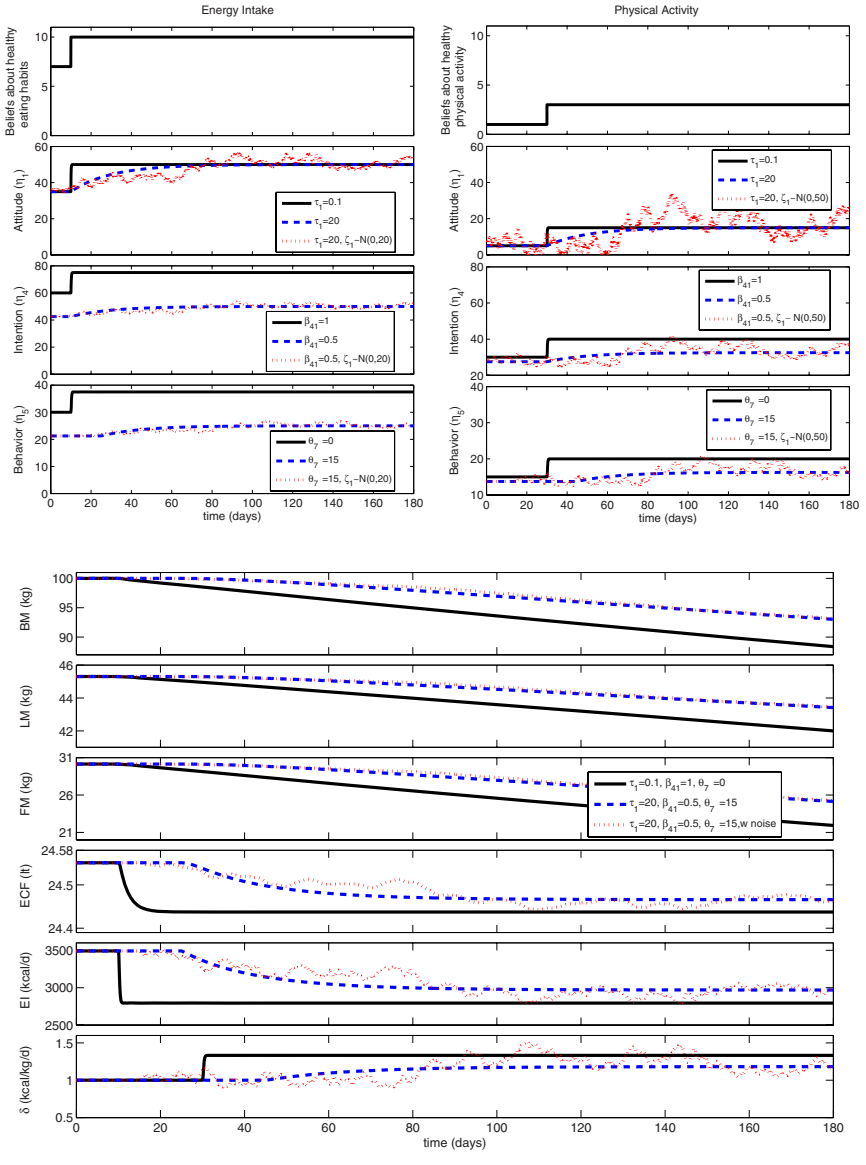


Fig. 3. (top) Step responses for the inventory systems of energy intake behavior EI-TPB and physical activity behavior PA-TPB, for interventions on beliefs about the outcome b_1; (bottom) changes on body compartments and total effect of intervention on EI and PA. Simulations for the following intervention cases: (i) complete assimilation $\tau_1 = 0.1, \beta_{41} = 1, \theta_7 = 0$, (ii) partial assimilation $\tau_1 = 20, \beta_{41} = 0.5, \theta_7 = 15$, and (iii) partial assimilation as in case (ii) but with noise $\zeta_1 \sim N(0, 20)$ (for EI), and $\zeta_1 \sim N(0, 50)$ (for PA). Further parameters: $\xi_2 = \xi_3 = 50$, $\theta_1 = \cdots = \theta_6 = 0$, $\theta_8 = 0$, $\tau_2 = \cdots = \tau_5 = 0.1$, $\gamma_{ij} = 1$, $\beta_{42} = \beta_{43} = \beta_{54} = 0.5$ and $\beta_{53} = 0$.

After a step change in the variable ξ_1 is introduced, the magnitude level of the inventories for attitude, intention and behavior change, respectively. Observe that the attitude response in scenario (ii) takes a larger number of time steps to reach the steady-state when compared to scenario (i). This occurs because of the different τ_1 values, the larger value of τ_i represents a longer transition of the system to the new steady state. Moreover, the level of intention in scenario (ii) is much lower than in scenario (i). The smaller β_{ij} the larger the depletion. Finally, the change in behavior in scenario (ii) starts much after the change in behavior in scenario (i). The larger the value for θ_i, the longer the delay.

A number of conclusions can be drawn from these simulation results. The most significant is the contrast between the results of case (i) versus case (ii) after a six-month time period. The weight for the participant in case (i) decreases by almost 10 kg over the six month period, whereas for case (ii) the weight loss is only 7 kg. The behavioral "lag" has resulted in a substantially lower achievable weight loss for the participant. However, despite the presence of stochastic disturbances in case (iii), these do not result in significant differences on total body weight loss when compared to case (ii).

5 Summary and Conclusions

A dynamical systems model for a behavioral intervention associated with weight loss has been proposed, which provides a potentially useful framework for understanding and optimizing this class of interventions. By being able to test the effect of intervention components on outcomes of interest over time, an intervention scientist can use this information to optimally decide on the ordering and strength of intervention components, and better predict both the inter- and intra-individual variability that will be reflected in these interventions.

An extended version of the dynamic TPB model has been developed where all endogenous variables are latent as opposed to observed variables; it has not been presented in this paper for reasons of brevity.

The simulation results point to the need for data from experimental trials or observational studies that can be used to estimate parameter coefficients in these models and validate the modeling framework. We are currently exploring how methods from the field of system identification [15], coupled with data resulting from participant diaries or ecological momentary assessment, can be used for this purpose. A long-term goal is to develop adaptive behavioral interventions for preventing weight gain or loss in patients with obesity or malnutrition, relying on control systems engineering principles [17].

Acknowledgments. The assistance of Drs. Kevin Hall and Carson Chow from the National Institute on Diabetes and Digestive Diseases (NIDDK) in developing the energy balance model is gratefully acknowledged. We appreciate helpful suggestions provided by Dr. Inbal Nahum-Shani at Penn State. Support for this research has been provided by the Office of Behavioral and Social Sciences Research of the National Institutes of Health (OBSSR) and the National Institute on Drug Abuse (NIDA) through grants R21 DA024266 and P50 DA010075.

References

1. Ajzen, I., Madden, T.J.: Prediction of goal-directed behavior: attitudes, intentions, and perceived behavioral control. J. Exp. Soc. Psychol. 22, 453–474 (1986)
2. Baranowski, T., Cullen, K.W., Nicklas, T., Thompson, D., Baranowski, J.: Are current health behavioral change models helpful in guiding prevention of weight gain efforts? Obesity Research 11, 23S–43S (2003)
3. Blanchard, C.M., Fisher, J., Sparling, P.B., Shanks, T.H., Nehl, E., Rhodes, R.E., et al.: Understanding adherence to 5 serving of fruits and vegetables per day: A theory of planned behavior perspective. J. Nutr. Edu. Behav. 41(1), 3–10 (2009)
4. Bollen, K.A.: Structural equations with latent variables. Series in probability and mathematical statistics. Wiley, Chichester (1989)
5. Symons Downs, D., Hausenblas, H.A.: The theories of reasoned action and planned behavior applied to exercise: A meta-analytic update. J. Phys. Act. Health 2, 76–97 (2005)
6. FAO/WHO/UNU Expert Consultation, Human energy requirements, Food and Nutrition Technical Report Series 1, FAO, Rome (October 2001)
7. Fishbein, M., Ajzen, I.: Belief, attitude, intention and behavior: An introduction to theory and research. Addison-Wesley, Reading (1975)
8. Forbes, G.B.: Lean body mass-body fat interrelationships in humans. Nutr. Rev. 45, 225–231 (1987)
9. Garrow, J.S.: Energy balance and obesity in man, 2nd edn. Elsevier/North-Holland Biomedica Press, Amsterdam (1978)
10. Hall, K.D.: Computational model of in vivo human energy metabolism during semistarvation and refeeding. Am. J. Physiol. Endocrinol Metab. 291, E23–E37 (2006)
11. Hall, K.D., Jordan, P.N.: Modeling weight-loss maintenance to help prevent body weight regain. Am. J. Clin. Nutr. 88, 1495–1503 (2008)
12. Hall, K.D., Chow, C.: A simple dynamic model of body weight and composition change. personal communication
13. Keim, N.L., Blanton, C.A., Kretsch, M.J.: America's obesity epidemic: Measuring physical activity to promote an active lifestyle. J. Am. Diet. Assoc. 104(9), 1398–1409 (2004)
14. Merrill, A.L., Watt, B.K.: Energy value of foods - basis and derivation. In: Agriculture Handbook. U.S. Department of Agriculture, vol. 74 (1973)
15. Ljung, L.: System Identification: Theory for the User. Prentice Hall Information and System Sciences Series. Prentice Hall, Englewood Cliffs (1987)
16. Raykov, T., Marcoulides, G.A.: A First Course in Structural Equation Modeling, 2nd edn. Erlbaum, Mahwah (2006)
17. Rivera, D.E., Pew, M.D., Collins, L.M.: Using engineering control principles to inform the design of adaptive interventions: A conceptual introduction. Drug and Alcohol Dependence 88, S31–S40 (2007)
18. Schwartz, J.D., Wang, W., Rivera, D.E.: Optimal tuning of process control-based decision policies for inventory management in supply chains. Automatica 42, 1311–1320 (2006)
19. Trumbo, P., Schlicker, S., Yates, A.A., Poos, M.: Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids. J. Am. Diet Assoc. 102(11), 1621–1630 (2002)

COLBERT: A Scoring Based Graphical Model for Expert Identification

Muhammad Aurangzeb Ahmad and Xin Zhao

Department of Computer Science and Engineering, University of Minnesota
mahmad@cs.umn.edu, zhao0111@umn.edu

Abstract. In recent years a number of graphical models have been proposed for Topic discovery in various contexts and network analysis. However there is one class of document corpus, documents with ratings, where the problem of topic discovery has not been explored in much detail. In such document corpuses reviews and ratings of documents in addition to the documents themselves are also available. In this paper we address the problem of discovery of latent structures in document-review corpus which can then be used to construct a social network of experts. We present a graphical model COLBERT that automatically discovers latent topics based on the contents of the document, the review of the document and the ratings of the review.

Keywords: Expert Identification, Topic Modeling, COLBERT.

1 Introduction

Graphical Models for discovering latent structure in document corpora has been applied in a number of different settings. Thus given a document corpus one can discover latent topics in the corpus based on the content of the documents and additional information if it is available. The relationships that are discovered can be one to one (author and document), one to many (multiple authors for the same topic) or many to many (multiple authors for multiple topics). Examples of such relationships and their respective contexts include document corpus data[21][5][13] and e-mail datasets[12]. In this paper we present a graphical model, the COLBERT (COrolated Latent BEhavior Related Topic) Model, which discovers latent topics in document categories by taking into account document reviewers and the ratings of the reviews. By taking into account 'groups' formed by the reviewers we also exploit the social network for the topic discovery task. We use the epinions dataset which consists of product reviews and ratings of the reviews. We note that product category discovery is analogous to topic discovery in a document corpus.

We use the epinions dataset for our experiments. Epinions is a website which contains information about a large number of products ranging from books to movies to software. These products are grouped together into categories, sub-categories and super-categories. Users of epinions can post reviews of these products. The quality of these reviews can be further evaluated by other reviewers. Additionally the website stores a wide range of information related to the products and the users e.g., user's

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 180–188, 2010.

rating of products, ratings of reviews, the social network of the users, textual information regarding the product, textual and numerical reviews of the product, ratings of specific aspects of the product e.g., in the case of movies this can be the 'suspense factor', 'action factor' and 'special effects.' The Epinions.com data thus provides a rich avenue for exploring multi-relational data. In the present setting we consider the review, review of reviews, the author and the reviewer information.

To avoid confusion with regards to the user who is the author of the document (which is a review in this case) and a user that reviews the review we use the term ’primary reviewer’ to designate the user who originally reviewed the product and the term 'secondary reviewer' to refer to the user who reviewed the product review. Since the graphical model gives us Author-Topic pairs and groups based on reviews and ratings as well, we can use this information to build a social network which is contingent upon the similarity between topics reviewed by primary and secondary reviewers. While the problem of expert identification in social networks has been explored before, to the best of our knowledge the subject of expert identification via latent topic discovery has not been addressed to a great extent. With respect to topic modeling and social networks, it is possible to adopt two different approaches: One can take into account the latent group structure of the community while generating topics [16], also known as social topics. Alternatively one can first discover topics, the authors and reviewers associated with the topics and then construct a social network from this information. In this paper we take the later approach. The rest of the paper is organized as follows: In section 2 we describe related work in topic modeling, social networks and expert identification. In section 3 we describe our graphical model and in section 4 we describe the experiments and the results. Conclusion and future work is described in section 5.

2 Related Work

Although the work described in this paper mainly builds upon topic modeling, we apply topic modeling to other domains and thus in addition to topic modeling we also describe related work in expert identification and social network analysis. A number of graphical models have been created in recent years to study different aspects of structured data. The study of structured data has a long history in the field of information retrieval. One of the first well known models was Latent Semantic Indexing (LSI) that creates a term-document matrix that describes the occurrences of terms in a corpus documents. The matrix is then transformed into a relation between terms and concepts [6]. An important improvement over LSI was the pLSI (probabilistic LSI) model where the idea is to model topics as distributions over words and documents are assumed to be generated by the activation of different topics. A more general model was the Latent Dirichlet Allocation (LDA) which used variational EM techniques for parameter estimation. It should be noted that the pLSI can be considered equivalent to the LDA model under a uniform Dirichlet prior distribution. While these and other models only considered textual information for topic discovery, more sophisticated models included additional information. Thus, for example, in the Author-Topic model [13] each author is represented as a multinomial distribution over words and each document is a distribution over topics which are a

mixture of distribution of authors for that document. Similarly in the Group-Topic model[20] discovers groups and topic pairs in a network. The Author-Recipient-Topic Model extends the Author-Topic model for modeling message data[13]. It takes into account the senders and the recipients of a message and consequently model topic discovery is partially driven by the social network of the people in the corpus. The aforementioned models assume that topics are not correlated which may not always be a reasonable assumption. The Correlated Topic Model models pairwise correlations amongst topics and the Pachinko Allocation Model [22] models correlations across multiple topics. Another model that takes into account the ratings on documents was presented by Blei et al [23].

The problem of expert identification in social networks has been addressed in the case of social networks constructed via e-mails. Schwartz and Wood et al.[18] analyzed e-mail flows to identify groups of individual with common interests. The ContactFinder [10] was a system that found the right person to forward a query using the text and addresses of messages on bulletin boards. In a similar manner the ExpertFinder system used topic keywords and frequencies of mentions of a person near the keyword to determine expertise scores and ranks [11]. Campell et al. [4] described a graph-based ranking approach that takes into account the content of email communication to determine experts in a network. Ahmad et al. [1] describe an Ant Colony Optimization based technique for expert identification in social networks. Balog et al.[2] describe two strategies for expert identification: In the first method the expert is identified based on the documents that they are associated with. In the second approach they identify the documents associated with the topics and then identify experts associated with the topics.

There is a vast body of literature on social network analysis [19], the body work which is most relevant to this paper is community detection or community extraction in social networks. Since social networks can be represented as graphs one can use graph partitioning techniques [15] or other graph based techniques [14] to discover communities in social networks. Spectral bisection methods [17] improve an initial partition of the network by optimizing the number of inter and intra community edges via a greedy search. These methods employ the eigenvectors of the graph Laplacian. Hierarchical clustering [19] is another method used for community detection. The main idea is to define a similarity metric between the vertices of the graph. Edges are then added between the vertices based on the level of similarity. Girvan and Newman's [7] algorithm and its variants are based on divisive methods where edges are removed from the graph instead of being added to it. These methods can detect community structure without having to specify the number of communities beforehand.

3 The COLBERT Model

In this section we propose a graphical model for topic extraction, the COLBERT model for author-reviewer-score based topic modeling in reviewer data. A topic is defined as a distribution of words over the text corpus. The main theme of the topic is reflected in the keywords that are generated for the topic. These are the words that have a high probability of being associated with the topic and thus can be used to

characterize a topic. In a similar manner a document is represented as a mixture of distribution of topics. The greater the probability of association of a topic with a document the greater is the representation of the keywords that are associated with the topic in that document. It should be noted that a document can be associated with multiple topics.

The current model builds upon other well known graphical model like the Author-Topic (AT) model and the Author Recipient Topic (ART) model. Just like the Author-Topic model the COLBERT model can also be used to infer authors and topics. A crucial difference between these two models is that in the Author-Topic model there can be multiple authors for a given document while in our case there is always one author for any given document but there are also multiple secondary reviewers for a topic. In other ways the model is also similar to the Author-Recipient-Topic (ART) model as the reviewers in our setting can be said to play a similar role as the recipients in the ART model. On the other hand a crucial difference between the two is that the primary and the secondary reviewer jointly determine the topic in our setting. COLBERT also incorporates scores given by the secondary reviewers. No such analogue exists for the ART model.

The standard plate notation used for graphical models is given for the COLBERT model in Figure 1. In the plate notation each circle represent a variable, the dark circles represent observables, the Greek letters represent the priors on the distributions and the subscript on associated with a rectangle, called a block, represents the number of times that it occurs. Thus Figure 1 describes that the set of authors A is a distribution over the set of topics z, each topic T is a distribution over words W, each topic-score T-S is a distribution over the reviews. The author a, the topic words w in a document w and the score s for a review are the observables in the model. The topic z is a latent variable. α is a prior on the distribution of authors, β is the prior on the distribution of topics and γ is a prior on the distribution of the topic-review pairs.

The assignment of words to topics is done by probabilistic sampling, the most commonly used technique is Gibbs sampling. The conditional probability that for a topic given the word, all other topics, reviewer score, author and the priors is given by the following equation.

$$P(Z_i = t \mid w, z_{-i}, r, s, a, \alpha, \beta, \gamma) \propto \frac{C_{at}^{AT} + \alpha}{\sum_{t'} C_{at}^{AT} + A\alpha} \frac{C_{wt}^{WT} + \beta}{\sum_{w'} C_{w't}^{WT} + W\beta} \frac{C_{srt}^{SRT} + \gamma}{\sum_{r'} C_{sr't}^{SRT} + R\gamma}$$

where CAT represents the count matrix for the author-topic pairs, CWT is the count matrix for the word-topic pairs and the CSRT is the count matrix for the scores-reviews-topic pairs. The generation scheme for the COLBERT model for each document can be described as follows:

1. A topic z is generated based on the primary reviewer which is observable.
2. The word w is generated from the topic z.
3. The reviewer is generated based on the topic z and the score s.

In the COLBERT model a topic can be conceptualized as a distribution over authors. Similarly words can be conceptualized as distributions over topics and reviewers can be thought of as mixture of distributions over topics and scores. It should be noted that the discrete variables as shown in Figure 1 have Dirichlet priors.

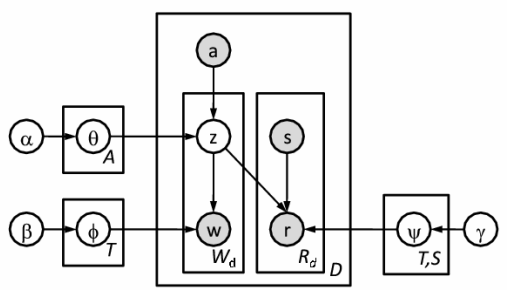


Fig. 1. Plate notation for COLBERT. the set of authors *A* is a distribution over the set of topics *z*, each topic *T* is a distribution over words *W*, each topic-score *T-S* is a distribution over the reviews. The author *a*, the topic words *w* in a document *w* and the score *s* for a review are the observables in the model. The topic *z* is a latent variable. α is a prior on the distribution of authors, β is the prior on the distribution of topics and γ is a prior on the distribution of the topic-review pairs.

In the epinions dataset the observable variables are as follows: The primary reviewer of the product, the set of secondary reviewers, words associated with the review and the scores given to the review. Since we incorporate the secondary reviewer's scores in the generate model, a clarification with regards to the semantics of the model is necessary here. Given that COLBERT is a generative model the choice of using the score as 'generating' a reviewer may appear odd at first. It may seem more straightforward to assume that the score is generated by the secondary reviewer and not vice versa. Our rea-son for using a different semantic interpretation is that assuming that the primary reviewer and the secondary reviewer together generate the topic is equally arbitrary since such an approach would imply a causal relation from the secondary reviewer and the topic.

Yet another approach could be that one could have only the topic generate the score but this would leave out important information related to the topic. Thus in our semantic interpretation of the model, the secondary reviewer is generated by the score and the topic which implies that given the topic and a set of finite possible states for the scores the user 'chooses' to become a reviewer for the topic. The subtle point here is that we treat the score given to a document (review) as a reaction the secondary reviewer has to the review and the score associated with it.

4 Experiments and Results

We used the epinions data for experiments on the COLBERT model. Performance is measured by comparing the average query time i.e., the minimum distance that it will take for a randomly selected node and a random topic to reach a node which is relevant to that topic, for a network constructed from COLBERT and a network based on links between primary and secondary reviewers. Epinions is a website which contains information about thousands of products which can be reviewed by the users of the website. The product can range from toys to books to movies etc. Users of

Table 1. Examples of Topic-Author-Reviewer sets discovered by COLBERT

Keywords (Topic 84)		**Keywords (Topic 3)**		**Keywords (Topic 1)**	
east	0.03458	christian	0.021591	Children	0.027092
heroic	0.02578	God	0.018965	Age	0.02052
warrior	0.02425	Bible	0.01065	Movie	0.02052
emperor	0.02342	World	0.010358	Suitable	0.018276
free	0.02154	Faith	0.009775	Suitability	0.018115
fight	0.02100	message	0.009337	Plot	0.017474
honor	0.02066	christians	0.008754	Older	0.016032
deception	0.02000	Special	0.007587	good	0.013467
save	0.01984	Effects	0.007441	clear	0.012024
hero	0.01945	nonchristian	0.007149	opinion	0.011864
Authors (Topic 84)		**Authors (Topic 3)**		**Authors (Topic 1)**	
user24	0.065041	user659	0.444366	user98	0.075329
user899	0.060976	user34	0.113951	user34	0.053055
user112	0.052669	user534	0.033465	user68	0.043929
user231	0.051432	user351	0.020333	user211	0.03604
user53	0.044362	user4	0.018215	user7	0.033101
user25	0.042595	user231	0.012567	user1	0.031555
user127	0.04241	user547	0.011155	User561	0.030162
user545	0.031283	user77	0.009178	user11	0.028306
user45	0.029516	user410	0.009037	user53	0.02645
user48	0.026158	user99	0.008755	user46	0.025677
Reviewers (Topic 84)		**Reviewers (Topic 3)**		**Reviewers (Topic 1)**	
user87	Off Topic	user449	Off Topic	user111	Off Topic
user104	Not Helpful	user316	Not Helpful	user88	Not Helpful
user104	SW Helpful	user555	SW Helpful	user88	SW. Helpful
user119	Helpful	user94	Helpful	user1210	Helpful
user111	Very Helpful	user786	V. Helpful	user117	Very Helpful
user119	Most Helpful	user786	M. Helpful	user41	Most Helpful

epinions can express their satisfaction or dissatisfaction about products, write a numerical and textual review of the product. The website also gives the reviewers the option to write reviews of reviews. As a consequence users who have been highly rated by many other users can build up a reputation amongst other users. Additionally a secondary reviewer has the option of giving one of five ratings to the review. The products on the website are divided into general categories like Electronics and Restaurants to finer and finer grained categories. The data was collected by crawling the Epinions.com website. For the purpose the current set of experiments and one particular category i.e., movies was collected. The category of movies and videos is further divided into smaller categories like horror, comedy, science fiction etc. The dataset used for the experiments has the following characteristics: There are 4676 documents, 54,061 unique words, 2,648 unique authors, 7,047 reviewers and 1,552,810 words. The reviews in our dataset had a maximum of 200 secondary reviewers.

The results from the COLBERT model are given in Table 1. Each column represents a Topic and each topic has three blocks, where the first block refers to the top ten words associated with that topic, the second block refers to the top ten primary reviewers associated with that topic and the last block refers to one reviewer from the top five reviewers associated with each scoring category. Notice that in this case one can have duplicates since the scoring patterns of people can vary. Thus Topic 84 is talking about heroic movies, while Topic 3 is talking about Christian movies and Topic 1 is talking about children movies.

The problem of expert identification in a social network can be described as follows. Given a social network of *N* actors where a subset of actors E are experts in a number of topics such that E << N, the task is to identify the set of actors which are experts in those topics. In almost all the settings where this problem is addressed it is often assumed that the set of topics is known beforehand either by a centralized repository or in a distributed manner. We also note that the case where the topics are not known before hand have been addressed previously [1] but our approach is a more general approach based on latent topic discovery. Given a topic T_i, and set of scores S_{ij} for the documents dj related to the topic, we define the expertise of an actor E_{ai} in the social network as follows:

$$E_{ai} = \frac{\sum_{d_j \in T_i} \sum_{r_{ij} \in d_j} S_{ij}}{\max(s_{ij}) * |r_{Ti}|}$$

Where r_{ij} is a reviewer who has reviewed documents written by user *a* for topic T_i and r_{Ti} is the total number of people who have reviewed a document on the topic T_i. The reason that the numerator is divided by the quantity r_{Ti} in the denominator is to avoid bias in the case where a set of reviewers rate each other highly and thus they do not end up with the high score, $\max(s_{ij})$ denotes the maximum possible score for a document. This quantity is 5 in our case. As stated previously the metric that we use to evaluate performance is the average query time i.e., the distance that it will take for a randomly selected node and a random topic to reach a node for the COLBERT network and a network constructed by observing the primary and the secondary reviewers. The summary of the results for 200 such cases are given in Table 2.

Table 2. Comparison of Query Times

Network Type	**Avg. Query Time**
COLBERT Network	3.54
Primary-Secondary Reviewer Network	6.85

The table indicates that the average query time for COLBERT is much less as compared to a network constructed by just looking at the reviewers themselves. It should also be noted that it may appear that the ART model may be used in a similar manner for determining experts. However the ART model does not take into account the ratings of documents since it is best suited for messaging data like e-mails where the concept of scoring e-mails by the sender would be meaningless. Most of the literature on trust in social networks assumes that users who are friends with one

another -are likely to have similar characteristics or profiles [8]. We explore this hypothesis by constructing a social network as follows: An edge is created between two users if they gave a similar usage pattern *i.e.,* what kind of topics the users usually like to review. Since COLBERT discovered many latent topics epinions dataset it is possible to build social networks for these topics based on the criteria of similar usage patterns. Thus consider the output from the COLBERT model which consists of a set of keywords that constitute a topic, a set of primary reviewers and a set of secondary reviewers. An obvious way to construct a social network would be to connect the nodes which either review or rate the same topic.

5 Conclusion and Future Work

The availability of multi-relational datasets where there are different types of many-to-many relationships with the entities involved presents a new challenge for developing methods that can detect structure in multi-relational data. Graphical models that address this challenge are being developed. In this paper, we described a graphical model COLBERT which can detect latent topics from document reviewer data. Previous models have focused on capturing the primary structure or interaction between different entities related to a corpus e.g., authors in the case of Author Topic Model or author and recipients in the case of the ART Model. There are other models which also take into account the community of the actors in the social network. In the COLBERT model we also consider the second order interaction or structure in the data in the form of scored reviews of reviews. The model is based on the idea that people of common interests will write reviews about similar products which in turn will be reviewed by people who have similar interests.

The scores associated with the secondary reviews could then be used as a marker of how good the reviewed person is regarding a latent topic. We also noted that the results obtained from the COLBERT model can be used in other domains like expert identification and in constructing social networks. Datasets like the Epinions.com have a rich set of modalities and here we have provided preliminary results from the COLBERT model, there are a number of ways in which this model can be extended e.g., one can take into account the social network information while discovering the topics. Comparing the latent communities with the communities based on trust would be an interesting task as it can reveal how 'generic' trust relates to trust in specific domains. Another modality that can be explored is the individual ratings of specific characteristics of products. The temporal dimension of product review datasets can also be explored. Since timestamp information is available for these datasets one can get a 'snapshot' of the dataset at different times, determining the overall trends in document reviews and study the 'corpus Zeitgeist' by building these models at those time snapshots.

References

1. Ahmad, M., Srivastava, J.: An Ant Colony Optimization Approach to Expert Identification in Social Networks. In: First International Workshop on Social Computing, Behavioral Modeling and Prediction Pheonix, Arizona (2008)

2. Balog, K., Azzopardi, L., de Rijke, M.: Formal models for expert finding in enterprise corpora. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 43–50 (2006)
3. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
4. Campbell, C.S., Magio, P.P., Cozzi, A., Dom, B.: Expertise Identification using email communications. In: CIKM 2003, pp. 528–531 (2003)
5. Chemudugunta, C., Smyth, P., Styevers, M.: Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model. In: Proceedings Neural Information Processing Systems, NIPS (2007)
6. Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the Society for Information Science 41(6), 391–407 (1990)
7. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821–7826 (2002)
8. Golbeck, J.: Trust and Nuanced Profile Similarity in Online Social Networks. MINDSWAP Tech Report TR-MS1291 (2007)
9. Griffiths, T., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences (2004)
10. Krulwich, B., Burkey, C.: The ContactFinder agent: Answering bulletin board questions with referrals. In: AAAI 1996 (1996)
11. Mattox, D., Maybury, M., Morey, D.: Enterprise expert and knowledge discovery. Technical report (1999)
12. McCallum, A., Corrada-Emmanuel, A., Wang, X.: Topic and Role Discovery in Social Networks. In: Proceedings International Joint Conference on Artificial Intelligence, IJCAI (2005)
13. Steyvers, M., Smyth, P., Rosen-Zvi, P., Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery. In: Proceedings Knowledge Discovery and Data Mining, KDD (2004)
14. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. 38, 321–330 (2004)
15. Newman, M.: Fast algorithms for detecting community structure. Phys. Rev. E 69, 066133 (2004)
16. Pathak, N., Delong, C., Erickson, K., Banerjee, A.: Social Topic Models for Community Extraction. In: The Second SNAKDD Workshop, August 24-27 (2008)
17. Pothen, A., Simon, H., Liou, K.-P.: Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl. 11, 430–452 (1990)
18. Schwartz, M.F., Wood, D.C.M.: Discovering shared interests using graph analysis. Commnications of the ACM 36(8), 78–89 (1993)
19. Wasserman, S., Faust, K.: Social Networks Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
20. Wang, X., Mohanty, N., McCallum, A.: Group and Topic Discovery from Relations and Their Attributes. In: Proceedings Neural Information Processing Systems, NIPS (2006)
21. Zhou, D., Ji, X., Zha, H., Lee, C.: Giles Topic Evolution and Social Interactions: How Authors Effect Research. In: CIKM 2006 (2006)
22. Li, W., McCallum, A.: Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, June 25-29, pp. 577–584 (2006)
23. Blei, D., McAuliffe, J.: Supervised topic models. In: Advances in Neural Information Processing Systems, vol. 21 (2007)

An Agent-Based Model for Studying Child Maltreatment and Child Maltreatment Prevention

Xiaolin Hu[1] and Richard W. Puddy[2]

[1] Department of Computer Science, Georgia State University, Atlanta, GA 30303
[2] Richard W. Puddy, PhD, MPH - Atlanta, GA 30341

Abstract. This paper presents an agent-based model that simulates the dynamics of child maltreatment and child maltreatment prevention. The developed model follows the principles of complex systems science and explicitly models a community and its families with multi-level factors and interconnections across the social ecology. This makes it possible to experiment how different factors and prevention strategies can affect the rate of child maltreatment. We present the background of this work and give an overview of the agent-based model and show some simulation results.

Keywords: Agent-based modeling, Complexity science, Child maltreatment, Child maltreatment prevention.

1 Introduction

Child maltreatment (CM) is a serious problem in magnitude and burden in the United States and around the world. According to state Child Protective Service (CPS) agencies, about 900,000 children are confirmed as having been maltreated each year in the United States (DHHS, 2004). In a national survey, 14.2% of men and 32.3% of women reported histories of sexual abuse and 22.2% of men and 19.5% of women reported histories of physical abuse (Briere & Elliott, 2003). Total direct (e.g., hospitalization, chronic health problems, social service system) and indirect (e.g., special education, lost productivity to society) costs of CM in the U.S. were estimated at $103.8 billion annually in 2007 (Wang & Holton, 2007). Substantial documentation exists in scientific literature of the association between CM and a broad range of emotional, behavioral, and physical health problems, e.g., depression and anxiety; conduct disorder and antisocial behavior; and diabetes, obesity, and reduced cognitive functioning. A recent CDC commentary in the Journal of the American Medical Association suggests that progress in preventing the nation's worst health problems – such as obesity, diabetes, and heart disease – can be made by investing in programs that promote raising infants and young children in healthy, safe, stable, and nurturing surroundings (Mercy & Saul, 2009). The article suggests that investments in early intervention/prevention programs can counter adverse experiences in childhood, promote optimal development, and reduce disparities in health.

Despite the importance of CM prevention, many of the current methodologies employed to understand and prevent maltreatment have not fully advanced the field to

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 189–198, 2010.

the point of making significant impact at the population level. In addition, funding, safety, and ethical issues prohibit engaging in research of this scope. The work of CM prevention is particularly challenging due to the dynamic and complex nature of this phenomenon. This complexity results from a system containing multi-level (individual, relationship, community, societal) factors across the social ecology, diversity of actors (such as families, schools, government agencies, health care providers) that potentially affect maltreatment, and multiplicity of mechanisms and pathways that are not well studied or well understood. By acknowledging that CM prevention is affected by a dynamic system of interacting variables with continuous feedback loops into the broader system, additional methodologies seem necessary to capture this non-linear and delayed response.

Complex systems science and agent-based modeling offer tremendous promise in this area because they have proven to be a powerful framework for exploring systems with similar characteristics (Hammond, 2009). Complex systems science is built on systems theory and incorporates elements of multi-level analysis, the transdisciplinary approach to prevention science, and social simulation. Some basic characteristics include: order flows from interactions, not from central control; systems are naturally adaptive and creative; the whole is greater than the sum of the parts; and small changes may produce big effects. Complex systems science has been used in other areas of public health to address problems such as heroin, cardiovascular disease, diabetes, mental health, and tobacco. As an example, complexity modeling of obesity has resulted in tools that help uncover the underlying dynamics of the problem, and help assist in identifying which areas will be more amenable to policy intervention and where leverage may best be applied for any particular policy goal (Hammond, 2009). Often this process of simulation modeling reveals critical leverage points that take into account a system's counterintuitive tendencies, unintended consequences, and time delays, therefore opening new avenues for fundamental improvement.

Using a complexity science-informed approach, we developed an agent-based model that simulates the dynamics of child maltreatment and child maltreatment prevention. This model primarily focuses on the community level of the social ecology, but also incorporates both the individual and relationship levels. It allows users to simulate how different factors and prevention strategies at the community level could affect the rate at which child maltreatment occurs. In this paper, we present the background of this work and give an overview of the agent-based model and show some simulation results.

2 The Social Ecology of Child Maltreatment

To understand the benefits of complexity science-informed approach, it is useful to first consider that the field of CM prevention has adopted the social ecological model (SEM) as an organizing conceptual framework for its work (Belsky, 1980; National Research Council , 1993; Dahlberg & Krug, 2002). The SEM is a systems perspective meaning that attention should be directed both to distinguishable parts and interconnections. Factors at the individual level are related to factors at the community and societal levels. Strategies need to be targeted at all four levels of the social ecology (individual, relationship, community, societal) to ultimately impact the rate of CM at a

population level. However, this conceptual understanding has resulted in only small shifts in how CM prevention is studied and how programs operate. It is still typical for programs to only focus on one level or part of the social ecology at a time (Kelly, 2006; Stokols, 1992). Those programs are predominately at the individual and relationship levels (Freisthler, Merritt, & LaScala, 2006).

In a report by Daro, Barringer, and English (2009) recently commissioned for the National Quality Improvement Center on Early Childhood Education (QIC-EC), they report that characteristics of successful child maltreatment primary prevention programs include those that offered a variety of service components across the social ecology. These include child development (via home visits, quality child care), family development (comprehensive health and mental health services, parenting education, nutrition education, health care and referrals, family support), and community building. Approaches of this type are more likely to sustain prevention efforts over time than any single intervention. There has been growing recognition that to truly prevent CM requires the development of the means to address the community and societal level factors underpinning the maltreatment of children (Tomison and Wise, 1999). This suggests the adoption of holistic prevention strategies with a focus on 'whole of community' approaches designed to influence a broad network of relationships and processes within the family and across the wider community. Unfortunately, there are few programs with demonstrated effectiveness at the community or societal levels of the SEM.

Complexity science provides a new set of tools and perspectives that promise to move forward the study of and implementation of practices that truly take a systemic perspective. An expanded lens regarding systems helps investigate the myriad of interconnections that exist between and among each level of the SEM. This would aid in the SEM being more fully implemented through new pathways for practice and research.

3 The Agent-Based Model of Child Maltreatment

Given the foregoing rationale, the community level was selected as the most relevant level of analysis for developing the agent-based model of child maltreatment, hereafter referred to as ABM-CM. The ABM-CM explicitly models a community of agents, each of which corresponds to a family unit in the community. We employ a resource-based conceptual model to simulate the occurrence of child maltreatment. Specifically, each agent is a family unit that includes a parent-child relationship (or caregiver-child relationship, used interchangeably in this paper). Thus the agent model deals with two basic components which are referred to as *parent care* (denoted as P) and *child need* (denoted as N). The *parent care* generally refers to the care that a parent can provide to take care of the child. This includes any care that the parent may obtain from supportive resources, such as the family resource and the community resource, if they exist. The *child need* represents the amount of care that a child requires for proper health and well-being. Intuitively, if an agent's parent care is enough to meet the child need, there is no maltreatment; otherwise, there exists unmet child need and thus the agent exhibits child maltreatment. Fig. 1 shows the major elements and the dynamics of the ABM-CM. Below we describe these elements. Our description focuses on explaining the concepts, while omitting the implementation details and leaving the mathematical equations to the Appendix of this paper.

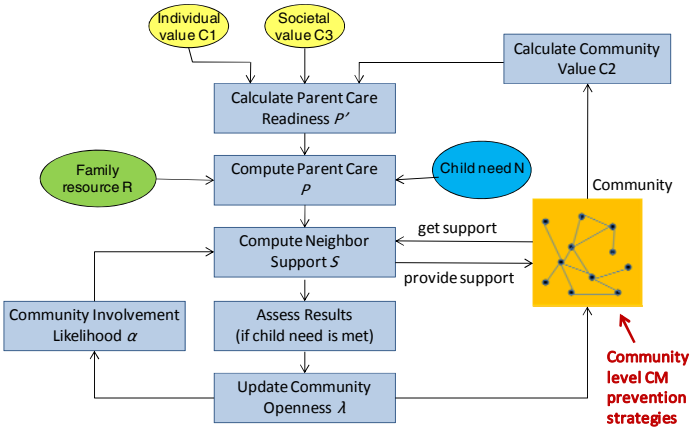


Fig. 1. Elements and dynamics of the agent model

3.1 Child Need

The *child need* represents the amount of care a child requires. To model that different children have different needs, we use a *child need base* to represent the average of an agent's child need. Different agents have different child need base, reflecting that each child's need is different. For example, a child with a developmental disability will have a higher child need base. During the simulation, an agent's actual child need is dynamically generated by oscillating within a range based on its child need base. In the current implementation, an agent's child need base is assigned to the agent in the beginning of the simulation and does not change as the simulation proceeds.

3.2 Parent Care Readiness and Family Resources

Child maltreatment is a complex phenomenon affected by a multiplicity of factors at different system levels (e.g., individual, family, community, and societal) of the social ecology. In the ABM-CM, this social ecological aspect is mainly modeled by how an agent calculates its parent care. To serve this purpose, we introduce a *parent care readiness* (denote as P') into the agent model. The idea is that each agent has a level of readiness for parent care, which depends on the parent's individual characteristics and the community environment and societal context where the agent lives. The parent care readiness is independent of the child need. In other words, this is the level of parent care that an agent would like to provide due to its readiness instead of the child need. As a result, this level of parent care is always provided by the agent, unless it surpasses the available *family resources* (described later) of the agent.

An agent's parent care readiness is affected by three factors at three different system levels: individual, community, and societal. We name these three factors as *individual value* (denoted as *C1*), *community value* (denoted as *C2*), and *societal value* (denoted as *C3*) respectively. The individual value reflects the parent's biological and personal characteristics (e.g., age, education background, history of abuse); the community value represents the wellness of the community as perceived

by the agent; and the societal value reflects the societal context where the agent lives. It is important to note that these "values" are abstractions from the various factors in the real world to serve the modeling purpose. Different agents have different individual values and community values, but they all share the same societal value. An agent's individual value is assigned in the beginning of the simulation and does not change during the simulation. However, this could be extended in the future, for example, to reflect an agent can learn and improve itself over time. Different form the individual value, an agent's community value is dynamically calculated at every time step based on the agent's perception of the community's wellness. Different agents perceive the community differently because of their different social networks. Section 3.3 gives more details about how the community value is calculated. The societal value is defined as a constant and shared by all the agents. An agent's individual value, community value, and societal value all contribute to the agent's parent care readiness. In the current model, the parent care readiness is calculated as a linear function of the three values (see Equation (1) in Appendix).

While the parent care readiness denotes the level of parent care that an agent prefers to provide, the actual parent care is also constrained by the *family resources* the agent has. The family resources are loosely defined as the maximum capacity of a family that an agent can use to provide child care. They are affected by a family's economic and social situations. For example, a parent has more family resources if there are family members living in the house to help the child care; a parent has more resources if the family's economic condition is good and supportive for providing care. In the agent model, each agent has a different capacity (randomly generated in the beginning of the simulation) of family resources and this capacity does not change over time. Family resources alone are not enough to promote positive child development, however. Much depends on how the family translates the available resources into positive child-caring behaviors (Megawangi et. al., 1995). In the ABM-CM, this role is served by the parent care readiness, which indicates how well an agent prepares itself to use the family resources for providing child care. As a result, an agent's actual parent care is defined by its parent care readiness as long as the readiness does not exceed the family resources. When an agent's parent care readiness exceeds its family resources, the agent cannot provide the parent care as it prefers to because the limitation of its family resources. In this case, the actual parent care is constrained by the family resources (see Equation (2) in Appendix).

3.3 Social Network and Community Openness

The ABM-CM views a community not only as the living environment, but also a potential resource for obtaining support in providing child care. A community's social structure is defined by the social network. Each agent is connected to some other "neighbor agents" through its social network. In the current implantation, the social network among agents is a scale-free network. It is generated using the preferential attachment algorithm of Barabasi and Albert (1999) in which the probability a new node links to existing nodes is increasing in the number of links the existing node already has, creating a positive feedback through which popular nodes become even more popular. This network structure can be easily replaced by other types of network that better fits a community in the real world. The social network structure is

initialized in the beginning of the simulation and does not change over time. However, this could be extended in the future to assign weights to the network connections and to allow connections to be dynamically added and removed. A dynamically changing social network structure would make it possible to simulate events that families dynamically establish connections with other families in the process of seeking to build a supportive network for child care.

An agent perceives its community through its social network. Thus the agent's community value C2 is defined by both the community's overall wellness and the agent's perceptions of its neighbor agents, if they exist. Specifically, an agent's community value C2 is calculated as the average of the overall community wellness and all its neighbor agents' parent cares (see Equation (3) in Appendix).

Being part of a community, each agent has a *community openness* attitude (denoted as λ) towards the community. The community openness defines how much the agent would like to be involved in the community. Specifically, the community involvement refers to two types of activities: 1) asking for support from its neighbors if its child need is not satisfied; 2) providing support to other agents if being requested and having remaining family resources. In general, the higher an agent's community openness is, the more likely the agent will be involved in the community; the lower an agent's community openness is, the less likely the agent will be involved in the community activities. In the agent model, this likelihood is specified by a *community involvement probability* (denoted as α), which is calculated based on the agent's community openness (see Equation (4) in Appendix).

As a simulation proceeds, an agent's community openness is dynamically updated based on the agent's experience in the community. In the current model, if the agent asks for support from neighbors and successfully gets enough support to meet its child need, it increases its community openness; otherwise if the agent asks for support but does not get enough support, it decreases its community openness (see Equation (5) in Appendix). The dynamically changing community openness will then influence the agent's likelihood of community involvement.

3.4 The Dynamics – How the Agent Model Works

Having described the elements of the agent model, this section describes how the agent model works. The major activities that an agent goes through in one simulation step are shown in Fig. 1. In the beginning of every simulation step, an agent generates a child need according to its child need base. This child need defines the amount of parent care that needs to be provided in order not to cause child maltreatment. Meanwhile, the agent computes its parent care readiness from its individual value, community value, and societal value. Among them the community value is calculated from the overall community wellness and all the neighbor agents' parent cares. This parent care readiness represents the agent's preferred parent care in this time step. The actual parent care is then calculated by comparing the parent care readiness with the family resource and selects the smaller one between them. After that the agent compares the parent care with the child need to check if the child need is met. If the parent care is less than the child need, the agent has a need to get more support (from its neighbor agents). Otherwise, the child need is satisfied. Even when an agent needs to get more support, the agent may or may not actually ask for support from its

neighbor agents. This is because of the agent's community openness that defines the likelihood of community involvement. In the simulation, the decision of whether the agent asks for support or not is based on the agent's community involvement probability (computed from the agent's community openness). At this stage an agent may also receives request from its neighbors to provide support. When this happens, the agent provides support only when it still has remaining family resource and it is willing to provide the support. Again, the willingness is based on the agent's community involvement probability. Finally, the agent assesses the result to see if its child need is met, and then updates its community openness based on the result as described in Section 3.3.

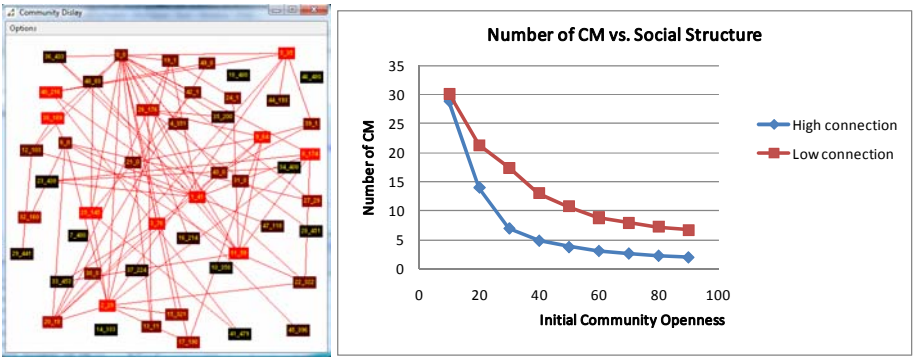
3.5 CM Prevention Modeling

Modeling the impact of CM prevention depends on how the prevention strategies operate on the community or the agents. A preventive strategy can impact a community through multiple pathways, such as *resource*, e.g., adding a community resource center from which families can get support; *knowledge*, e.g., increasing families' awareness of community-based supports through publicity; *belief*, e.g., changing families' community openness attitude through social events; and *social structure*, e.g., building new connections or enforcing existing connections among families. Special types of agents will be developed to model these pathways for simulating the impact of different prevention strategies. For example, to model that a child care center is added in the community, a special agent for the child care center can be created and included in the simulation. This agent is connected to every agent in the community, reflecting that it is accessible to all families in the community.

4 Simulation Results

We carried out a series of simulated experiments (simulations) using the ABM-CM by systematically varying the initial conditions and various model parameters to see how this would affect simulation results. That way, we were able to explore a variety of plausible scenarios that could result, given the assumption we made in setting up the model. The simulations presented in this section aim to demonstrate the major features of the ABM-CM and focus on the qualitative aspect, e.g., to show the trend of CM rate. Because of this, the model parameters and the time step used in the simulations have arbitrary units, intended for exploring how the model system behaves under varying conditions without predicting specific values in the real world.

Our first simulation shows how the social structure of a community can impact the number of child maltreatment. In this experiment, we simulate a community with 50 agents. All agents' family resources are 80; their child need bases are randomly generated between 50 and 70; their individual values are randomly generated between 40 and 80; and their societal values are set to 50. We note that in this setting, agents' family resources are big enough. Nevertheless, because of agents' different child need bases as well as individual values, some agents are not able to meet their child needs. To show the impact of community's social structure, we varied the initial values of agents' community openness (we set all agents to have the same initial value) in



(a) The community (b) Number of CM vs. social structure

Fig. 2. Simulating the impact of a community's social structure

different simulations. During a simulation, depending on an agent's and its neighbor agents' characteristics, the agent's community openness may increase or decrease, which then changes the agent's community involvement likelihood for getting support from or providing support to its neighbors. The simulations show that after a period of transition time, the community is able to "stabilize" and each agent's community openness stays at a certain "stable" level. Fig. 2(a) shows a simulation snapshot with agents shown as boxes, agents' social connections shown as links, and the different levels of community openness shown by different colors. We run the simulations for 2000 steps and measured the number of unmet child needs from step 1000 to 2000 (the first 1000 steps are considered as the transition period). Fig. 2(b) shows the average number unmet child need in every step when the initial values of agents' community openness increase from 10 to 90. We also varied the density of the community's social connections by using two different social networks (still the same set of agents). One social network (denoted as *high connection*) has more connections than the other (denoted as *low connection*). The results in Fig. 2 shows that 1) as agents' community openness increases, the number of unmet child need decreases; 2) the community with more social connections has less child maltreatment (even with the same set of agents).

Our second experiment illustrates the temporal dynamics when a child care center is added in and then removed from a community. Fig. 3 shows the average community value (red), the number of agents with unmet child need (blue), and the community openness of a selected agent (dark cyan) over time. In this experiment, the community initially has no child care center and the percentage of agents having unmet child need is about 60% (30 out of 50 agents). A child care center was added to the community at time 1000. As a result, Fig. 3 shows that the number of agents with unmet child need gradually decreases; the community openness of the selected agent increases; and the overall community value gradually increases too. Because of the support from the child care center, agents increase their community openness and as a result the overall wellness of the community is improved. At time 2000 the child care center was removed and the community went back to the initial situation of having no child care center. Fig. 3 shows that when the child care center was removed the

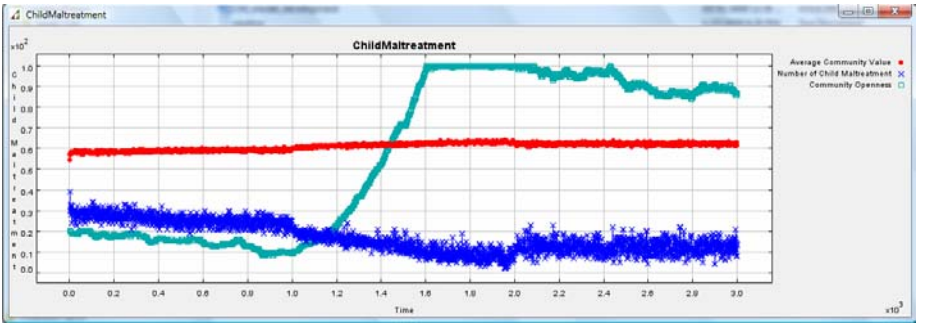


Fig. 3. Simulating the impact of adding and removing a child care center

number of agents with unmet child needs increases (to around 15). However, this number is significantly smaller than that in the beginning stage of the simulation. The selected agent's community openness was also able to maintain at a high level. This experiment shows that when a community's overall wellness is improved, the community can sustain its wellness (because agents support each other) even when some resources in the community are taken away.

5 Conclusion

This paper presents an ABM-CM for studying child maltreatment and child maltreatment prevention and shows some preliminary results. As one of the first efforts in applying a complexity science-informed agent-based modeling approach to the field of child maltreatment prevention, this work builds a starting point from which many future research and developments can be carried out. Some immediate extensions of this work include comprehensive analysis and evaluation of the model, incorporating risk and protective factors of child maltreatment, and aligning the model with real world data and carrying out model validation.

References

Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)

Belsky, J.: Child maltreatment: An ecological model. American Psychologist 35(4), 320–335 (1980)

Briere, J., Elliott, D.M.: Prevalence and psychological sequelae of self-reported childhood physical and sexual abuse in a general population sample of men and women. Child Abuse and Neglect 27, 1205–1222 (2003)

Dahlberg, L.L., Krug, E.G.: Violence-a global public health problem. In: Krug, E., Dahlberg, L.L., Mercy, J.A., Zwi, A.B., Lozano, R. (eds.) World Report on Violence and Health, pp. 1–56. World Health Organization, Geneva (2002)

Daro, D., Barringer, E., English, B.: Key Trends in Prevention: Report for the National Quality Improvement Center on Early Childhood Education (QIC-EC). Chapin Hall at the University of Chicago, Il (2009)

DHHS, Department of Health and Human Services (US), Administration on Children, Youth, and Families. Child maltreatment (2004), http://www.acf.hhs.gov/programs/cb/pubs/cmo4/index.htm (cited August 10, 2006)

Freisthler, B., Merritt, D., LaScala, E.: Understanding the ecology of child maltreatment: A review of the literature and directions for future research. Child Maltreatment 11, 263–280 (2006)

Hammond, R.A.: Complex systems modeling for obesity research. Preventing Chronic Disease: Public Health Research, Practice, and Policy 6, 1–10 (2009)

Kelly, J.G.: Becoming ecological: An expedition into Community Psychology. Oxford University Press, New York (2006)

Megawangi, R., Zeitlin, M.F., Garman, D.: Structural models of family social health theory. In: Zeitlin, F., et al. (eds.) Strengthening the family: implications for international development, pp. 182–237. United Nations University Press, Tokyo (1995)

Mercy, J.A., Saul, J.: Creating a healthier future through early interventions for children. Journal of the American Medical Association 301, 1–3 (2009)

National Research Council, Panel on Research on Child Abuse and Neglect. Understanding child abuse and neglect. National Academy Press, Washington (1993)

Stokols, D.: Establishing and maintaining healthy environments: Toward a social ecology of health promotion. American Psychologist 47, 6–22 (1992)

Tomison, A.M., Wise, S.: Community-based approaches in preventing child maltreatment. NCPCH Issues Paper no. 11, AIFS, Melbourne (1999)

Wang, C.T., Holton, J.: Total estimated cost of child abuse and neglect in the United States. Prevent Child Abuse America, Economic Impact Study (2007)

Appendix: Mathematical Equations Used in the Model

(1) Parent care readiness P': $P' = 0.4\times C1+0.4\times C2+0.2\times C3$

(2) Parent care P is the minimum between parent care readiness P' and family resource R: $P = \min\{P', R\}$

(3) Community value $C2$: $C2 = (P_{ave}+\sum P_i) / (n+1)$, where n is the number of neighbor agents; P_i is the parent care value for neighbor agent i; P_{ave} is the average parent care of the entire community: $P_{ave} = (\sum P_j) / N$, where N is the total number of agents in the entire community.

(4) Community involvement probability α: $\alpha = 0.01\times \lambda$

(5) Update of community openness λ: $\lambda(t) = \lambda(t-1) + 0.5$ if getting enough support to meet child need N; otherwise $\lambda(t) = \lambda(t-1) - 0.5$. The range of λ is between 0 and 100: $\lambda \in [0,100]$

Gryphon: A Hybrid Agent-Based Modeling and Simulation Platform for Infectious Diseases

Bin Yu, Jijun Wang, Michael McGowan,
Ganesh Vaidyanathan, and Kristofer Younger

Quantum Leap Innovations,
3 Innovation Way, Suite 100,
Newark, DE 19711, USA
{byu,jw,mtm,gv,ky}@quantumleap.us

Abstract. In this paper we present Gryphon, a hybrid agent-based stochastic modeling and simulation platform developed for characterizing the geographic spread of infectious diseases and the effects of interventions. We study both local and non-local transmission dynamics of stochastic simulations based on the published parameters and data for SARS. The results suggest that the expected numbers of infections and the timeline of control strategies predicted by our stochastic model are in reasonably good agreement with previous studies. These preliminary results indicate that Gryphon is able to characterize other future infectious diseases and identify endangered regions in advance.

1 Introduction

Various approaches have been developed to understand and predict the spread of infectious diseases and the impact of treatment and control strategies [1]. These range from compartmental models represented by sets of differential equations [2,3] to highly complex individual-based models which represent daily activities and connections of individuals via transmission networks [4]. Compartmental models can be easily solved, but they cannot model adaptive behaviors of individuals and complex interactions of different groups of populations during disease outbreaks. While individual-based models like EpiSims can capture the spread of diseases with high-fidelity, modeling large populations often resorts to supercomputers and makes it impractical for quick what-if analyses of interventions or treatments under different conditions.

In this paper we present Gryphon, a hybrid agent-based stochastic modeling and simulation platform for characterizing the geographic spread of infectious diseases and the effects of various mitigation strategies in a GIS environment. As a flexible, computationally efficient modeling and simulation platform, Gryphon has been used successfully in several real time exercises such as Cobra Gold 2008 in Thailand (bird flu, pandemic flu, and multi-lateral military exercises); Operation Caring Response to aid the humanitarian response to Cyclone Nargis in Myanmar (hurricane and epidemic disease outbreaks); and most recently as a primary tool to assist the U.S. Northern Command (USNORTHCOM) and the United States Department of Health and Human Services (DHHS) in modeling and managing the impact of the spread of the H1N1 virus.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 199–207, 2010.

Gryphon integrates agent-based modeling with a stochastic structured-population susceptible-exposed-infectious-recovered (SP-SEIR) model. This hybrid approach provides several advantages over each pure method by combining rich modeling capabilities of agent-based modeling and low computational overhead of differential (or difference) equations. Therefore, Gryphon enables multiple rapid what-if analyses to be performed using singular or multiple interventions and allows the users to optimize their pandemic responses. Compared to recent efforts on equation-based infectious disease modeling [5,6], Gryphon can support more complex and intensive user interactions for both modeling and interventions at runtime. These include pausing the simulation and modifying parameters for a specific group during runtime, modifying the probability that a sick individual travels from his home city to any other city, and enforcing temporary travel restrictions.

We study both local and non-local transmission dynamics of stochastic simulations based on the published parameters and data for SARS. SARS (Severe acute respiratory syndrome) is a respiratory disease first identified in Guangdong, China and became a severe health threat to more than 30 countries in 2003. Many studies have been reported to study the local temporal development of the SARS epidemic in one or more countries [7,8], but deficiencies were still lying in those models because of their separate space and time methodology and the lack of stochastic process for local and global disease transmission. The goal of this study is to validate the stochastic modeling and simulation of Gryphon for both local and non-local disease transmission dynamics using the SARS epidemic data. Instead of estimating all parameters in the stochastic SEIR model, we focus on the basic reproductive rate R_0, which is "the average number of secondary cases caused by an infectious individual in a totally susceptible population".

The rest of the paper is organized as follows. In Section 2 we present the design of hybrid agent-based modeling, stochastic disease model and data sets used by Gryphon. The results of validation study are presented in Section 3. Section 4 concludes the paper with some directions for future research.

2 Methods

2.1 Hybrid Agent-Based Modeling

A group of individuals associated with a geographic location (e.g., a country) is modeled as a primary group agent. A primary group agent can be decomposed into several secondary group agents. Each of the secondary group agents can be further decomposed into multiple tertiary group agents. Translocation is the process of decomposing each primary group into various secondary groups and populating locations with the corresponding secondary groups. The mixing of secondary groups at a location can be localized mixing or non-localizing mixing. Localized mixing refers to the manner in which members of all secondary groups at a location interact with one another. Non-localized mixing is the manner in which members of secondary groups at different locations indirectly interact with one another or with environments to spread disease such as indirect transmission of cholera via water. In this paper only localized mixing is considered.

Different from equation-based models such as SP-SEIR, the hybrid agent-based model does not have a migration matrix to determine the mixing rates among

different groups. Instead, the mixing process is naturally driven by the behaviors of different groups. The behaviors of an agent include two parts: active and reactive. Active behaviors of an agent are modeled by a set of decision rules such as movement patterns, condition-based behaviors caused by interventions and environmental changes. The reactive behaviors of an agent in the context of infectious diseases refer to localized and non-localized mixing for a location, where the numbers of individuals at different disease states change constantly due to the interaction with other agents at the location.

Each simulation time step consists of three steps in the order of pre-step, step, and post-step. In pre-step, a secondary group agent may change its behaviors in response to either interventions or environmental changes. In step, secondary group agents at a location mix with each other based on a given disease model. In post-step, the system will update the state of each secondary group agent based on the calculation of the disease model. Subsequently, each secondary group agent notifies its primary group agent of the state changes. At the end of post step, all secondary groups at each location are cleared and the translocation process of each primary group agent is executed to prepare for next simulation time step.

2.2 Disease Model

We use a discrete-time stochastic susceptible-exposed-infectious-recovered (SEIR) model to simulate the localized mixing of all secondary group agents at a location, where S(t), E(t), I(t) and R(t) represent the number of susceptible, exposed, infectious, and recovered individuals, respectively, at a location at time t. The total population at the location $N(t) = S(t) + E(t) + I(t) + R(t)$ is assumed to be a constant (birth and death are ignored). Specifically, the stochastic SEIR model is specified by the following difference equations.

$$S(t+h) = S(t) - B(t) \tag{1}$$
$$E(t+h) = E(t) + B(t) - C(t) \tag{2}$$
$$I(t+h) = I(t) + C(t) - D(t) \tag{3}$$
$$R(t+h) = R(t) + D(t) \tag{4}$$

where h represents the time interval between two continuous simulation steps and h is set to 1 day.

$B(t)$ is the estimated total number of infections resulting from individuals in the $I(t)$ state. For a given infectious person, the number of new infections from $N(t)$ is sampled from a binomial distribution as $M(t) = Binomial(Binomial(Poisson(c), S(t)/N(t)), p)$, where c is the mean number of daily contacts per person and p is the probability that a contact produces infection. Given $M(t)$, $B(t)$ can be calculated as the sum of $M(t)$ for all individuals at I state. Since a Poisson distribution is a special case of a Binomial distribution, the compound distribution $Binomial(Binomial(Poisson(c), S(t)/N(t)), p)$ can be reduced to $Poisson(\beta IS(t)/N(t))$, where β is the transmission rate and $\beta = c * p$ [9]. The number of individuals becoming infectious C(t) in a day can be represented by a binomial distribution $Binomial(E, \alpha)$, where $1/\alpha$ is the length of the mean latent period. Similarly, the number of recoveries $D(t)$ can be represented by $Binomial(I, \gamma)$, where $1/\gamma$ is the length of the mean infectious period.

The values of mean latent period and mean infectious period are 0.85 day and 2.95 days, respectively. The daily transmission rate β is estimated from the basic reproductive rate R_0 as $\beta = R_0 * \gamma$.

2.3 Data Sets

Important events in the timeline of the 2003 SARS epidemic in Hong Kong and other Asian countries are as follows

- February 15, 2003: Official report of a 33-year male and a 9 year old son in Hong Kong with Avian influenza (H5N1).
- March 12, 2003: First global alert about atypical pneumonia in Vietnam and Hong Kong was issued by World Health Organization (WHO).
- March 15, 2003: Second global alert about name of SARS and case definition was issued by WHO.

One simple way to model the transmission dynamics and control strategies is to change the basic reproductive rate R_0. Therefore, instead of using one value for R_0, we have a pairwise value (R_H, R_L) for R_0 to reflect the level of effectiveness of control strategies after WHO warnings. The value of R_0 is switched from R_H to R_L in the experiments based upon one of the WHO global alerts. The range of R_H is $\{2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5\}$ and the range of R_L is $\{0.6, 0.7, 0.8\}$.

The travel data sets are generated from the International Air Transport Association (IATA) (http://www.iata.org) database, which contains the number of available sets between any two given countries. The country data sets, including population, latitude and longitude for each country, are generated from the website (http://www.geonames.org).

3 Results

In general, there are two ways to approximate R_0: by assuming an exponential increase in the number of cases over time (e.g., [10]) or by fitting a specific model that summarizes

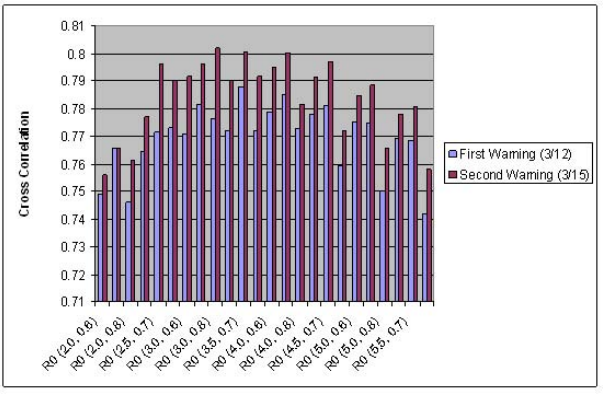


Fig. 1. The cross-correlation coefficient for different pairwise R_0, where the value of R_0 is switched on 3/12 and 3/15, respectively

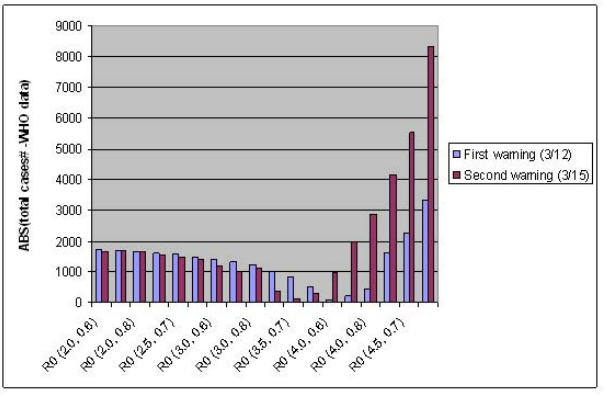


Fig. 2. The difference of simulated accumulative $case\#$ and actual accumulative $case\#$ for pairwise R_0, where the value of R_0 is switched on 3/12 and 3/15, respectively

assumptions about the epidemiology of an infectious disease (e.g., [7]). In this paper we follow the second approach to estimate the spread of a disease based on R_0. Moreover, we choose the metric of cross correlation to quantify the similarity between simulation data and WHO data.

3.1 Parameters for Local Transmission Dynamics

The first experiment studies the basic reproductive rate for local transmission dynamics in Hong Kong. Figure 1 describes the cross-correlation coefficient between the mean of the simulated data for 100 rounds and the WHO data for different pairwise R_0 values, where we seed two infected individuals in each simulation. According to the WHO data, there are only two infections in Hong Kong on February 15, 2003. Note that the two data series in Figure 1 are aligned to calculate the maximal cross-correlation coefficient between the simulated data and the WHO data. We can find that, for Figure 1, the cross-correlation coefficient is consistently higher when R_0 is switched on March 15, 2003. This indicates that those serious control measures such as quarantine and isolation are implemented in Hong Kong only after WHO issued the second global warning on March 15, 2003.

However, it is hard to find the proper value of R_0 only from Figure 1. The reason is that cross-correlation coefficient only models the shape of two data series. As we can see from Figure 1, it is difficult to tell whether pair (3.0, 0.7) is better than pair (3.5, 0.7) on modeling the spread of SARS in Hong Kong. One idea is to use accumulative case numbers as the second measurement to model the scale of the curves. Figure 2 describes the difference of simulated accumulative $case\#$ and actual accumulative $case\#$ for pairwise R_0, where the value of R_0 is switched on 3/12 and 3/15, respectively. From Figure 2, we find that the accumulative case number for (3.5, 0.7) is much closer to the WHO data based on the second global warning. The experimental results show that the combined two metrics, cross-correlation coefficient and cumulative case number, can effectively estimate R_0 values from temporal patterns in an observed epidemic.

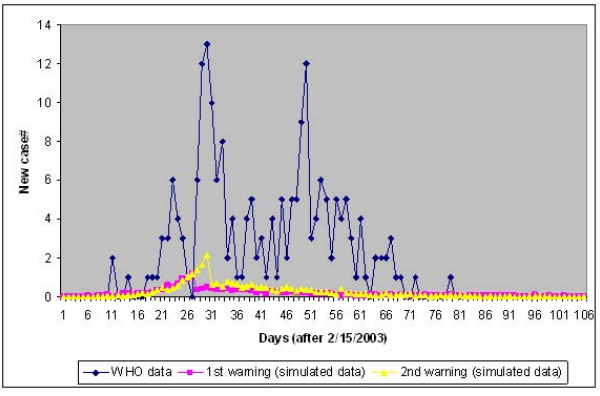


Fig. 3. The new $case\#$ of SARS in Singapore from WHO data and the mean new $case\#$ of SARS from the simulations (100 rounds) for the pairwise R_0 at (3.5, 0.7)

3.2 Parameters for Non-local Transmission Dynamics

The non-local transmission dynamics for SARS at the country level may depend on two factors: the airline travel and the probability that a sick individual travels from his home city to other cities. In this experiment we seed two infections in Hong Kong on February 15, 2003 and we examine the disease outbreak in two Asian countries: Singapore and Japan. We assume that the probability that a sick individual with SARS travels is 0.5. Figure 3 shows the new $case\#$ of SARS in Singapore from the WHO data and the mean new $case\#$ of SARS from simulations. We can see that, based on the airline travel data and the probability that a sick individual travels, the stochastic simulation engine significantly underestimates the SARS outbreak in Singapore. The peak of the mean new case number is one for the first warning and two for the second warning.

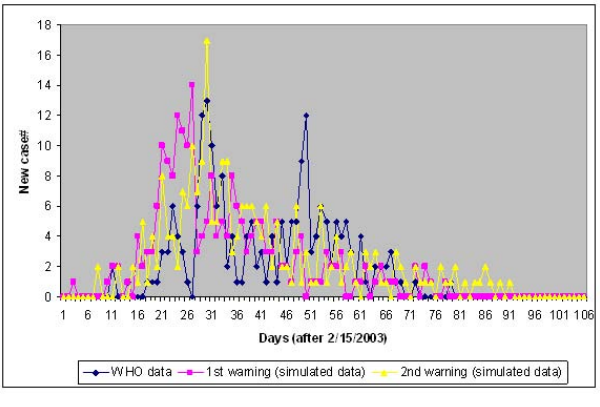


Fig. 4. The new $case\#$ of SARS in Singapore from WHO data and the best match new $case\#$ of SARS from the simulations for the pairwise R_0 at (3.5, 0.7)

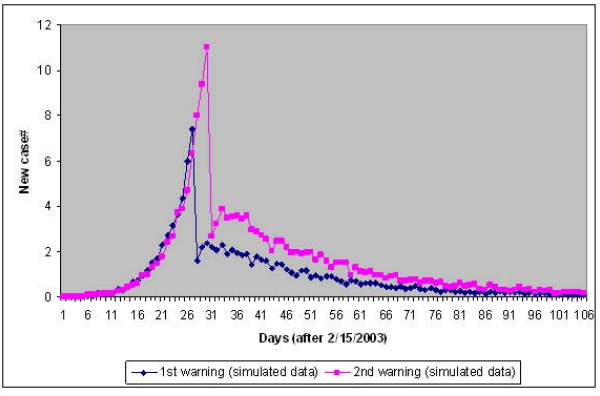


Fig. 5. The mean new $case\#$ of SARS in Japan from the simulations (100 rounds) with the pairwise R_0 at (3.5, 0.7)

However, some specific simulations did capture the dynamics of the SARS outbreak in Singapore. Figure 4 describes the best matched epidemic curve in Singapore from stochastic simulations. The cross correlation coefficient between the best matched simulation data and the WHO data is around 0.6. This indicates that the spread of the SARS epidemic in Singapore is probably the worse case scenario predicted by stochastic simulations. Most likely the first few infections are super-spreaders (a super-spreader is a person having many contacts) and they transmit the disease to large numbers of people.

As shown in Figure 3 and Figure 4, the mean new case number predicted from the stochastic simulation engine can overestimate the SARS epidemic. On the other hand, the stochastic engine can also underestimate the SARS epidemic in some countries like Japan in 2003. The daily airline traffic (in 2005) between Hong Kong and Singapore is 38000 seats/per day, while the daily airline traffic between Hong Kong and Japan is 133000 seats/per day. Based on the traffic data as well as the potential of disease transmission from mainland China, the estimated SARS epidemic should be more severe in Japan than that in Singapore.

Figure 5 shows the mean new $case\#$ of SARS in Japan generated by 100 rounds of simulations. The stochastic simulation engine predicts that there will be about 98 SARS cases for the first warning and 160 SARS cases for the second warning in Japan. This value is close to the mean total case number of SARS predicted by the continuous-time stochastic model published in [6]. However, there is no reported SARS case in Japan during the 2003 SARS outbreak. The actual scenarios in Singapore and Japan motivate us to rethink more complex processes of the spatial and temporal transmission as well as different modes of transmission. These include but not limited to the role of the super-spreader and the heterogeneous mixing among different social groups. For example, Japanese tourists may not well mix with the Chinese community in both Hong Kong and mainland China. The high standard of hygiene conditions in Japan may also prevent the spread of SARS.

4 Conclusion

Gryphon provides a scalable, flexible and interactive disease modeling capability that combines agent-based modeling and mathematical modeling to perform rapid, reasonable fidelity simulations and what-if analyses. We studied the effectiveness of Gryphon, an agent-based stochastic simulation engine for infectious diseases using the historic SARS data. The estimated pairwise value of R_0 for Hong Kong is consistent with [10] by assuming an exponential increase in the number of cases over time, while the predicated total case number for non-local disease transmission is close to the one given in [6], in which Hufnagel et al. used a continuous-time stochastic SEIR model. The experimental results suggest that the expected numbers of infections as well as the timeline of enforced control strategies predicted by our stochastic engine are in reasonably good agreement with previous approaches.

In this validation study we simply use a pairwise R_0 to capture the control strategies deployed upon the first and second WHO warnings. We can see that the peak of the simulated data from Gryphon in Figure 5 drops very fast. This motivates us to develop the next generation of Gryphon technology for data-driven stochastic simulations, where the basic reproductive rate R_0 is dynamically changing based on the available data during a disease outbreak. The data-driven Gryphon will serve as a real-time epidemiological environment for pandemic preparedness and response planning.

Acknowledgements

The authors would like to thank Jay Askren, Albert Boehmler, Julie Cowart, David Hample, Eric Jean and Dr. Steve Prior for their contribution to the system development. We would also like to thank Dr. Amy Kircher at the U.S. Northern Command for her assistance and valuable comments. This research has been sponsored by the Office of Naval Research (ONR) under Contract No. NOOO14-07-C-OO14.

References

1. Anderson, R.M., May, R.M.: Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford (1992)
2. Rvachev, L.A., Longini, I.M.: A mathematical model for the global spread of influenza. Mathematical BioSciences 75, 3–22 (1985)
3. Sattenspiel, L., Simon, C.P.: The spread and persistence of infectious diseases in structured populations. Mathematical BioSciences 90, 341–366 (1988)
4. Eubank, S., Guclu, H., Kumar, V.S.A., Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature 429, 180–184 (2004)
5. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.J., Vespignani, A.: Modeling the worldwide spread of pandemic influenza: Baseline case and containment intervention. PLOS Medicine (2007)
6. Hufnagel, L., Brockmann, D., Geisel, T.: Forecast and control of epidemics in a globalized world. PNAS 101(42), 14124–15129 (2004)
7. Chowell, G., Fenimore, P.W., Castillo-Garsow, M.A., Castillo-Chavez, C.: SARS outbreaks in ontario, hong kong and singapore: the role of diagnosis and isolation as a control mechanism. Journal of Theorectical Biology 224, 1–8 (2003)

8. Chowell, G., Castillo-Chavez, C., Fenimore, P.W., Kribs-Zaleta, C.M., Arriola, L., Hyman, J.M.: Model parameters and outbreak control for SARS. Emerging Infectious Diseases 10(7), 1258–1263 (2004)
9. Feller, W.: An Introduction to Probability Theory and Its Applications, vol. 1. Wiley, Chichester (1968)
10. Wallinga, J., Teunis, P.: Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. American Journal of Epidemiology 160(6), 509–516 (2004)

A Risk Factor Analysis of West Nile Virus: Extraction of Relationships from a Neural-Network Model

Debarchana Ghosh[1] and Rajarshi Guha[2]

[1] Department of Geography, Kent State University, 413 McGilvrey Hall, Kent, OH 44242
dghosh@kent.edu
[2] National Institute of Health Chemical Genomics Center, 9800 Medical Center Drive, Rockville, MD 20852
guhar@mail.nih.gov

Abstract. The West Nile Virus (WNV) is an infectious disease spreading rapidly throughout the United States, causing illness among thousands of birds, animals, and humans. The broad categories of risk factors underlying WNV incidences are: environmental, socioeconomic, built-environment, and existing mosquito abatement policies. Computational neural network (CNN) model was developed to understand the occurrence of WNV infected dead birds because of their ability to capture complex relationships with higher accuracy than linear models. In this paper, we describe a method to interpret a CNN model by considering the final optimized weights. The research was conducted in the Metropolitan area of Minnesota, which had experienced significant outbreaks from 2002 till present.

Keywords: West Nile Virus, risk factors, non-linear, neural network, interpretation.

1 Introduction

West Nile Virus (WNV), first isolated in Uganda in 1937, is a vector-borne infectious disease of global public health concern. The virus is transmitted to humans and other mammals by infected mosquitoes that acquire the virus by feeding on WNV-infected birds [1]. Among humans, infections typically vary from asymptomatic symptoms to mild illness with fever, rash, and headache. During the recent outbreaks in southern Romania (1999), Volga delta of Russia (1999), and Northeastern United States (between 1999 and 2000) there were instances of the more severe form of illness, i.e. West Nile encephalitis (inflammation of the brain), and meningitis (inflammation of the lining of the brain and spinal cord), both of which can be fatal. Since its epicenter in New York in 1999, the virus has spread rapidly in the United States causing seasonal epidemics and illness. The virus is also the largest epidemic of human West Nile neuroinvasive disease to date in North America [2].

It is challenging to understand the spread of WNV because it propagates via complex interrelationships between human, avian, and mosquito habitat systems

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 208–217, 2010.

coupled with risk factors. The broad categories of risk factors underlying WNV incidences are: environmental (temperature, precipitation, vegetation, hydrologic features, parks), socioeconomic (occupation, income, housing age and condition), built environment (catch basins, construction sites, ditches, scrap-tire stockpiles, sewers), and existing mosquito abatement policies. Previous models built to understand the relationships between disease occurrence and potential risk factors assumed *a priori* that there is a linear relationship between these risk factors and WNV incidences. However, it is difficult for linear models to incorporate the complexities of WNV transmission network. As a result, previous investigations were unable to rigorously understand the *non-linear* relationship between risk factors and disease outcome.

We built a computational neural network (CNN) model to capture the *nonlinear* relationship between hypothesized risk factors and WNV disease cases [3]. There are several advantages of using neural network algorithms: 1) ability to capture complex relationships, 2) good predictive capability, 3) does not require to specify *a priori* distribution, 4) no rigid assumptions of normality and homoskedasticity.

Neural network models also have shortcomings. The most important drawback is its *black box* nature or lack of *interpretability*. Along with accurate predictions, interpretability also plays an important role in modeling processes. For instance, it is crucial to understand the relationships between risk factors (predictors) and WNV disease outcome (response) for several reasons. First, some measure of interpretability is needed to provide a sense of confidence regarding the soundness of the model. Second, detailed interpretation would provide evidence to support the use of such models in future scenarios of WNV outbreaks. Third, the extracted correlations between risk factors and WNV occurrence could be useful for vector control policy recommendations. However, interpreting the encoded relationships in a CNN model is difficult, which often forces their use as a purely *predictive tool* rather than an *explanatory tool*. This is in contrast to linear models, which can be interpreted in a simple manner but have poor predictive ability. We observe that in many cases, where the interpretability of a model is a trade-off with predictive accuracy.

In this paper, we address this important research need and describe a method to interpret a CNN model by considering the final optimized weights. Currently, the method is restricted to the interpretation of 3-layer, feed forward, fully connected networks, though extension to more hidden layers is possible. The study is situated in the Twin Cities Metropolitan Area (TCMA) of Minnesota, United States. The virus first reached Minnesota in 2002, creating epidemiological 'hotspots' in the metropolitan area in 2003, 2006, and 2007.

2 Methodology

The proposed CNN interpretation methodology is analogues to the procedure used for interpretation of linear models using partial least squares (PLS) [4]. Hence we start with a summary of PLS method. The predictors for a linear model are used to build a PLS model. The model consists of latent variables, which are linear combinations of the original predictor variables. In the absence of over fitting, the number of latent variables is equal to the number of input variables. The results of the PLS analysis are

summarized by two tables. The first table reports the cumulative variances for each latent variable. Typically the first few latent variables or components explain a large portion of the total variation (80% ~ 90%). As a result, the remaining components are ignored. The second table lists the X-weights for each component. These correspond to the linear combination of coefficients for each input variable in a given component. Therefore analysis of these weights allows one to understand the importance and correlation (direction) of a given input variable to the value predicted by that component.

2.1 Preliminaries

The detailed CNN interpretation method is based on two assumptions. First, the hidden neurons are analogous to the latent variables in a PLS model and second, X-weights in a PLS model are similar to the connection weights in a CNN model. Clearly, these are not one-to-one correspondence because each neuron in the CNN passes through a *sigmoidal* (non-linear) transfer function from one layer to another. By considering the weights connecting the input factors to a specific hidden neuron, we can then interpret how each predictor correlates to the output of that hidden layer neuron. Finally, by defining the contribution of each hidden layer neuron to the output value of the neural network, we can determine which hidden layer neurons are important and which ones can be ignored.

First, we present a brief analysis of how the input values will, in general, relate to the output value. Since, the output value of a CNN for a given set of input value is obtained via a sigmoidal transfer function, the output value, O, is

$$O = \frac{1}{1 + \exp(-X)} \tag{1}$$

where X is the sum of weighted outputs from all the hidden layer neurons. If the output of each hidden layer neuron is denoted by x_j^H, $1 \leq j \leq n_H$, and the weight between each hidden layer neuron and the output neuron as w_j^H, $1 \leq j \leq n_H$, we can write X as $X = \sum_{j=1}^{n_H} w_j^H x_j^H$.

Equation (**1**) can be written as

$$O = \frac{1}{1+\exp\left(-\sum_{j=1}^{n_H} w_j^H x_j^H\right)}$$

$$\frac{1}{O} \sim exp\left(-\sum_{j=1}^{n_H} w_j^H x_j^H\right)$$

$$O \sim exp\left(w_1^H x_1^H + w_2^H x_2^H + \ldots\ldots\ldots + w_{n_H}^H x_{n_H}^H\right) \tag{2}$$

where we drop the constant term because it does not affect the general trend between the output value and the exponential term. From Equation (2), we can say that the components, $w_j^H x_j^H$, have a monotonically increasing function. Also, considering that the output from each hidden neuron will be a positive number, Equation (2) indicates

that if a particular hidden neuron has large weight between itself and the output neuron, then the output from that hidden neuron will dominate the sum. This allows us to *rank* the hidden neurons on the basis of the contribution to the output value. Furthermore the signs of the weights indicate whether the hidden layer neurons will affect the output value positively or negatively [4].

2.2 Combining Weights

The above discussion applies to connections between the hidden layer and output layer. Based on the similar reasoning, it can also be applied to the connection between the input and hidden layers. We denote the weights between the input layer neuron i and the hidden layer neuron j as $w_{ij,}$ where, $1 \leq i \leq n_I, 1 \leq j \leq n_H$, n_I is the number of input layer neurons and n_H is the number of hidden layer neurons. Now let us consider the value of the first predictor for a given observation and how it passes through the layers. As this value passes from the first input neuron to the first hidden layer neuron, w_{11} weight will be formed. The value from the first hidden layer neurons are then passed to the output neuron with the weight of w_1^H. Thus we can say, that, as the value passes from the input layer to the hidden layer and then to the output layer, it is affected by a combined weight of $w_{11}w_1^H$. Similarly, for the same input value passing through the second hidden neuron and then to the output neuron, we can write the associated combined weight as $w_{12}w_2^H$. In general, the combined weight between the i^{th} input neuron and the j^{th} hidden layer neuron will be $w_{ij}w_j^H$. Fig 1 further explains schematically the accumulation of hypothetical weights as the input value flows down the network.

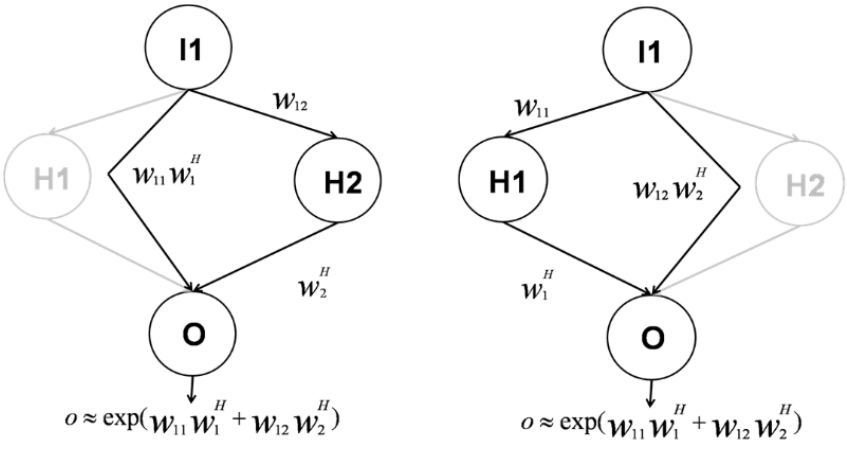


Fig. 1. Schematic Diagram of Combined Weights flowing down the layers in a hypothetical 1-2-1 CNN model for a given observation. The hypothetical 1-2-1 model has one input neuron, 2 hidden layer neurons, and 1 output neuron.

The absolute value and sign of $w_{ij}w_j^H$ is our main interest in terms of interpreting the correlation between the predictors and the response variable. The absolute value

of the weights between the hidden layer neurons and the output neuron might be an indication of which hidden neuron is more effective in terms of contribution to the final output value. Also, the sign of the weight indicates the trend of the output value. For example, if both the weights w_{11} and w_1^H are positive (or negative) we can expect that input values passing down that path will show a positive (or negative) correlation with the output value. If w_{11} and w_1^H are positive and negative respectively, one would expect that the net effect would be a negative correlation between the input values and output values.

2.3 Interpreting Combined Weights

In this section we will consider two possible ways to use the combined weights to interpret the relationships between the predictor variables and the response variable. From the preceding discussion we can represent the weights from a hypothetical 4-3-1 CNN model in a matrix form as shown in Table 1. Here, I1, I2, I3, I4 represents the four input neurons (predictors), $w_{ij,}$ denotes the connection weight between the i^{th} input neuron and the j^{th} hidden neuron, and w_j^H represents the weight between the j^{th} hidden neuron and the output neuron. For this example i ranges from 1 to 4 and j ranges from 1 to 3. The first step in interpreting the weight matrix is to decide the order of the hidden layer neurons in terms of their contributions to the output value. To do so we have followed the ranking technique explained by Guha *et al* [4]. The authors calculated the "squared contribution value" (SCV) for each hidden layer based on the combined connection weights between the input layer and the hidden layer and then between the hidden layer and the output layer. Individual SCV value ranges from 0 to 1 and they sum to 1 for all the hidden layer neurons. SCV values are functionally analogous (but not mathematically equivalent) to the cumulative variance explained by the latent variables in a PLS technique. Therefore, the SCV values clearly indicate the contributions of each hidden neuron and allow us to possibly ignore hidden neurons that have very small values of SCV.

Table 1. Tabular representation of combined weights for a hypothetical 4-3-1 CNN model

	Hidden Neurons		
	1	2	3
I1	$w_{11}w_1^H$	$w_{12}w_2^H$	$w_{13}w_3^H$
I2	$w_{21}w_1^H$	$w_{22}w_2^H$	$w_{23}w_3^H$
I3	$w_{31}w_1^H$	$w_{32}w_2^H$	$w_{33}w_3^H$
I4	$w_{41}w_1^H$	$w_{42}w_2^H$	$w_{43}w_3^H$

For this hypothetical 4-3-1 CNN model, let us assume that the order of importance of the hidden layer neurons is given by H1 > H2 > H3. Thus, the first hidden neuron is the main contributor to the output value. Next we consider the values in the first column (H1) to find out which input neuron is associated with the highest weight. If the value in a given row is higher than the others it implies that the corresponding input neuron (predictor) has contributed more to the hidden layer neurons. Since we

have already ordered the hidden neurons, this means that we can identify the contribution of each input neuron to the output value. Furthermore the sign of the weights will indicate whether high values of that input neuron correspond to low or high values of the output value.

2.4 Cross-Validation

We conducted model cross-validation by two techniques. First, Leave-one-out cross validation or LOO method is used as an internal technique [5]. The Q^2, obtained by using the LOO cross-validation procedure is an alternative to R^2. That is, the neural network model with the same structure (same number of input variables, hidden neurons, and output neuron) is generated using the whole dataset excluding one point. The response value for this point is then predicted using the model and this procedure is repeated for all the points in the dataset. The R^2 for these predictions is denoted by Q^2. Typically if the R^2 and Q^2 are in the range of 10-15 percent, the model is considered good with high predictive ability and generalizability.

Second, to ensure that the proposed CNN interpretation methodology provides valid interpretations, we compared the CNN model interpretation results to that of results obtained from the ordinary least square (OLS) model with same specification.

2.5 Data

The data for the year 2006 with 479 WNV infected dead birds was used to build the CNN model. After checking for colinearity, missing values, and variation, 32 potential risk factors were included in the model and were broadly categorized into four groups. Maximum daily temperature, daily precipitation, and land cover variables were grouped in the environmental category. Built environment factors included density of catch basins, density of ditches, area of parks (open green space), housing density, and housing age. Proximity to features such as lakes, wastewater discharge points, golf courses, trails, shrub swamps, wooded swamp, and bogs were also considered. The variables in the final category of vector control policies were frequency and percentage of public land survey (PLS) units treated for larvicide and adulticide.

3 Results and Interpretation

After training, building, and cross validating the CNN models, a 5-2-1 architecture neural network model was selected to understand the dynamics of WNV in the TCMA. The procedure and criteria involved in choosing the best CNN model is described in details in Ghosh *et. al* [3]. This paper, however, only discusses the interpretation of the selected CNN model. The 5-2-1 CNN model had a RMSE value of 1.78, R^2 of 0.75, and Q^2 of 0.62. The model had five input neurons each corresponding to one of the predictor variable, two hidden neurons, and one output neuron, which stored the predicted number of WNV infected dead birds for a zip code. Out of 32 risk factors, the predictor variables selected for this model were distance to bogs (miles), distance to lakes (miles), daily maximum temperature (F), age of houses (years), and percentage of developed medium density land cover class (Fig 2).

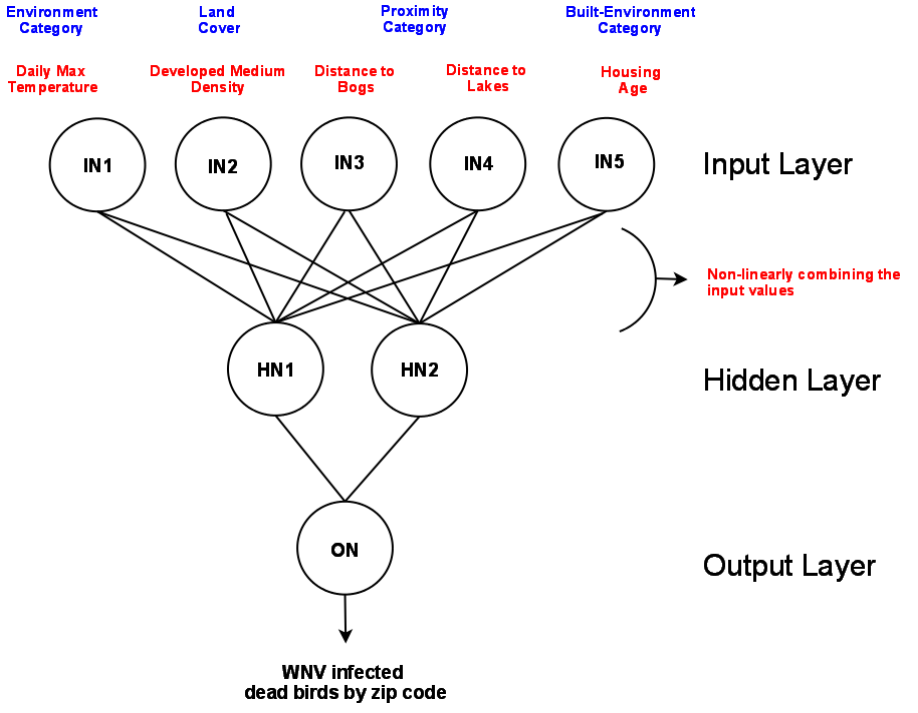


Fig. 2. Structure of West Nile virus analysis model using 3-Layer Feed-Forward Neural Network Algorithm

3.1 Model Interpretation

We first present a discussion of the results obtained from the OLS model with same model specification as that of the CNN model in Table 2. Among the variables which were statistically significant at p-value < 0.05, maximum daily temperature and medium density land cover were positively related to the predicted output. In other words, higher values of maximum temperature could lead to higher numbers of infected dead birds and therefore amplify the rate of WNV transmission. Areas in and around older houses were also at higher risk but this relationship was not statistically significant at the specified p-value. The distance to bogs was negatively related to the occurrence of WNV infected dead birds, i.e. areas near bogs or closer proximity to bogs would increase the risk of WNV infection. The last input factor, distance to lakes was positively related to the number of dead birds and was also not statistically significant. Here the positive correlation was interesting and needed further investigation because previous studies indicated that closer proximity to hydrologic features including lakes and wetlands, increased the risk of WNV infection in birds, mosquitoes, and human [6-7]. The following paragraphs compare these results with those obtained from the interpretation of the CNN model.

Table 2. Summary of the linear regression model developed for the WNV model

Variables	Estimate	Std. Error	*T*	*P*	Sig
(Intercept)	-1.332	0.068	-1.993	0.048	*
Distance to Lakes	0.0004	0.00004	0.914	0.362	
Distance to Bogs	-0.094	0.00005	-1.998	0.041	*
Age of Houses	0.003	0.0012	0.11	0.913	
Dev, Medium Density	0.098	0.0018	5.093	0.000	***
Daily Max Temperature	0.429	0.0006	7.139	0.000	***

Table 3. The combined weight matrix for the 5-2-1 West Nile virus model

	Hidden Neurons	
	1	**2**
Distance to Lakes	-197.34	78.02
Distance to Bogs	-1000.95	-857.77
Age of Houses	732.54	569.11
Dev, Medium Density	380.18	282.27
Daily Max Temperature	1433.95	902.86
SCV	**0.78**	**0.22**

The combined weight matrix obtained from the detailed interpretation of 5-2-1 CNN model is shown in Table 3. The columns correspond to the hidden neurons and were ordered by the SCV values shown in the last row of the table. The SCV values indicated that the first hidden neuron played an important role in explaining the model output of predicted number of infected dead birds. The second hidden neuron played a lesser role. The use of the SCV values in choosing the influential hidden neurons was analogues to the use of percentage of variance explained by the latent variables in a PLS approach. Considering the absolute values of weights in the second column of Table 4, we observed that the most weighted predictors were maximum daily temperature, distance to bogs, and housing age. Maximum daily temperature and housing age had large positive weights, indicating strong positive correlation between these risk factors and disease occurrence in birds. On the other hand, distance to bogs was associated with negative weights i.e., areas near bogs or closer proximity to bogs increased the risk of WNV infection. The other predictor variables, developed medium density land cover and distance to lakes had relatively smaller absolute values of weights with positive and negative signs respectively.

When we considered the second column (second hidden layer neuron), the order of predictor importance based on the absolute values of combined weights were same as that of the first hidden neuron. Except for distance to lakes, all the other predictor variables had same signs (correlation) as before, thus confirming the relationships between these predictor variables and the number of infected dead birds. For the first hidden neuron, the distance to lakes had negative correlation to the WNV infection in birds, i.e. with closer proximity to lakes, the risk of infection among birds increased.

However the relationship became positive under the second hidden neuron. Given the fact that the first neuron had significantly higher contribution to the predicted output than the second hidden neuron, and that the absolute values of weights corresponding to distance to lakes was double in the first hidden neuron (197.34) than the second hidden neuron (78.02), we can confidently say that the correlation between the proximity to lakes and the number of WNV infected dead birds was negative.

In summary, the correlations of the risk factors that played major roles in the predicted value of infected dead birds encoded by the CNN model were similar to that of the linear model. However there were a few differences, which needed further explanation. The main difference was the order of importance of the risk factors. For instance, the linear model indicated that the developed medium density land cover played a very important role in explaining the occurrence of WNV infection in birds, whereas the CNN model accorded it a less significant role. On the other hand, distance to bogs, a wetland type with high potential for mosquito breeding and bird habitats, played a significant role in the CNN model. Similarly, the age of houses had moved up in the order of importance over the land cover variable in the CNN model. These two predictors, characterizing urban features, were both positively related to the WNV infection. Another important difference between the two models was the correlation of distance to lakes. The linear model interpreted the correlation to be positive, whereas the CNN model, as hypothesized, defined the correlation between distance to lakes and the occurrence of WNV infection as negative. These differences are not surprising, since the CNN model correlated the predictor variables or the risk factors in a nonlinear fashion. The point to note here is that the reliability of results obtained from a nonlinear model, capturing the dynamics of a complex health outcome such as WNV, is higher than the results obtained from a linear model.

4 Discussion and Conclusion

This paper has both methodological and applied contributions to the field of health geography and public health. In the following paragraphs, we will first discuss the methodological contributions in terms of interpreting a neural network model and second, the applied contributions will include vector control policy recommendations to prevent the future spread of WNV in the TCMA. The detailed interpretation method, based on weights flowing down from the input layer to the output layer, provided means for understanding the relationships between input risk factors and occurrences of WNV incidences. The methodology is similar to the PLS interpretation method for linear regression model. The analogy to the PLS method is strengthened further when we consider that the hidden layer neurons are analogous to latent variables. The analysis of combined weights and the signs associated with each of the input factors in the weight matrix allowed deciphering the relationships between the predictor variables and the neural network output. Overall, the correlations between the risk factors and disease occurrence were similar to that of the OLS model, however, some of the differences indicated that the neural network model was better suited to capture the nonlinear relationships with greater accuracy than the linear regression model. The proposed method is quite general, as it requires the optimized weights from the network and can be applied to other types of neural network algorithms. This interpretation method

expands the role of CNN models in social and health sciences as both predictive and explanatory tool, thus alleviating the black box nature of the neural network methodology to some extent.

In terms of applied contributions, the findings from this paper could be used for vector control and prevention recommendations, such as, (1) provide sanitation guidelines to individual property owners in the developed medium density area: maintenance of houses, surroundings, and reducing potential mosquito habitats; (2) educational information on the importance of sanitation in the form of videos, and fact sheets distributed at fairs, schools and other public areas can also be effective; (3) regular and planned pest control treatment of older houses, (4) targeting vector control abatement programs, such as larviciding and adulticiding in areas around bogs and lakes.

References

1. CDC: Outbreak of West Nile like viral encephalitis - New York. Centers for Disease Control and Prevention (1999)
2. O'Leary, D., Marfin, A., Montgomery, S., Kipp, A.M., Lehman, J., Biggerstaff, B.: The epidemic of West Nile virus in the United States, 2002. Vector Borne Zoonotic Disease 4, 61–70 (2004)
3. Ghosh, D., Guha, R.: Identifying optimal risk factors for prediction and interpretation of West Nile virus incidences using Genetic Algorithm and Neural Network techniques. Computers, Environment and Urban Systems. Computers, Environment and Urban Systems Under Review (2010)
4. Guha, R., Stanton, D., Jurs, P.: Interpreting Computational Neural Networks QSAR Models: A Detailed Interpretation of the Weights and Biases. Journal of Chemical Information and Modeling 45, 1109–1121 (2005)
5. Golbraikh, A., Tropsha, A.: Beware of Q2. Journal of Molecular Graphics and Modeling 20, 269–276 (2002)
6. Cooke, W.I., Grala, K., Wallis, R.: Avian GIS models signal human risk for West Nile virus in Mississippi. International Journal of Geographic Information Science 5 (2006)
7. Ezenwa, V., Milheim, L., Coffey, M., Godsey, M., King, R., Guptill, S.: Land cover variation and West Nile virus prevalence: patterns, processess, and implications for disease control. Vector Borne Zoonitc Diseases 7, 173–180 (2007)

Coevolution of Epidemics, Social Networks, and Individual Behavior: A Case Study*

Jiangzhuo Chen[1], Achla Marathe[1,2], and Madhav Marathe[1,3]

[1] Network Dynamics and Simulation Science Laboratory,
Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061, USA
[2] Department of Agricultural and Applied Economics, Virginia Tech,
Blacksburg, VA 24061, USA
[3] Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA

Abstract. This research shows how a limited supply of antivirals can be distributed optimally between the hospitals and the market so that the attack rate is minimized and enough revenue is generated to recover the cost of the antivirals. Results using an individual based model find that prevalence elastic demand behavior delays the epidemic and change in the social contact network induced by isolation reduces the peak of the epidemic significantly. A microeconomic analysis methodology combining behavioral economics and agent-based simulation is a major contribution of this work. In this paper we apply this methodology to analyze the fairness of the stockpile distribution, and the response of human behavior to disease prevalence level and its interaction with the market.

Keywords: social network, epidemic, antiviral, behavioral economics, microeconomic analysis.

1 Introduction

A severe epidemic presents a serious threat as it can affect the lives of millions of people and cost billions through morbidity and mortality. Successful pandemic preparedness demands full participation from the private sector and the public sector. Private sector knows which activities and personnel are most critical in their operations. The public sector needs to ensure that all critical infrastructures and resources stay operational and accessible. The roles and responsibilities of the private sector and the government need to be well coordinated and understood in order to construct effective intervention strategies.

* We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by NSF Nets Grant CNS-0626964, NSF HSD Grant SES-0729441, CDC Center of Excellence in Public Health Informatics Grant 2506055-01, NIH-NIGMS MIDAS project 5U01GM070694-05, NIH MIDAS project 2U01GM070694-07, NSF PetaApps Grant OCI-0904844, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113 and NSF NETS CNS-0831633.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 218–227, 2010.

This research is motivated by our recent work as a part of a DHHS requested analysis undertaken by the NIH sponsored MIDAS program [9]. Recent expansion of antiviral drug production capacity and research based evidence of potential benefits of prophylactic antiviral drug has created a renewed interest in assessing the antiviral drug use strategies and potential stockpiling targets by the private and public sector [7]. The strategies for stockpiling antiviral drugs should ensure that sufficient quantities are available to support recommended interventions. The public sector stockpiles are expected to be available mostly for treatment purposes given the limited supply. Responsibility for purchasing antiviral drugs for prophylaxis will primarily be of the private sector and households [18,22]. In the current H1N1 pandemic, a few countries have allowed antiviral drugs to circulate through markets, e.g., in India people can purchase Tamiflu if they have a prescription [1]. In this work we study the conditions for market distribution of antiviral drugs to become desirable.

This research uses an economic perspective to study how disease prevalence, social contact networks and individual behavior co-evolve. A specific scenario is considered: given a limited stockpile of antivirals, is there an optimum antiviral allocation strategy between the market-based availability and public distribution via hospitals that minimizes the attack rate and recovers the cost of antivirals through the market. An economic perspective for the evaluation of public health measures is important as it helps separate the effects of public health policies from those of private decision making.

One of our major contributions is the study of the interaction between the prevalence of the disease, which is decreased by the demand for antivirals, and the demand for antiviral itself, which is increased by the extent of the disease. To our best knowledge this is the first study that uses individual based approach to analyze how behavioral changes occur in response to the growth of the disease and how these changes, in turn, affect the disease dynamics. The degree to which prevalence induces demand for antiviral drugs determines the prevalence elasticity of demand for prevention against the disease. Note that prevalence elastic demand behavior is both good and bad. When the prevalence is high, the demand for prevention makes the disease self limiting. However, when the prevalence is low, the demand for prevention is low which makes it progressively harder to eradicate the disease. In fact, the prevalence elasticity of demand explains the oscillatory behavior of the disease prevalence [14,15,20].

This study aims to isolate the effects of changes in individual behavior caused by the fear of getting infected, and the effects of changes in the social contact network, on the spread of the disease. Our results show that prevalence elastic demand can *delay the onset of the outbreak* and changes in social network caused by isolation measures can *reduce the peak of the epidemic curve* by a significant amount. It also shows that allocating the entire stockpile of the antivirals to just the public sector or just the private sector is a sub-optimal strategy.

In this work we apply a novel microeconomic analysis methodology that combines behavioral economics [21] and agent-based simulation. We model each individual's self intervention in an epidemic as economically motivated behavior

and simulate how the behavior in an aggregated form affects the epidemic dynamics.

The rest of the paper is organized as follows. We describe the infectious disease model, the contact network model, and various interventions including the antiviral market model in §2. Simulation settings are laid out in §3. We present our experimental results in §4 and conclude in §5.

2 Modeling Framework

This study assumes that a "flu-like" disease spreads in a population through person-to-person contact. The disease progression within a host follows the usual SEIR model, explained in §2.1. Person to person contacts form a *social contact network*, explained in §2.2. Interventions to mitigate the epidemic includes antiviral administration that decreases the probabilities of disease transmission, and social distancing measures that decrease person-to-person contact, which eventually decrease the probabilities of disease transmission too. Details of the interventions are described in §2.3.

2.1 SEIR Disease Model

In this study we follow the standard SEIR disease model from the mathematical epidemiology literature [2,16,19]. Each person in the model is in one of the following four health states at any time: *susceptible*, *exposed*, *infectious*, and *removed*. A person is in the susceptible state until he becomes exposed. If person v becomes exposed, he remains exposed for α_v days, called *incubation period*, during which he is not infectious. Then he becomes infectious and remains so for γ_v days, called *infectious period*, during which he may be *symptomatic* or *asymptomatic*. A person in the infectious state will probabilistically transmit the disease to any of his contacts that are in the susceptible state. An asymptomatic person is less likely to transmit the disease to other people than a symptomatic person. After γ_v days the infectious person v becomes removed (or recovered) and remains so permanently. Note that these state transitions are not reversible and are the only possible transitions. The SEIR disease model is explained in more detail in [5].

2.2 Contact Network

A contact network is a directed, edge-weighted graph $G(V, E, w)$. Nodes V correspond to individuals in the population. Edges E represent contacts between individuals during each day and edge weights are the contact durations. There exists edge $(u, v) \in E$ with weight $w(u, v)$ if and only if node u has contact of duration $w(u, v)$ with node v every day.

The probability of disease transmitting from an infectious node u to a susceptible node v is an increasing function of the infectivity of the disease and the contact duration $w(u, v)$. More details about the social contact network model and the disease progression on a contact network can be found in [5].

2.3 Interventions

Interventions include pharmaceutical interventions (PI) and non-pharmaceutical interventions (NPI). PI interventions include administering antivirals, antibiotics and vaccines. NPI interventions refer to actions that effectively change the social network without administering any drugs. This includes various kinds of social distancing, such as school closure and isolation. See the recent federal pandemic influenza plan [8] for discussions on the efficacy of social distancing measures.

At implementation level, a PI only changes the infectivity (if already in infectious state) or vulnerability (if still susceptible) of the intervened person. He maintains his usual daily activities. So the disease may transmit through the contact network via the same edges. But he becomes less likely to get infected and even if he is infected he will be less likely to infect other people. On the contrary, an NPI changes the person-person interactions. For example, if we isolate a household, then each household member only has contacts with other members of the household. Any contact edge between a household member and an outside individual will be removed during the isolation.

Interventions can also be classified as public health level measures and household level self-interventions. The former does not involve a micro-level decision making based on private information such as household income. It includes antiviral administration through hospitals to diagnosed individuals, and school closure. The latter involves household level decisions. We consider two kinds of household level interventions. One, each household can decide the amount of antivirals to demand in the market based on the price for antiviral, its budget, level of prevalence of the disease and its own risk aversion towards the disease. Each household can also decide to isolate all its members at home if one member shows symptoms.

Antiviral Market. In existing works the distribution of antiviral drugs is studied from the public health viewpoint. They are given to people, either targeted or random, chosen by the authorities [18,22]. Who will receive antiviral is decided at the population level. In this work we explore antiviral distribution through the market. In an antiviral market who get the drugs is mainly an individual level decision, although the government can still plays an role, e.g., by putting an upper bound on the antiviral price.

We assume inelastic antiviral supply in the market and elastic demand from each household. The demand for antiviral of household h on day t, $D_{t,h}$, depends on the current epidemic prevalence x_t, which is the fraction of the population in the infectious state on day t, the current market price P_t, and its current budget constraint $B_{t,h}$: $D_{t,h} = f(x_t, P_t, B_{t,h})$. Usually $\frac{\partial f}{\partial x_t} \geq 0, \frac{\partial f}{\partial P_t} \leq 0, \frac{\partial f}{\partial B_{t,h}} \geq 0$.

The demand function can take various functional forms, representing different risk aversion towards the disease. For example, it can be linear in prevalence: $D_{t,h} = \frac{B_{t,h}}{P_t}(\alpha + \frac{x_t}{\beta})$, upper bounded by $\frac{B_{t,h}}{P_t}$; or exponential in prevalence: $D_{t,h} = \frac{B_{t,h}}{P_t}(1 - e^{-\gamma x_t})$. In the latter case the demand for antiviral increases quickly even at low levels of disease prevalence, compared with the former case. Therefore households with exponential demand are more *sensitive* to the epidemic.

Coevolution of Epidemics, Social Networks, and Human Behavior. Figure 1 shows how they interact with each other: (i) Disease dynamics affect individual behavior. Higher disease prevalence and infection occurring to themselves or household members make people voluntarily take social distancing measures and buy antiviral drugs from markets. (ii) Individual behavior changes the social network. Social distancing measures taken by individuals change the connectivity of the social contact network. (iii) Social network changes affect epidemic dynamics. Less connected contact network leads to fewer disease transmissions. (iv) Individual behavior affects the epidemic evolution. Antivirals taken by individuals lower their vulnerability to the disease and their infectvity to other people. This helps in containing the epidemic.

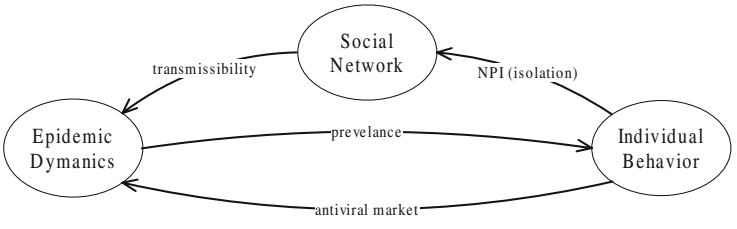


Fig. 1. Interactions among epidemic dynamics, social network, and individual behavior

3 Experimental Setup

In this section we describe our experiment environment and simulation settings. We introduce our simulation tool *EpiFast* in §3.1. Simulation data and parameters are described in §3.2.

3.1 EpiFast: A Fast Simulation for Epidemics

EpiFast is a fast discrete event simulation for disease propagation over a contact network. It uses a parallel algorithm, which enables scaling on distributed memory systems. We have implemented EpiFast in C++/MPI and it is proved to be fast and scalable, and to be able to handle large-scale realistic epidemic simulations. EpiFast uses a disaggregated, agent based model which can represent each interaction between individuals and hence study critical pathways of the diseases. Disaggregated models require neither partitions of the population nor assumptions about large scale regularity of interactions. See [13,10,11,12,17] for results and discussions on this topic. For the details of the underlying algorithm, parallel implementation, and performance of EpiFast, see [5].

3.2 Simulation Configurations

The study is performed on a synthetic population of New River Valley, Virginia. The synthetic population is a set of synthetic people and households, located

geographically, each associated with demographic variables recorded in the US census. Joint demographic distributions are reconstructed from the marginal distributions available in typical census data using an iterative proportional fitting technique [4]. The process guarantees that a census of our synthetic population is statistically indistinguishable from the original US census data [3,6,10].

The New River Valley population has about 150,000 people. We assume that a fixed amount, 15,000 courses, of antiviral drugs are to be allocated between hospitals and markets.

When an individual becomes infectious, with probability 2/3 he appears symptomatic and goes to hospital; otherwise he is asymptomatic. In the latter case he is 47% less likely to transmit the disease. At the hospital, the symptomatic individual is diagnosed with probability 60% and given one course of the antiviral, until the hospital allocation of antivirals runs out. There is no cost for receiving antivirals from the hospital. An individual who is not infected may be mis-diagnosed as infected. For that, we assume one such mis-diagnosis in every six people diagnosed as infected. Similar assumptions have been made in the DHHS sponsored study [9].

Each household, each day, observes the level of epidemic prevalence and the market price of the antiviral, and determines demand for the antivirals. If a household has just made a purchase and is still using the purchased antiviral, it has no demand. The price of the antivirals depends on the remaining amount of supply: $P_t = P_{\max} - (P_{\max} - P_{\min})\frac{Q_t}{Q_0}$, where $P_{\max}$ and $P_{\min}$ are the upper and lower bounds set by the government, and Q_t is the antiviral supply on day t. We set $P_{\max} = \$150, P_{\min} = \$50$. The budget for antiviral purchase is 1% of each household's income. In this study we use the exponential demand function and set $\gamma = 20$.

The market clearing mechanism works as follows. If a household has a positive demand for antiviral, it buys one course for each member, unless the market supply exhausts, or its budget is insufficient to buy one course per member. It may buy antivirals again after it finishes using what it had purchased, until its budget for the antivirals is used up.

Each course of antiviral is applied for 10 continuous days. When antiviral is applied to an infectious person, his probability of transmitting the disease is reduced by 80%; when applied to a susceptible person, his probability of being infected is reduced by 87%.

Isolation measure is taken by households depending on their members' health states as well as the overall level of the disease prevalence. We assume a global threshold for isolation. Isolation does not occur until the disease prevalence reaches 0.002. After the threshold is met, when a member of a household is diagnosed as infected, with compliance rate 40%, all members of this household isolate themselves at home, until the diagnosed member is recovered.

In this study, we try different levels of antiviral allocation between the public sector (hospitals) and the private sector (markets), starting from nothing allocated to hospitals, to all to hospitals, in 1K increments. The simulations are run on a cluster using 10 processors.

4 Experimental Results and Discussion

We are interested in the problem of optimal allocation, of a limited stockpile of antivirals, between the hospitals and the market, with the goal of controlling the epidemic. The observations are discussed in §4.1. The *coevolution* problem is studied by observing the progression of the epidemic when the household demand function is dependent or independent of the epidemic prevalence, as well as when changes to the social network occur. The results are discussed in §4.2.

4.1 Optimal Antiviral Allocation

The efficiency of the antiviral stockpile allocation between the hospitals and the market, in terms of decreasing the attack rate, is shown in Fig. 2. The attack rate decreases as more of the antiviral stockpile is allocated to the hospitals. This is because the antivirals allocated to the hospitals are targeted to those who are infectious, while the market stockpile is distributed to people according to the household income rather than their health states.

It may not be optimal to allocate all antiviral stockpile to hospitals, however. The simulation results indicate that when 40% of the stockpile is assigned to the hospitals, the attack rate reaches its minimum at 5%. After that, there is no significant decrease in attack rate from larger hospital allocation.

The number of antivirals distributed by the hospitals is approximately a constant fraction of the number of infected people. This is because only a fraction of the infected are symptomatic and of those, only a fraction go to the hospital and of those, only a fraction are correctly diagnosed. Therefore, there is a lower bound on the attack rate below which the attack rate cannot be reduced through the hospital allocation. In our simulation experiments, this lower bound is 5%. Further allocation of antivirals to the hospitals does not help reduce the attack rate because they do not get used.

In light of this knowledge, it will be more efficient to allocate part of the antiviral stockpile to the market to help recover its production cost. In fact, our simulations show that if the antiviral unit price varies between \$50 and \$150, the 40-60 allocation (40% to the hospitals and 60% to the market) scheme will help recover about \$629k, see Fig. 3. This implies that if the actual cost of the antivirals is \$42 or less, total revenue will be enough to break even.

4.2 Coevolution

We are particularly interested in understanding the interactions between the epidemics, human behavior and social networks. The following experiment is performed to capture this interaction. Two sets of scenarios are simulated. In the first set, two types of demand is considered. One in which the demand for antivirals is independent of the level of disease prevalence; and in the other, demand is an exponential function of the prevalence. Other parameters and assumptions are the same as in §3.2 in both sets. This set is expected to shed light on how human behavior and the epidemics interact. In the second set,

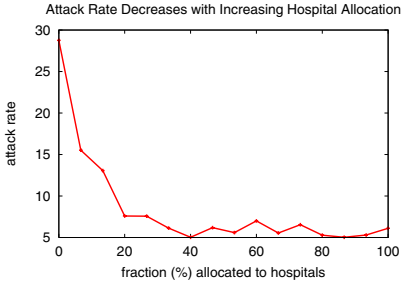


Fig. 2. Targeted antiviral intervention is more effective

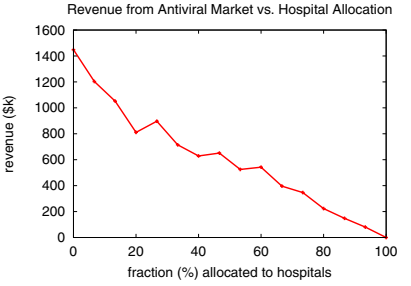


Fig. 3. Recover cost of antiviral production from the market

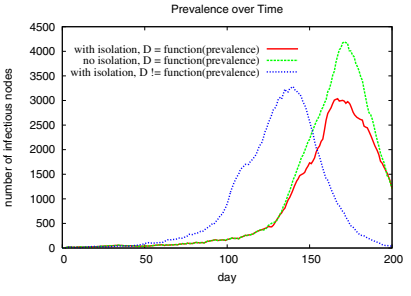


Fig. 4. Comparison of prevalence

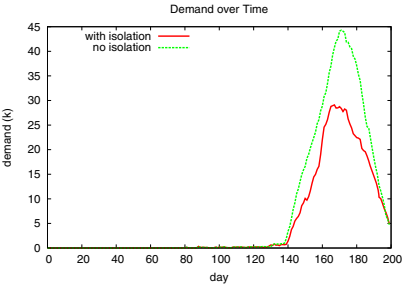


Fig. 5. Comparison of demand

two types of household isolation strategies are allowed. In one, the household members isolate themselves when a member is diagnosed as infected and in the other the household members do not isolate themselves. In both cases all other parameters and assumptions are kept the same. In both cases the demand is an exponential function of the prevalence. This set is expected to capture the effect of social networks on the epidemics.

Now, we compare the following three cases: (i) isolation is on and the demand function is exponential in prevalence; (ii) isolation is off and demand is exponential in prevalence; (iii) isolation is on and demand is independent of the prevalence. In case (iii), each day each household with probability 0.01 has a positive demand for antivirals and otherwise has a zero demand. If it has a positive demand, the demand is $D_{t,h} = \frac{B_{t,h}}{P_t}$.

Figure 4 compares the day-by-day prevalence under these three cases. It suggests that the dependence of demand on prevalence postpones the epidemic peak by about a month; and that isolation decreases the peak infections by more than 1000. In Fig. 5, we compare the day-by-day demand for antivirals under different cases. We omit case (iii) in the plot since the curve will be a flat line close to the x-axis. The figure suggests that antiviral demand increases by about 15K in the absence of isolation, which is mainly caused by higher prevalence.

It is also interesting to see which households purchase antiviral from the market. We sort the households by per capita income and draw the percentile graph. In addition, we plot the amount of antiviral purchase, and the number of infected in these households in Fig. 6. Surprisingly, with exponential demand function, only households with income of 70 percentile or more buy antivirals, and the purchase is dominated by the richest percentile households. Households of lower income do not buy because early in the epidemic the prevalence is low; later in the epidemic when the prevalence increases, the price increases as well, making it unaffordable for the people with lower income. In the case where the demand is independent of prevalence, the antiviral purchase seems to be distributed more evenly among households of different income levels, see Fig. 7.

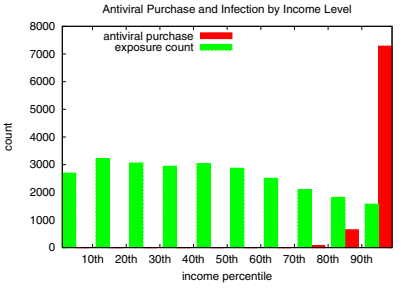


Fig. 6. Distribution of antivirals purchased and exposure counts by household income level. Demand is prevalence elastic.

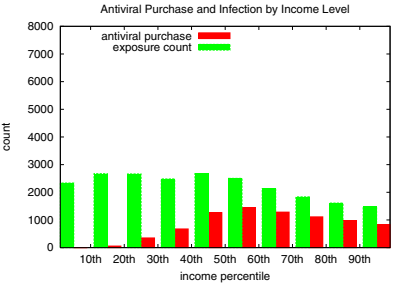


Fig. 7. Distribution of antivirals purchased and infections by household income level. Demand is not prevalence elastic.

5 Conclusions

We study the problem of allocating a limited supply of antiviral drugs between the private sector and the public sector to minimize an epidemic. We find that while the public sector should be given priority, it can be beneficial to distribute part of the stockpile through an antiviral market. This may help recover the cost of antiviral manufacturing, and achieve the overall goal of containing the epidemic at the same time. Our simulation results also suggest significant interaction and coevolution of epidemics, social networks, and human behavior.

References

1. Centre approves restricted retail sale of Tamiflu (2009), http://www.livemint.com/2009/08/28001825/Centre-approves-restricted-ret.html
2. Bailey, N.T.: The mathematical theory of infectious diseases and its applications. Hafner Press, New York (1975)

3. Barrett, C., Bisset, K., Leidig, J., Marathe, A., Marathe, M.: Estimating the impact of public and private strategies for controlling an epidemic: A multi-agent approach. In: Proceedings of the 21st IAAI Conference (2009)
4. Beckman, R.J., Baggerly, K.A., McKay, M.D.: Creating synthetic baseline populations. Transportation Research, Part A: Policy and Practice 30, 415–429 (1996)
5. Bisset, K., Chen, J., Feng, X., Kumar, V.A., Marathe, M.: EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of the 23rd International Conference on Supercomputing (ICS), pp. 430–439 (2009)
6. Bisset, K., Marathe, M.: A cyber environment to support pandemic planning and response. DOE SciDAC Review Magazine (13) (Summer 2009)
7. Dept. of Health and Huamn Services. Guidance on antiviral drug use during an influenza pandemic (2009), http://www.flu.gov/individualfamily/vaccination/antiviral_use.pdf (accessed on November 6, 2009)
8. Dept. of Health and Huamn Services. HHS pandemic influenza plan (2007)
9. Epstein, J., Eubank, S., Lipsitch, M., Hammond, R., Bergstrom, C., Goldstein, E., Marathe, A., Raifman, M., Lewis, B.: Modeling of distribution alternatives of home antiviral drug stockpiling. In: NIH MIDAS Meeting (June 17, 2008)
10. Eubank, S., Guclu, H., Kumar, V.A., Marathe, M., Srinivasan, A., Toroczkai, Z., Wang, N.: Modeling disease outbreaks in realistic urban social networks. Nature 429, 180–184 (2004)
11. Ferguson, N.L., Cummings, D.A.T., Cauchemez, S., et al.: Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437, 209–214 (2005)
12. Ferguson, N.L., Cummings, D.A.T., Fraser, C., Cajka, J.C., Cooley, P.C., Burke, D.S.: Strategies for mitigating an influenza pandemic. Nature 442, 448–452 (2006)
13. Germann, T., Kadau, K., Longini Jr, I.M., Macken, C.A.: Mitigation strategies for pandemic influenza in the United States. PNAS 103(15), 5935–5940 (2006)
14. Gersovitz, M., Hammer, J.S.: Infectious diseases, public policy, and the marriage of economics and epidemiology. The World Bank Research Observer 18(2), 129–157 (2003)
15. Kremer, M.: Integrating behavioral choice into epidemiological models of the AIDS epidemic. The Quarterly Journal Of Economics 111(2), 549–573 (1996)
16. Kuznetsov, Y., Piccardi, C.: Bifurcation analysis of periodic SEIR and SIR epidemic models. Journal of Mathematical Biology 32, 109–121 (1994)
17. Meyers, L., Newman, M., Martin, M., Schrag, S.: Applying network theory to epidemics: Control measures for outbreaks of mycoplasma pneumonia. Emerging Infectious Diseases 9, 204–210 (2003)
18. Monto, A., Pichichero, M., Blanckenberg, S., et al.: Zanamivir prophylaxis: an effective strategy for the prevention of influenza types A and B within households. J. Infect Dis. 186, 1582–1588 (2002)
19. Murray, J.D.: Mathematical Biology: I. An Introduction, 3rd edn. Springer, Heidelberg (2007)
20. Philipson, T.: Economic epidemiology and infectious diseases. In: Culyer, A.J., Newhouse, J.P. (eds.) Handbook of Health Economics, vol. 1, ch.33, pp. 1761–1799. Elsevier, Amsterdam (2000)
21. Rabin, M.: A perspective on psychology and economics. European Economic Review. 46, 657–685 (2002)
22. Welliver, R., Monto, A., Carewicz, O., et al.: Effectiveness of oseltamivir in preventing influenza in household contacts: a randomized controlled trial. JAMA 285, 748–774 (2001)

User Generated Content Consumption and Social Networking in Knowledge-Sharing OSNs

Jake T. Lussier, Troy Raeder, and Nitesh V. Chawla

Interdisciplinary Center for Network Science and Applications,
Department of Computer Science and Engineering, University of Notre Dame, USA

Abstract. Knowledge-sharing online social networks are becoming increasingly pervasive and popular. While the user-to-user interactions in these networks have received substantial attention, the consumption of user generated content has not been studied extensively. In this work, we use data gathered from digg.com to present novel findings and draw important sociological conclusions regarding the intimate relationship between consumption and social networking. We first demonstrate that individuals' consumption habits influence their friend networks, consistent with the concept of *homophily*. We then show that one's social network can also influence the consumption of a submission through the activation of an extended friend network. Finally, we investigate the level of reciprocity, or balance, in the network and uncover relationships that are significantly less balanced than expected.

1 Introduction

The recent emergence of online social networks (OSNs) has affected the manner in which web content is both created and used. In some cases, web sites have created entirely new *classes* of virtual content from social networks on Facebook and dynamic career profiles on LinkedIn, all the way to separate virtual worlds (Second Life). Regardless of the specific application, nearly all OSNs allow users the opportunity to create and consume content. Given the incredibly diverse range of existing OSNs, this content, commonly referred to as User Generated Content (UGC), may be the only common thread in all these networks.

As such, it is useful to characterize OSNs by the role that the UGC plays. One useful distinction, as described by Guo et al. [9], is whether an OSN is networking oriented or knowledge-sharing oriented. Networking oriented OSNs are those in which the formation and sustenance of social links are the primary concern and the sharing of UGC is only a consequence of this. Some examples of these networks are Facebook, Twitter, MySpace, and LinkedIn. In knowledge-sharing oriented networks, on the other hand, the creation and consumption of UGC is most important, and people only form social ties in order to facilitate these processes. Two examples of these networks in popular culture are Digg and Youtube. It should be noted that there is no hard line between the two classes of networks. For example, a Facebook user may occasionally friend another user simply to share information, or a blogger might friend another blogger because

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 228–237, 2010.

of a "real-life" friendship with no intent to share information. Still, it is the primary purpose and role of an OSN which defines it.

We consider Digg, a knowledge-sharing oriented OSN. While there have been a few papers analyzing Digg content consumption [11,19], these works deal primarily with the characterization of future consumption behavior based on past behavior, touching only tangentially on network aspects. Our work brings networking to the forefront and thus presents a more complete understanding of the relationship between UGC and social networking.

Contributions
The details about the data acquired from the social bookmarking site `digg.com` are given in Section 2. The key contributions of the paper are as follows:

1. Using a statistical measure of distributional divergence, we show evidence of *homophily* in Digg, wherein friends tend to digg stories of similar topics and non-friends' tastes are less similar. This implies that friendship on such online networks is largely a phenomenon of common interests (see Section 3.1).
2. We show that stories achieving especially high levels of consumption do so by activating the submitter's second degree friend network (see Section 3.2).
3. Using a measure of *reciprocity*, we show that UGC consumption activity can be very *imbalanced*, meaning that A consumes B's content much more readily than B consumes A's content. This implies that there are highly unbalanced dyads, leading to an important distinction between real-world human social networks versus UGC derived social networks (see Section 3.2).

2 Data Specifics

Before delving into consumption and social networking, we first present a brief depiction of `digg.com`, our particular data set, and the social network we constructed. Launched in 2004, `digg.com` was intended to democratize digital media. Digg allows users to discover and share content from anywhere on the web by pasting a URL; indicating whether it is a story, video, or image; and providing a short description. Other users then comment on the content, or simply "digg" (like) or "bury" (dislike) it. Once a submission has earned enough diggs, it becomes "popular" and jumps to the homepage in its category. Stories that are not yet popular are listed in the "upcoming" section. Finally, Digg allows users to add others to their social networks. If user A adds user B to his or her network of friends, A becomes a *fan* of B. This unidirectional link allows the initiator to monitor the other's activity. Specifically, once A selects B as a friend, A can see any stories that B submits or diggs through a special "friend" interface. If B reciprocates and returns A's friendship, then A and B are called *friends*. Since its launch, Digg has grown to over two million users and has prompted the creation and growth of other social networking sites centered on story creation and dispersion.

We performed a single crawl of digg.com, which returned 6,073,456 friend relationships and 564,193 users. We then constructed a network in which users are

nodes and fan relationships are directed links. More specifically, we consider a directed edge from A to B if A has added B to his network of friends. The resulting network has 564,193 nodes, 6,073,456 edges, an average clustering coefficient of 0.075, and an average degree of 16.146. The node-degree distribution has a clear heavy tail: 78% of users have degree less than five, about 0.1% of users (567) have degree $> 1,000$ and only eight users have degree $> 10,000$. Thus, although most users have relatively few connections there are a substantial number of users who are extremely well connected.

3 Social Networks and Consumption

As stated previously, UGC consumption is a primary driving force in knowledge-sharing oriented OSNs. In this section, we will substantiate this claim by illustrating the relationships between consumption and social networking.

3.1 Consumption Patterns and Friend-Making

In order to study the relationship between social networking and consumption, we need a means of quantifying the difference between two individuals' consumption patterns. A Digg story is classified into one of several *containers* (such as Technology or World & Business) that broadly describe subject matter. If a user diggs or comments on a set of stories $\mathcal{S}$, the set $\mathcal{S}$ will form a *distribution* across the various containers. We would say that two users are similar if they comment or digg similar stories, and we can measure this similarity based on the distribution of the stories they digg.

To do this, we employ *Hellinger distance* [5], a measure of distributional divergence that is both *bounded* and *skew insensitive*, meaning that its value does not depend on the number of samples from either of the distributions being compared. The Hellinger distance $d(a, b)$ between two distributions a and b is defined as

$$d(a,b) = \sqrt{\sum_i \left(\sqrt{\frac{a_i}{|a|}} - \sqrt{\frac{b_i}{|b|}} \right)^2}$$

where i, in this case, runs across all possible containers, a_i represents the number of diggs or comments by user a in container i, and $|a|$ is the total number of diggs or comments by user a.

Thus, users who tend to consume similar content have smaller Hellinger distances. Calculating this distance for all pairs of friends and all pairs of non-friends, we can determine whether friends tend to have similar consumption patterns. Friend and non-friend distributions of Hellinger distances are shown in Figure 1. As can be seen, the distribution of distances for friends is shifted far to the left, indicating that consumption patterns may influence friend-making behaviors. This is consistent with the sociological concept of *homophily*: that individuals tend to befriend people similar to themselves.

Fig. 1. Distributions of Hellinger Distances for friends and non-friends

3.2 Controlling the Consumption of UGC

Often, the creators of user-generated content have a vested interest in the consumption of their content: people want their work to be seen. Now while Section 3.1 illustrated how consumption influences friend-making, the social-network structure within Digg, discussed in Section 2, also allows friendship networks to indirectly affect consumption. When a user submits, comments on, or diggs a story, that story becomes visible to his or her fans through a page known as the "friends interface." This gives the user's direct connections the opportunity to read the story and then comment on it or digg it. If a submitter promotes a story well and it receives enough diggs, it is "promoted" to the front page of Digg, where it is prominently displayed to casual visitors of the site.

We now study the impact of this control. Specifically, we study the importance of the submitter's friend network on the promotion of stories. Following typical Digg terminology, we will henceforth refer to stories that have been promoted as "popular" stories and those that have not been promoted as "upcoming" stories. Figure 2 plots the number of diggs in each hour of a story's lifetime for both popular and upcoming stories. It is immediately clear that appearing on the Digg front page makes a substantial difference in the consumption pattern of a story: the upcoming stories receive most of their diggs at the very beginning of their lifetime, and their consumption activity decays monotonically and rapidly (with a slight blip at 24 hours). For popular stories, digg activity increases for the first several hours then slowly decays.

For a simple explanation of this gradual increase, see Figure 3. The top graph shows comment and digg activity for popular stories relative to the times that they were promoted. We see that the time immediately after promotion is by far the busiest time for a story, with a steady, gradual decrease thereafter. The bottom graph plots the age at which stories become popular. We see a wide range of ages, with some stories achieving promotion almost instantly, while others lingered for over two days. However, most stories that achieve popularity do so very quickly: more than half of our popular stories were promoted within

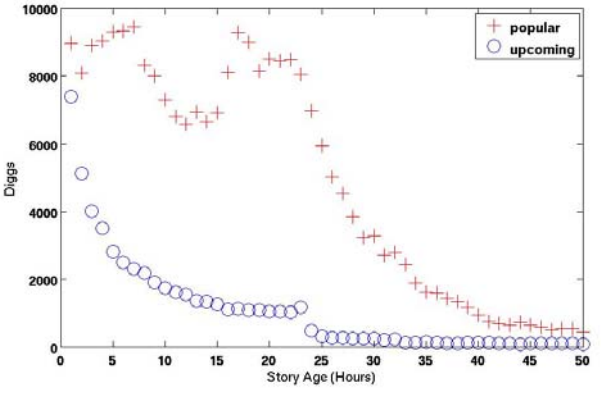


Fig. 2. Digg activity over the lifetime of popular vs. upcoming stories

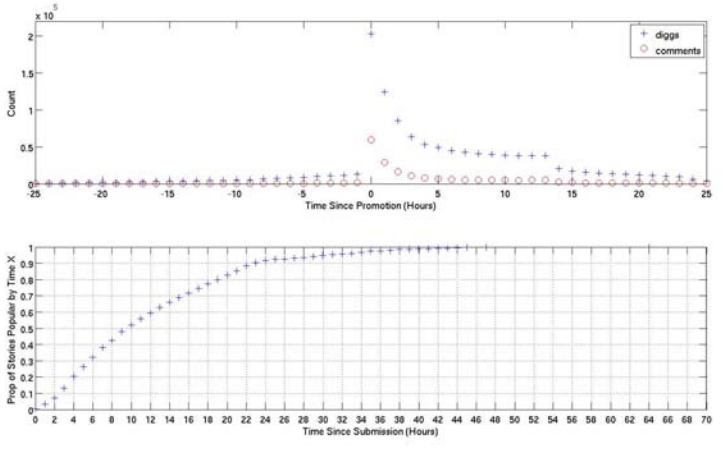


Fig. 3. [Top] Digg activity relative to promotion time. [Bottom] Promotion time.

10 hours of submission, and over 90% are promoted within a day. This range accounts almost exactly for the peak we see in Figure 2.

How do submitters achieve such rapid recognition for their stories? One way would be to post the Digg link on a heavily trafficked web page and rely on its visitors to digg the story. Another possibility is by means of the network mechanisms described, relying on friends and fans to spread the story. The remainder of the section studies the effects of these networks.

The most basic evidence of a network effect on popularity is a difference in consumption patterns before and after promotion. If a person's network plays a role in the promotion of stories, we would expect that a substantial proportion of a story's diggs prior to promotion would come from members of the author's network. After promotion, by contrast, we would expect diggs fairly evenly across

Table 1. Proportion of a story's diggs by shortest-path distance from the submitter

Distance	Pre-Promotion	Post-Promotion
0	0.0115	0.0000
1	0.4611	0.0296
2	0.2812	0.1310
3	0.1579	0.4702
>3	0.0881	0.3690

user population at large. Table 1 shows that this intuition plays out: over 46% of all diggs in the pre-promotion period come from direct friends of the submitter. Of additional interest is the importance of a user's second network. This group (friends-of-friends of the submitter) contributes a larger proportion of the diggs in the pre-promotion period than after promotion, suggesting that diffusion through the friend network contributes to the success of popular stories.

The above result is reasonable, as the friends interface mechanism naturally supports such diffusion. Given a relationship $t \rightarrow u \rightarrow v$ between users t, u, and v, any story that t submits and u diggs will be presented to v through the friends interface. A more complete picture of this phenomenon is given in Figure 4, which shows Digg activity for popular stories as a function of the age of the story (in minutes) for several different shortest-path distances between the submitter and the person digging the story. Here, a very surprising result emerges. In popular stories, the submitter's friend network is activated *almost immediately*. While direct friends of the submitter dominate a story's digg activity for almost two hours after submission, direct-friend diggs reach their peak incredibly quickly ($\approx$ 10 minutes) and then slowly decay as the friend network becomes saturated. By contrast, we see very few diggs early on from users that are three or more steps away from the submitter. These users dominate much later in the story's lifetime as it becomes more popular and more universally accessible.

Activity among second neighbors of the submitter is elevated from very early in the story's lifetime before leveling off, providing additional evidence that the diffusion supported by the friends interface is a significant factor early on in the lifecycle of popular stories. This suggests a modification to the promotion-prediction model of [11], which casts promotion as a function of *interestingness* of the story and *number of fans* of the submitter. It may be more appropriate to consider second-neighborhood size (number of people $\leq$ two levels out) due to this diffusion effect. When friends of the submitter (who generally act quickly) digg a story it becomes visible to their friends. Some fraction of these friends will digg the story, but the effect becomes less noticeable as friends-of-friends hear about the story through other means.

While the preceeding analysis of consumption and social networking dealt with a network of friends, it is also useful to consider a network in which nodes represent users and directed edges represent an individual digging or commenting on another user's submission. Such a network allows us to consider the relationships not only between friends, but more generally between any two users who interact. We calculate a measure of *reciprocity*, or relationship balance, for any

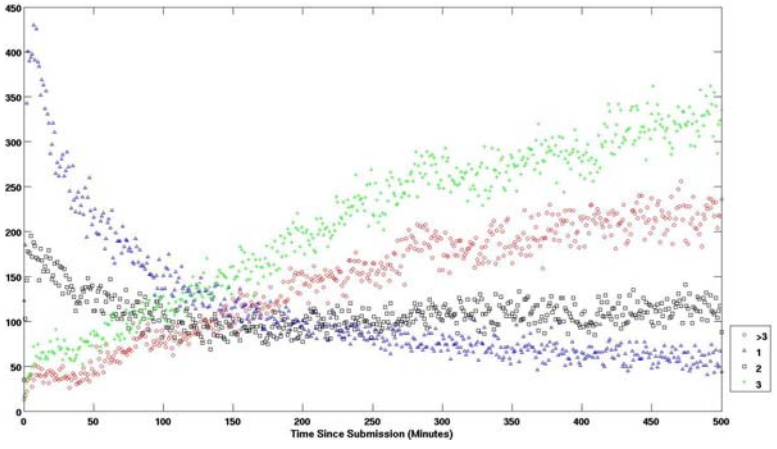


Fig. 4. Digg activity as a function of story age for different shortest-path distance

user pair for which at least one user has submitted at least five diggs or three comments. The general equation for reciprocity for any users a and b divides the number of comments or diggs from a to b by the comments or diggs from b to a. Moreover, we add one to the numerator and denominator as a smoothing factor. The actual equation is given by:

$$reciprocity(a,b) = \frac{consumptions_{a \to b} + 1}{consumptions_{b \to a} + 1} \tag{1}$$

After doing this for all eligible pairs, we then take the logarithm of all the reciprocities so as to transform them into pairs of equal but opposite values. The distribution of these values is then binned into 100 equally sized intervals, as shown in Figure 5.

For both digg and comment distributions, many relationships are either even or mildly uneven, but of particular interest is a small set of users whose relationships are extremely uneven. For example, there are 1,003 dyads (user pairs) in the data for which digg reciprocity is at least 20. That is to say, between two users Alice and Bob, Alice diggs Bob's submissions 20 times more than Bob diggs Alice's submissions. Going out further, we find 456 relationships with reciprocity at least 30 and 302 with reciprocity at least 40.

These heavily imbalanced relationships represent a critical distinction between the (quasi-)social networks developed on UGC sites and the relationships in real-world human social networks or other OSNs. Extreme imbalance, such as 40-to-1 reciprocity, is incredibly uncommon in human social networks [6] as such relationships are generally believed to be unstable (if Bob calls Alice 40 times for every time Alice calls Bob, Bob will tire of maintaining the relationship). On OSNs like Digg and iReport however, these relationships are critical to the intended function of the site, as non-reciprocity helps generate the "buzz" associated with popular articles. If all relationships on Digg were perfectly reciprocal, the only

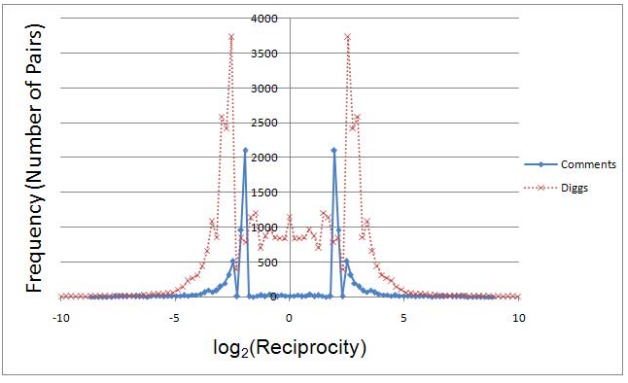


Fig. 5. Distributions of comment and digg reciprocities

determining factor in the success of a story would be the size of the submitter's network.

4 Related Work

Much of the work done in knowledge-sharing oriented OSNs focuses on the formation [7], diffusion [2,8,15,16], and growth [12,14,17] of social networks. Those studies that relate directly to UGC either focus on creation patterns alone or deal only superficially with consumption patterns. Cheng et al. [4] study Youtube and conclude that "related" videos have strong correlations with each other. Leskovec et al. [13] study the diffusion of news across web sites and discover that blogs generally lag mainstream news sites by only a few hours. Guo et al. [9] study UGC creation patterns and find regular temporal patterns and stretched-exponential posting behavior, suggesting that a small set of power users in knowledge-sharing oriented OSNs cannot dominate as they can in a network fitting a power-law. Agrawal et al [3] propose a method for identifying influential contributors to blogs. Certain aspects of the model (number of *in-links* a blog post receives and the number of comments it generates) are directly related to consumption. However, the authors do not study these consumption patterns directly; they merely use them as part of a larger model. Lastly, Guo et al. [10] touch on ideas related to consumption when they examine media access patterns. However, media access is simply the viewing of any form of media available on the web whether it is user generated content or not.

The studies that have examined the consumption of user-generated content focus primarily on characterizing the future consumption patterns of stories based on past consumption. Wu and Huberman [19,20] model the popularity of stories on `digg.com` and find that the number of diggs N_t that a story receives after time t is modeled by a simple multiplicative process. Hogg and Lerman

[11] develop a stochastic modeling framework for user-generated content and use digg.com as an example.

5 Conclusions and Future Work

We studied *consumption* of user-generated content in OSNs in the context of the social bookmarking website digg.com. In contrast to other works, which have focused primarily on characterizing future consumption patterns based on past consumption, we focused more on the interplay between social network formation and content consumption.

In doing so, we showed that similar consumption patterns imply a higher likelihood of friendship. This finding provides evidence of *homophily* (the tendency of people to choose friends similar to themselves) in the Digg network.

In studying the effect of the Digg friendship network on the promotion of popular stories we have two significant findings. First, stories that are successfully promoted to the Digg front page tend to activate the submitter's friend networks very quickly, with friends of the submitter often digging a submission within minutes. Second, we find that second-level neighbors (friends-of-friends) of the submitter also figure prominently in the very early life of a story.

Finally, we studied the level of *reciprocity* or balance in Digg relationships. We found a small number of relationships with exceptionally high levels of imbalance, meaning that person A diggs person B's stories far more frequently than B diggs A's stories. We hypothesized that, while high levels of imbalance are typically unsustainable in human relationships, they are critical in OSNs because people consume much more content than they generate and a heavy-tailed popularity distribution (a small quantity of hugely popular content) is desirable.

Acknowledgments. Our thanks to Kaitlin Clark, Adam Lusch, and Michael Moriarty for providing the friend data for digg.com. This work was supported in part by NSF DHB-0826958 and the Arthur J. Schmitt Foundation. Jake Lussier was also an Ateyeh Undergraduate Scholar.

References

1. Adamic, L., 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, New York (2005)
2. Adar, E., Adamic, L.: Tracking information epidemics in blogspace. In: Proceedings. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 207–214 (2005)
3. Agrawal, N., Liu, H., Tang, L., Yu, P.S.: Identifying Influential Bloggers in a Community. In: Proceedings of WSDM 2008 (2008)
4. Cheng, X., Dale, C., Liu, J.: Statistics and social network of youtube videos. In: 16th International Workshop Quality of Service, 2008. IWQoS 2008, pp. 229–238 (2008)

5. Cieslak, D.A., Chawla, N.V.: Detecting fractures in classifier performance. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp. 123–132. IEEE Computer Society, Washington (2007)
6. Gouldner, A.W.: The norm of reciprocity: A preliminary statement. American sociological review, 161–178 (1960)
7. Gross, R., Acquisti, A.: Information revelation and privacy in online social networks. In: ACM workshop on Privacy in the Electronic Society, pp. 71–80. ACM, New York (2005)
8. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of the 13th international conference on World Wide Web, pp. 491–501. ACM, New York (2004)
9. Guo, L., Tan, E., Chen, S., Zhang, X., Zhao, Y.E.: Analyzing patterns of user content generation in online social networks. In: Proceedings of KDD 2009, pp. 369–378. ACM, New York (2009)
10. Guo, L., Tan, E., Chen, S., Xiao, Z., Zhang, X.: The stretched exponential distribution of internet media access patterns. In: Proceedings of the twenty-seventh ACM symposium on Principles of distributed computing, pp. 283–294. ACM, New York (2008)
11. Hogg, T., Lerman, K.: Stochastic Models of User-Contributory Web Sites. In: AAAI Conference on Weblogs and Social Media (2007)
12. Kumar, R., Novak, J., Tomkins, A.: Structure and evolution of online social networks. In: Proceedings of KDD 2006, pp. 611–617. ACM, New York (2006)
13. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the Dynamics of the News Cycle. In: Proceedings of KDD 2009, pp. 497–506. ACM, New York (2009)
14. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: Proceedings of KDD 2008, pp. 462–470. ACM, New York (2008)
15. Leskovec, J., Adamic, L.-A., Huberman, B.-A.: The Dynamics of Viral Marketing. ACM Transactions on the Web (2007)
16. Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N., Hurst, M.: Cascading behavior in large blog graphs. In: SIAM International Conference on Data Mining, SDM 2007 (2007)
17. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proceedings of the National Academy of Sciences 102(33), 11623–11628 (2005)
18. Quinlan, J.R.: C4. 5: programs for machine learning. Morgan Kaufmann, San Francisco (2003)
19. Wu, F., Huberman, B.A.: Novelty and collective attention. Proceedings of the National Academy of Sciences 104(45), 17599 (2007)
20. Wu, F., Huberman, B.A.: Popularity, novelty and attention. In: ACM Conference on Electronic Commerce (2008)

Where Are the Academic Jobs? Interactive Exploration of Job Advertisements in Geospatial and Topical Space

Angela M. Zoss[1], Michael Conover[2], and Katy Börner[1]

[1] Cyberinfrastructure for Network Science Center, School of Library and Information Science, Indiana University, Bloomington, IN 47405
[2] School of Informatics and Computing, Indiana University, Bloomington, IN 47405
{amzoss,midconov,katy}@indiana.edu

Abstract. This paper details a methodology for capturing, analyzing, and communicating one specific type of real time data: advertisements of currently available academic jobs. The work was inspired by the American Recovery and Reinvestment Act of 2009 (ARRA) [2] that provides approximately $100 billion for education, creating a historic opportunity to create and save hundreds of thousands of jobs. Here, we discuss methodological challenges and practical problems when developing interactive visual interfaces to real time data streams such as job advertisements. Related work is discussed, preliminary solutions are presented, and future work is outlined. The presented approach should be valuable to deal with the enormous volume and complexity of social and behavioral data that evolve continuously in real time, and analyses of them need to be communicated to a broad audience of researchers, practitioners, clients, educators, and interested policymakers, as originally suggested by Hemmings and Wilkinson [1].

Keywords: RSS feeds, data mining, data visualization, science map, visual interfaces, job market.

1 Introduction

According to the U.S. Bureau of Labor Statistics (`http://www.bls.gov`) the U.S. unemployment rate rose to 9.8% in September 2009 from 9.7% in August, 2009. More than 214,000 people lost their jobs within one month. This is the highest unemployment rate since June 1983 when it was 10.1%. Academia, industry, and government are all affected. Many universities cut staff lines, reduced salaries by up to 20%, or have hiring freezes. Students that graduate this year or postdocs that are interested in moving on will face major competition for few jobs. Understanding the job market is an essential element of both informed career choices and scientific policy making.

The work presented here aims to capture and visually communicate exactly what academic job opportunities currently exist. Data from major job advertisement services was captured, processed and analyzed. The geospatial and topical

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 238–247, 2010.

distribution of available jobs is communicated to a broad audience using two different base maps: a map of the world and a map of all sciences. Methodological challenges comprise the robustness of real-time data analysis including stopwording and matching algorithms and the legibility of visualizations. Practical challenges relate to the different temporal distributions and formats of diverse RSS feeds, the automatic identification of geolocations and topics, map labeling, search, and interactivity.

Solutions are expected to be valuable for other projects that aim to use multifaceted, linked visualizations to help understand, model, and predict complex social and behavioral data in real-time. The remainder of the paper is organized as follows: Section 2 discusses related work, section 3 details the datasets used, section 4 presents data preparation and analysis, section 5 introduces the interactive visualizations, and section 6 discusses the strengths and limitations of the presented work together with an outlook to future work.

2 Related Work

The data analysis and visualization work presented here draws inspiration from projects in many fields, including those that deal with real-time data analysis and interactive visualizations and those that focus specifically on job market data.

2.1 Real-Time Data Analysis and Interactive Visualizations

There are few tools and services that support real-time data analysis. Among them is Google Trends [3], which is a service offered by Google that provides longitudinal data about Google searches performed on specified terms and topics. Users enter one or more search terms, and Google Trends produces a report with a plot of the usage frequency over time, information about the geographic distribution of the searches, and related Google News stories. Data can be exported as a CSV file, and some restrictions can be made on the report, i.e., limiting the report to a specific geographic region or time period.

Visualizations such as the Map of the Market by SmartMoney [4] provide up-to-date information on the size and trends of more than 500 stocks using a tree map visualization. The maps are updated every 15 minutes (with a 20 minute delay) based on stock data provided by ComStock Partners, Inc.; historical prices and fundamental data by Hemscott, Inc.; earnings estimates by Zacks Investment Research; and insider trading data provided by the financial division of Thomson Reuters. Stocks are grouped by industry. The size of a rectangle (an individual company) represents its market capitalization. Color gradation depicts the level of losses (bright red is -6 percent) or gains (bright green is +6 percent). Hovering the mouse over a rectangle brings up the company's name and advises whether its stock price is going up or down. Clicking on a rectangle provides more detailed information. Newsmap by Marcos Weskamp takes groupings from the Google News aggregator and displays it as a tree map in real-time [5]. Here, size is

used to indicate the number of articles dealing with a particular topic. Color codes show what larger news category (e.g., business, entertainment) each topic belongs to.

The systems discussed so far visualize small to medium size datasets. However, there is an urgent need to make sense of larger amounts of data to understand their topic coverage and context. For example, the *Science Related Wikipedian Activity* map [6,7] uses a base map of all Wikipedia articles. Overlaid are 3,599 math, 6,474 science, and 3,164 technology relevant articles. Four smaller maps show articles size coded according to article edit activity, number of major edits from January 1st, 2007 to April 6th, 2007, number of bursts in edit activity, and indegree, e.g., the number of times other articles link to an article. These visualizations serve to highlight current trends and predict future editing activity and growth in science, technology, and mathematics related Wikipedia articles. Similarly, *A Topic Map of NIH Grants 2007* shows all 60,000 grants awarded by the National Institutes of Health (NIH) in 2007 [8,9]. It supports search, zoom and pan, color coding, and differential labeling for the different scales. By exploring this map, one can see what topics of research are being heavily pursued, how the topics relate to one another, and what research topics each institute is funding.

2.2 Job Market Data Analysis and Visualizations

There are a number of online sites that visualize employment (or unemployment) data. Among them are the Flowing Data Bleeding Country maps [10], the Slate interactive map of employment data [11], Recovery.gov maps of recovery funding and unemployment [12], indeed.com [13], coolworks.com [14], mapyourjob.com [15], and jobmaps.us [16].

Most sites focus specifically on the geospatial visualization of job data. Indeed.com aggregates jobs from many major job sites and displays them as circles of varying sizes that have been normalized by the population of the location. That is, a circle indicates that, e.g., 49 jobs have been posted for every 1000 people living in the city. The Slate interactive map also used circles of varying sizes but does not normalize by population density. The Bleeding Country and Recovery.gov maps color-code geographic regions (states or counties) by unemployment rate. Other jobs maps, such as those at coolworks.com, mapyourjob.com, and jobmaps.us, use a single flag on the map for each job, making it difficult to evaluate the strength of the job market in a particular location from a distant zoom level. Some sites list very brief snippets of the jobs that are being displayed on the map, but mapyourjob.com in particular includes a detailed table for the jobs listed. Each incorporates some sort of search or filter by topic except for Indeed.com, which is more of a static visualization of quarterly data than a way to browse individual jobs. Thus, many of the available maps incorporate both geospatial data and topical filters. To our knowledge there exists no site that serves topic maps of jobs.

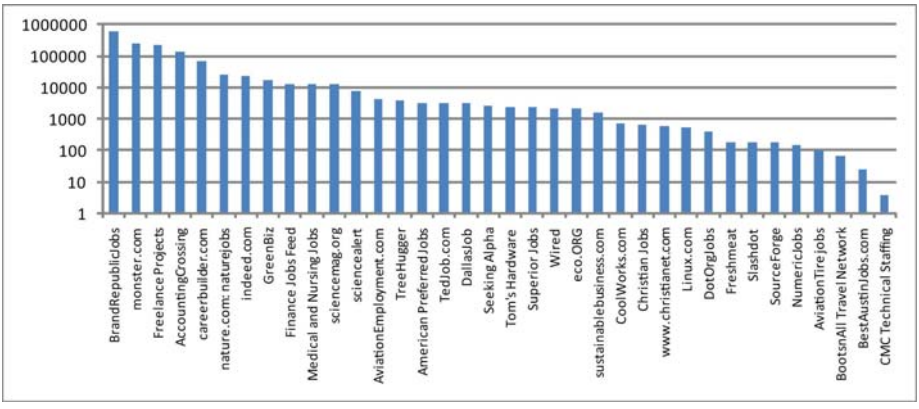


Fig. 1. Number of posts collected from each job site between December 2008 and October 2009. Each job site may have multiple RSS feeds, dividing jobs by topic or geographic area.

3 Data Sets

Since December 2008, we have been collecting job postings from 380 RSS feeds, representing 36 distinct career sites (Figure 1). The feeds were selected from the BestCollegesOnline.com list of the 100 Best RSS Feeds for Recent College Grads [17]. Among these are Monster.com, CareerBuilder, Indeed, and other specialized job sites, many of which publish multiple RSS feeds. Closer examination of the data, however, revealed that there is considerable variety in the type and quantity of text included in the RSS feed items for each site. The feed items are often abridged descriptions of the jobs with links to a more complete job posting. The RSS feed itself may only contain the name of the job and the first sentence or two of the job description. Because of the limitations of using so short of a description for text-based analysis, a more complete sample data set was collected to prototype the system. The sample data set includes over 3,500 full-text, location-specific, time stamped job postings from Nature Jobs [18]. These posts were parsed and stored in a relational PostgreSQL database. The HTML files from these sites have clearly-delineated fields for important information, such as post title, location, employer, etc. These were harvested using screen scraping techniques. Ideally, the data collected would have high coverage (U.S. or world-wide) and high quality of location, topic, and salary data with few missing or unidentifiable values and added flags for those jobs that are funded by ARRA. This level of detail and delineation, however, is not yet available from major job posting sites.

4 Data Preparation and Analysis

For the sample data set, we extracted a timestamp for the posting, the title, the source (company or agency) responsible for the post, the full HTML-formatted

text of the job description, and a URL linking to the site on which the post originally appeared.

In order to geo-locate jobs, the Google Maps geocoding API [19] was used to transform plaintext location strings into rich geographic information, including latitude and longitude.

In order to science-locate jobs, the job descriptions were processed as follows. First, the full text of each job description was stopword-filtered and tokenized into 1-, 2- and 3-grams. These n-grams were then scored for relevance using TFIDF (term frequency-inverse document frequency) term weighting, and these weights were summed to create a total strength of association between the job posting and the node or nodes to which the used keywords belong.

5 Visualization

Many of the sites discussed in the related work section use circle size coding to effectively denote the number of jobs, and we adopted this in our visualizations. Similarly, we adopted zoom, pan, search, and request of detail functionality following Shneiderman's Visual Information-Seeking Mantra [20].

Contrary to the other sites mentioned, our visualization provides two complementary views of the data: a geospatial view and a topical view. The geospatial view helps answer, "Where are the jobs?" The topical view helps answer, "What jobs exist?" Common visual metaphors are leveraged to give the user a sense of consistency. Specifically, because both visualizations are maps, users transfer an understanding of spatial relationships between the two. At the code level, both maps use the same Javascript library and server-side web service such that interaction mechanisms from one visualization are readily available to the other. The primary interaction affordances shared by the two visualizations include the circular markers (or icons) that are size-coded to represent job density, higher resolution (more markers) at lower zoom levels, a common search interface, and identical detail-on-demand behavior. Moreover, both of these visualizations were created with the Google Maps API, and user tests have demonstrated that this familiar interaction framework affords users an immediate understanding of the basic functionality of the interface, allowing them to begin exploring its features more easily.

When users click on an icon on either map, an Information Window pops up to show a list of the jobs that have been associated with that position (that is, with a location for the geographic visualization or with a scientific domain in the Map of Science visualization). When a user clicks on one of the job titles in the Information Window, the secondary window on the right of the web page displays more detailed information about the job.

5.1 Geospatial Visualization

The geospatial visualization behaves much like a traditional Google Map, with individual circle markers representing clusters of posts in a given geospatial

area. Semantic zoom [21] is employed using MarkerManager [22] to ensure equal information density at different zoom levels. For example, at high zoom levels, where much of the U.S. is visible at a given time, fewer markers are displayed, associating jobs only with states and not with individual cities (Figure 2, left).

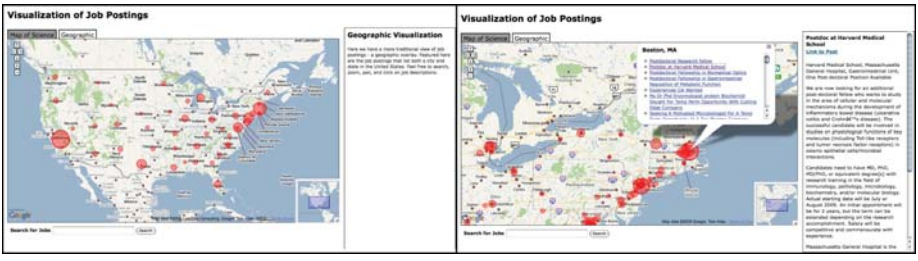


Fig. 2. U.S. level view of job postings on the geographic visualization, clustered by state (left). Lower zoom level of job postings on the geographic visualization – state clusters have broken apart into individual geographic locations at this level (right).

When users zoom in, however, the visualization displays a finer-grained representation of the geographic area, and markers corresponding to individual localities become visible (Figure 2, right).

Further research is warranted to investigate the best scaling technique for the geospatial visualization, whether it is using a density measure (number of jobs per unit population) or scaling linearly or by a power law to approximate real-world populations.

5.2 Map of Science Visualization

The Google Maps API allows users to create a custom map with custom tile sets, thus appropriating the standard pan and zoom actions from the map metaphor to explore other types of images. Here, we have created a custom Google Map with the UCSD Map of Science as a basemap. The UCSD Map of Science is the product of a large study by Klavans and Boyack supported by the University of California San Diego [23, 24, 25]. It uses 7.2 million papers and over 16,000 separate journals, proceedings, and series from the Web of Science by Thomson Scientific and Scopus by Elsevier over the five year period from 2001 to 2005. Bibliographic coupling using both highly cited references and keywords was applied to determine the similarity of journals. Using a hierarchical, multi-step clustering procedure, journals were grouped into 554 clusters, represented by 554 individual nodes in the network. Links denote strong bibliographic coupling relations. In its traditional format, the UCSD map has different sizes for the nodes to indicate the volume of publications from the 2001-2005 data set. Here we only use the structure of the network; see Figure 3.

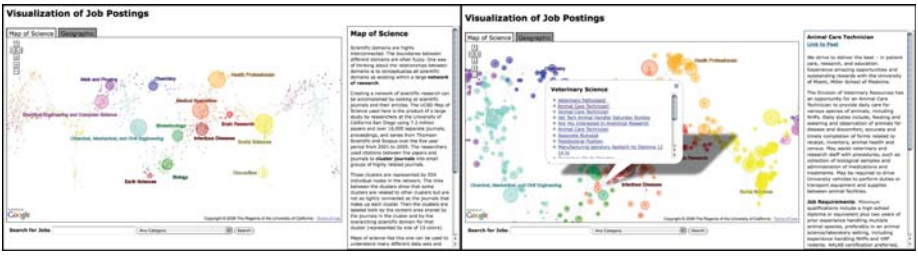


Fig. 3. High (left) and low (right) zoom views of the Map of Science visualization. The map is circular, so areas of the map are repeated side to side as users scroll back and forth. Postings are clustered by the 13 main scientific domains at the high zoom level and the 554 subdisciplines at the lower zoom level.

The clusters are grouped into the 13 overarching scientific domains identified by the analysis (e.g., "Math and Physics", "Humanities"). Each cluster also has its own name (a descriptive name of the subdomain, like "plant physiology") and a set of keywords (an average of about 130 keywords per cluster). Keywords located within a job posting can then be used to "science-locate" a job positing. However, if a posting contains keywords from several Map of Science clusters, that posting will appear in multiple nodes in the visualization. As can be seen in Figure 4, most job posts are associated with multiple nodes, though only a few have more than 15 associations. A larger data set with job posts from a variety of different job sites may produce a different distribution of associations between posts and nodes.

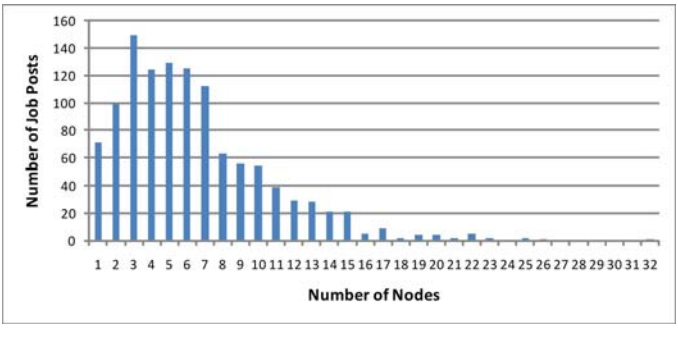


Fig. 4. Number of associations between Nature job posts and nodes. The y-axis is a count of the job posts that are associated with the number of nodes on the x-axis.

The process of creating a Google Map with custom tiles is partially outlined on Google's Map Overlays [26] page. More detailed instructions at Mapki [27] and a Photoshop script for creating custom tiles [28] were also very helpful.

Finally, because the online documentation is a bit sparse and often uses different versions of the Google Maps API, we heavily relied on and modeled after one particular guide by Matthew Muro [29] and several examples, like the previously mentioned NIH map [8] and Google examples like the Tile Detector [30] and the LabeledMarker Marker Hider [31]. The tiles were created from a PostScript file of the base map that includes colored nodes of a uniform size.

6 Discussion and Outlook

The presented work makes several contributions. We have created geospatial and topical visualizations of job opportunities throughout the United States and all sciences. These complementary visualizations use similar visual metaphors to afford the user unique insights into the continuously evolving scientific job market. Moreover, the approach enables decision makers and job seekers to get a high level overview of the relative distribution of employment opportunities in different domains, while at the same time providing a more detailed perspective of data from a prominent employment site that is both easy to use and insightful.

Future work will improve the online service by using higher quality job data sets, improve geolocation and the cleaning of job descriptions for keyword matching, and optimize the visual display of larger amounts of jobs. A significant challenge of this project involves the mapping of job data to nodes in the Map of Science. The Map of Science term data have been automatically extracted from the text in journal publications. Language use in scientific publications naturally differs from language use in job postings. Overly-broad words such as “chemists” or “economy” are uniquely associated with a single cluster in the Map of Science, whereas these words may appear more commonly in job postings. Analyses of the growing job posting data set could help establish baseline information about the differences in language use and suggest modifications to the keyword set that would produce more meaningful matches.

The connections between disciplines and the presence of multiple nodes per discipline may also cause conceptual problems for users. Because the Map of Science was created from bibliometric data and the journals were clustered by computational analysis, disciplines have multiple nodes, nodes have multiple journals and keywords, and one job might map to multiple nodes. On the other hand, some nodes may not have any job data associated with them. Without documentation, the connections between nodes, the size of “empty” nodes, and the differences between nodes of the same discipline may not be clear. The positions and shapes of continents and countries are extensively taught and used in school. Similar training might be required to fully utilize maps of science.

Additionally, the correct encoding mechanism to communicate the density of job postings in the geographic visualization warrants further evaluation. Important considerations include whether to use size-coding or color-coding and whether or not to scale raw values linearly, by a power law, or in relation to the size of a scientific domain or geographic location.

The next phase in the progression of the project is to conduct a usability study to establish standards and explore competencies of navigation for both of

the visualizations. Showing the increase or decrease of jobs over time is a major challenge that will be addressed in the usability study. The final online service will contribute meaningful data and trend analyses to labor market research, especially when grounded by and compared with other data sets from agencies like the Bureau of Labor Statistics.

Acknowledgements

This project received a great deal of support from members of the Cyberinfrastructure for Network Science Center at the School of Library and Information Science at Indiana University. Consultations with Bruce Herr II and Russell J. Duhon were helpful in the conceptualization and implementation of this project.

This work is funded in part by the National Science Foundation under grant IIS-0715303 and the National Institutes of Health under grants RM-07-004 and 1U24RR029822-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Hemmings, J., Wilkinson, J.: What is a public health observatory? Journal of Epidemiology and Community Health 57, 324–326 (2003)
2. The American Recovery and Reinvestment Act of 2009: Saving and Creating Jobs and Reforming Education (2009), http://www.ed.gov/policy/gen/leg/recovery/implementation.html
3. Google Trends, http://www.google.com/trends
4. Map of the Market at SmartMoney.com, http://www.smartmoney.com/map-of-the-market/
5. Newsmap, http://newsmap.jp
6. Herr II, B.W., Holloway, T., Hardy, E.F., Boyack, K.W., Börner, K.: Science Related Wikipedian Activity. 3rd iteration (2007); The power of forecasts, places and spaces: Mapping science edn., vol. 3, Places and Spaces: Mapping Science, Bloomington IN and Albuquerque (2007), http://scimaps.org
7. Math, Science, & Technology Articles in Wikipedia Visualization, http://www.gigapan.org/viewGigapan.php?id=4305
8. Herr II, B.W., Burns, G., Newman, D., Talley, E.: A Topic Map of NIH Grants 2007. 5th iteration (2009); Science maps for science policy makers, places and spaces: Mapping science edn, Places and Spaces: Mapping Science, Bloomington (2007), http://scimaps.org
9. Herr II, B.W., Talley, E.M., Burns, G.A., Newman, D., La Rowe, G.: The nih visual browser: An interactive visualization of biomedical research. In: Proceedings of the 13th International Conference on Information Visualization (IV 2009), Barcelona, Spain, July 14-17, pp. 505–509. IEEE Computer Scoiety, Los Alamitos (2009)
10. Unemployment in the United States, 2004 to Present (2004), http://projects.flowingdata.com/america/unemployment/
11. An interactive map of vanishing employment across the country, http://www.slate.com/id/2216238/

12. Where is the Money Going?, http://www.recovery.gov/transparency/pages/home.aspx
13. Where are the Jobs? – Indeed.com, http://www.indeed.com/jobtrends.jsp
14. Job Map - CoolWorks.com, http://www.coolworks.com/job-map/
15. MapYourJob.com: Find or Post your job on our map!, http://www.mapyourjob.com
16. JobMaps = Indeed Job Search + Google Maps, http://jobmaps.us
17. 100 Best RSS Feeds for Recent College Grads, http://www.bestcollegesonline.com/blog/2008/08/28/100-best-rss-feeds-for-job-seekers/
18. Science Jobs: Scientist Recruitment & Vacancies: Nature Jobs, http://www.nature.com/naturejobs/index.html
19. Services - Google Maps API - Google Code, http://code.google.com/apis/maps/documentation/services.html#Geocoding
20. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the 1996 IEEE Symposium on Visual Languages, pp. 336–343. IEEE Computer Society, Washington (1996)
21. Furnas, G.W., Bederson, B.B.: Space-scale diagrams: Understanding multiscale interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 234–241. ACM Press, Denver (1995)
22. MarkerManager v1.0 Reference, http://gmaps-utility-library-dev.googlecode.com/svn/tags/markermanager/1.1/docs/reference.html
23. Boyack, K.W., Klavans, R.: Map of Scientific Paradigms. 2nd iteration (2006); The power of reference systems, places and spaces: Mapping science edn. Places and Spaces: Mapping Science, Albuquerque, NM and Berwyn, PA (2006), http://scimaps.org
24. Klavans, R., Boyack, K.W.: Is there a convergent structure to science? In: Torres-Salinas, D., Moed, H. (eds.) Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid. CSIC, pp. 437–448 (2007)
25. Maps of Science, http://www.mapofscience.com/
26. Map Overlays - Google Maps API - Google Code, http://code.google.com/apis/maps/documentation/overlays.html
27. Add Your Own Custom Map - Google Mapki, http://mapki.com/index.php?title=Add_Your_Own_Custom_Map
28. Automatic Tile Cutter - Google Mapki, http://mapki.com/index.php?title=Automatic_Tile_Cutter
29. Custom Google Maps, http://webtide.wordpress.com/2008/08/27/custom-google-maps/
30. Google Maps JavaScript API Example: Tile Detector, http://code.google.com/apis/maps/documentation/examples/tile-detector.html
31. LabeledMarker v1.3 Reference, http://gmaps-utility-library-dev.googlecode.com/svn/tags/labeledmarker/1.3/docs/reference.html

Assessing Group Interaction with Social Language Network Analysis

Andrew J. Scholand[1], Yla R. Tausczik[2], and James W. Pennebaker[2]

[1] Sandia National Laboratories*
Box 5800, Albuquerque, NM 87185
ajschol@sandia.gov
[2] Department of Psychology
University of Texas, Austin, TX 78712
{tausczik,pennebaker}@mail.utexas.edu

Abstract. In this paper we discuss a new methodology, social language network analysis (SLNA), that combines tools from social language processing and network analysis to assess socially situated working relationships within a group. Specifically, SLNA aims to identify and characterize the nature of working relationships by processing artifacts generated with computer-mediated communication systems, such as instant message texts or emails. Because social language processing is able to identify psychological, social, and emotional processes that individuals are not able to fully mask, social language network analysis can clarify and highlight complex interdependencies between group members, even when these relationships are latent or unrecognized.

Keywords: social language processing, social network analysis, network structure, communication, content analysis, group.

1 Introduction

This research, addressing a technical means to make a socially informed group assessment, was motivated by interest in successful group interaction and collaboration. Much knowledge and expertise in successfully executing work quickly is not written down because this information is developed, shared, and acted upon in an operational context through informal conversations among and between groups of individuals. The fundamental research proposition is that digital records arising from interactions in the 'as-is' organization can be analyzed to create an approximate but meaningful representation of the work-centered social dynamics within the group. 'Meaningful' in this context implies facets of information relevant to interpersonal dynamics, aspects of distributed cognition and group work, and the development of organizational power and control.

* Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 248–255, 2010.

In earlier work, we used social language network analysis (SLNA) to diagram a hierarchy of professional respect and predict close personal friendships [6] and to identify dominant cognitive themes in work-related conversations [7]. In this paper, we turn the focus of SNLA to relations within a group (peer to peer) and between leadership and staff.

2 Background

This work leverages social language analysis performed by the Linguistic Inquiry and Word Count (LIWC) software program developed by University of Texas researchers James W. Pennebaker, Roger J. Booth, and Martha E. Francis [4]. LIWC is a program for quantitative text analysis that uses a word count strategy for both the analysis of content (what is being said) and style (how it is being said). Word count strategies are based on the assumption that the words people use convey psychological information over and above their literal meaning and independent of their semantic context. In this sense, they are "top down" in that they explore text within the context of previously defined psychological content dimensions or word categories. (In contrast, word pattern strategies such as latent semantic analysis mathematically detect "bottom-up" how words co-vary across large samples of text, typically to determine the degree to which two texts are similar in terms of their content.) LIWC searches for over 2300 words or word stems previously categorized by independent judges into over 80 linguistic dimensions. These dimensions include standard language categories (e.g., articles, prepositions, pronouns– including first person singular, first person plural, etc.), psychological processes (e.g., positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), relativity-related words (e.g., time, verb tense, motion, space), and traditional Freudian content dimensions.

3 Data Source

The National Infrastructure Simulation and Analysis Center (NISAC) developed a programmable collaboration library to facilitate secure collaborative interaction by geographically distributed decision-makers [3]. The collaboration framework offers the usual collaborative services (chat and file transfer) as well as the ability to publish multiple images for collaborative text and graphical annotation. These capabilities focus primarily on *synchronous* capabilities that allow the integration of multiple perspectives and quick convergence on a shared view of a problem to facilitate high-pressure, time-constrained analyses. This framework has been used since 2003 by the geographically distributed Computational Economics Group to plan, stage, execute, debug, and interpret high performance computing simulations of the national economy subject to regional disruptions. The group also used the tool to evaluate simulation initialization specifications derived from data fused across multiple government and commercial data sources. These work-related instant message conversations between 18

team members were collected for this analysis from September 2006 to November 2007.[1] In this period, there were 14,416 separate statements totaling 170,197 words. The participants included 7 females and 11 males, varying in age from 22 to 64 years old. Four other chat participants were excluded due to contributing less than 250 words in public chat during the period of the study. The number of words contributed per participant ranged from a maximum of 56464 to a minimum of 253, with a mean value of 9435 (standard deviation of 15755) and a median of 2298.

4 Relational Analysis Results

The pattern of conversations in chat illustrate a dichotomous structure underlying the group interaction that in turn affects language use. Figure 1 shows a tree constructed by clustering (Johnson's hierarchical clustering [1] with weighted average clustering) individuals based on the number of conversations they had with each other as a measure of similarity. Individuals who are connected to each other near the left hand side of the dendogram shown in Figure 1 were involved in a higher number of conversations with each other. The gray rectangle superimposed on the dendogram divides the group into two subgroups of equal size. The upper subgroup represents the highly connected 'core' of the group. The lower subgroup represents (a portion of) the more loosely interconnected periphery. Although there is some pairwise structure in the peripheral subgroup, these ties mostly occur at weaker levels than the ties in the core subgroup.

The first person plural pronoun group includes the pronouns 'we,' 'us,' and 'our' as well as various plural and possessive variants. The LIWC program computes the relative ratios of these words to all words spoken in the recorded conversations. Due to variations in speech patterns by age and gender, these metrics are normalized by computing the average of non-zero 'We' pronoun usage percentages for each speaker and deducting this average from those values. Zero values for various individuals indicate no conversations occurred between the speaker and that individual, and so these zeros were left unmodified. The resulting pattern of strongly positive links, indicating conversations between individuals with normalized 'We' pronoun usage values above 2%, were all observed to originate in the core group and link to the peripheral group. Conversely, examining arcs with percentages of -0.75% or lower (relative to each speaker's average usage of 'We' pronouns), 17 of the 22 arcs (77%) both originate and terminate within the core group. The use of 'We' pronouns appears to be substantially less within the subgroup of individuals who comprise the core of this group.

The essence of this finding, then, is that 'We' pronoun usage is inversely related to the degree to which members belong to the group. Those individuals engaging in the most conversations within the group use pronouns from the 'We' group most infrequently when chatting with other frequent conversation partners. To test this association statistically, the Quadratic Assignment Procedure (QAP) [2] can be

[1] The use of these data has been reviewed and approved by Sandia's Human Studies Board in Research Protocol SNL0806.

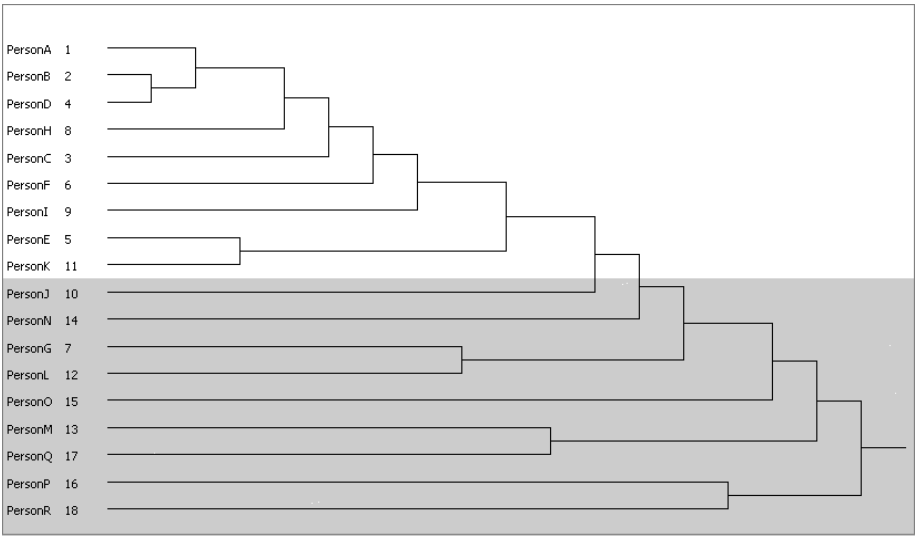


Fig. 1. Johnson Hierarchical Clustering of Conversation Count

Table 1. QAP Correlation of 'We' use and Conversation Count data

Statistic	Value
Pearson Correlation:	-0.314
Significance:	0.000
Permutation Average (50000 permutations):	0.000
Permutation Standard Deviation:	0.086
Minimum Permuted Value:	-0.291
Maximum Permuted Value:	0.306

applied to measure the degree of correlation between the normalized 'We' pronoun use matrix and the conversation count matrix.[2] First, the corresponding cells of the two adjacency matrices are correlated using ordinary Pearson correlation. Second, a large number (50,000) of randomly re-arranged matrices are correlated to assess if the observed match is likely by pure chance. If the proportion of random trials that would generate a coefficient as small as the statistic actually observed is small enough, typically below 0.05, the hypothesis of no association is rejected. Table 1 shows that randomly permuted matrices on average have no correlation whatsoever (Pearson Correlation of 0.000), and therefore the observed inverse correlation of -0.314 is highly significant statistically.

[2] The conversation count data was zero-meaned before being processed, so that the Pearson's coefficient is computed for centered data.

Table 2. Pearson Correlation of LIWC Categories and Conversation Count data

Category	Examples	Value	Signif.	Avg.	S. D.	Min.
biological processes	breakfast, cafeteria, pizza	-0.330	0.000	0.000	0.082	-0.289
smileys	:-)	-0.325	0.000	0.000	0.082	-0.269
body	eye*, face, sleep*	-0.319	0.000	0.000	0.075	-0.252
first person plural	we, us, our	-0.314	0.000	0.000	0.086	-0.291
health	ache*, exercis*, pills	-0.312	0.000	0.000	0.071	-0.249
inclusive	both, come, inclu*	-0.285	0.000	0.000	0.077	-0.282
ingestion	ate, chew*, coffee	-0.269	0.000	-0.001	0.088	-0.305
family	family, husband, wife*	-0.264	0.000	0.000	0.061	-0.272
third person plural	their*, them, they've	-0.244	0.000	-0.001	0.074	-0.240
discrepancy	besides, if, problem*	-0.237	0.000	0.000	0.077	-0.258

To put these findings in context, Table 2 lists this category with the other LIWC categories that are most strongly associated with group structure. All of these associations are negative, suggesting that the core subgroup focuses on these categories primarily in communications with the peripheral subgroup rather than among themselves. The topics suggest attention to health and wellness (health, body, ingestion, biological processes), non-work issues (family), and minimizing communication misunderstandings (smileys) in these communication channels. As discussed above, there are multiple components suggesting outreach and perhaps attempts to verbally assimilate the periphery into the core, including the 'we,' 'they,' and 'inclusive' categories. The 'discrepancy' group is the only category suggesting a specific work-related focus; discrepancy words are used to differentiate concepts.

5 In-Group and Management Relationships

Participants completed a questionnaire rated their attitudes toward each of their collegues by evaluating nine different statements on a seven point Likert scale. Eleven of the 18 participants whose messages were recorded in the public chat forum responded to our request to complete surveys. These participants, 4 female and 7 male, ranged in age from 22 to 64. The ratings each participant gave to each of the other group members can be considered that person's conceptualization of their dyadic relationship. Chats with those individuals is then an instance of the class of conversations that occurs at that degree of relation. For example, assume Person A rates Person B as a close personal friend at Likert scale degree 7, 'a great deal.' When Person A chats to Person B, then, this language can be considered to be representative of chat between very close friends. If Person E similarly rates Person K as a very close friend, chat from Person E to Person K would also be categorized as being between very close friends. Aggregating all of the language from individuals who rated their conversational partner at each Likert scale level yields a sample of language across the spectrum of sentiment

for a given question. We subjected this partitioning of the chat language sample to LIWC content analysis and correlated the results to the scale level.

A common observation in the organizational studies literature is that negative information is increasingly filtered as it moves up the management chain [5]. Figure 2 provides empirical support for this assertion. The graph on the left side of Figure 2 shows a negative correlation between the LIWC category 'Negemo' and higher social status. People spoke with fewer negative terms when conversing with individuals they perceived to be of higher status. The graph on the right shows a reinforcing effect, namely that people in this group tended to use more positive terms ('Posemo') the less well they knew the person. Hence, management, having both higher social status[3] and being less well known by most staff members than their peers, receive both less negative information and more positive information from staff.

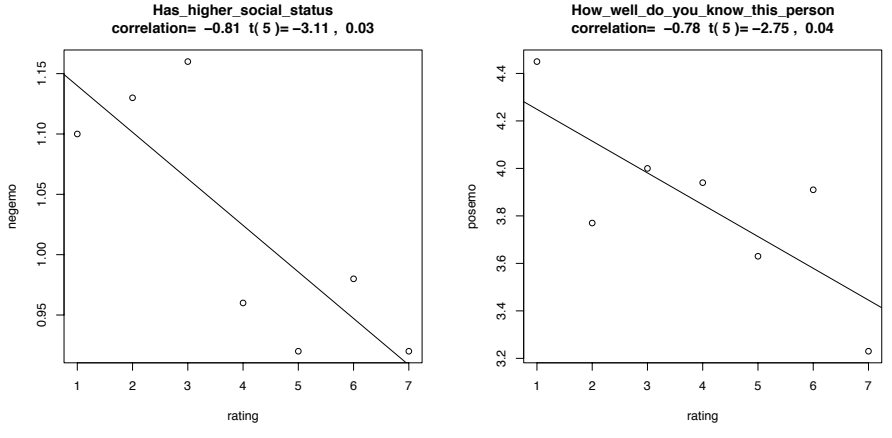


Fig. 2. Emotional Filtering as a Function of Status and Familiarity

How group members rated the communication skills of their peers strongly predicted the types of conversations held with those peers. The graph on the left side of Figure 3 shows a link between how easy to talk to a person is perceived to be and the extent to which conversations with that person contain tentative words. People are willing to share thoughts and interpretations they are less sure of when their conversational partner is easy to talk to. Conversely, the graph on the right side of Figure 3 shows that more difficult to approach individuals are met with more "causative" language. Conversations are initiated with these individuals predominantly when there is a reason to approach them.

[3] The manager and team leads in this work group were rated as having high social status in the questionnaire.

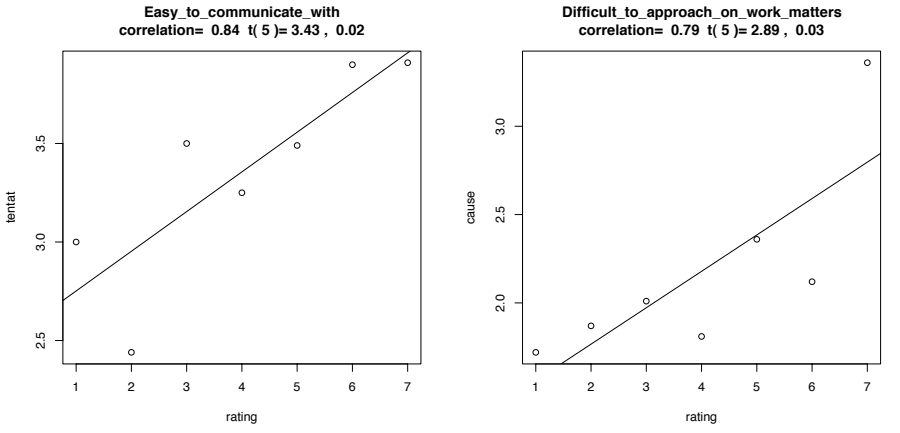


Fig. 3. Type of Conversation as Function of Communication Skills

6 Conclusions

This work combines the high fidelity assessment of relationships between entities made possible by language analysis with the contextual framework of social network processing to both predict underlying structural relations and retrospectively describe patterns of group interaction. By selectively extracting, combining, and processing different psychological, social, and emotional linguistic markers it is possible to map the rich relationships within and across groups, making difficult tasks such as managing organizational change, organizational design, and interorganizational relationships easier.

A study of language use between the well-connected core and periphery of a group of 18 scientific collaborators highlights efforts at outreach and assimilation, including exceptional use of pronouns such as 'we' and 'they,' and other 'inclusive' language. Only one work-related LIWC category ('discrepancy' words; used to differentiate concepts) was used in a similar statistically significant manner, suggesting work is primarily accomplished in subgroups rather than across the group as a whole. Correlating LIWC measures with survey elicited evaluations of co-workers revealed people in this organization are more willing to share tentative interpretations when their conversational partner is easy to talk to. In contrast, conversations are typically initiated with less socially skilled individuals only when there is a specific reason to approach them. Vertical information flow within the group is also shown to be influenced by both status and social distance, with high status and aloof leaders receiving both decreases in negative information and increases in positive information from staff.

Acknowledgments. Thanks to the Army Research Institute (W91WAW-07-C-0029) and the Sandia Laboratory Directed Research & Development Seniors Council for the funding that made this publication possible.

References

1. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 32(3), 241–254 (1967)
2. Krackhardt, D.: QAP Partialling as a Test of Spuriousness. Social Networks 9(2), 171–186 (1987)
3. Linebarger, J.M., Scholand, A.J., Ehlen, M.A., Procopio, M.J.: Benefits of Synchronous Collaboration Support for an Application-centered Analysis Team Working on Complex Problems: a Case Study. In: GROUP 2005, pp. 51–60. ACM Press, New York (2005)
4. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic Inquiry and Word Count: A computerized text analysis program, Austin, TX (2007), www.LIWC.net
5. Reason, J.: Managing the Risks of Organizational Accidents. Aldershot, Ashgate (1997)
6. Scholand, A.J., Tausczik, Y.R., Pennebaker, J.W.: Social Language Network Analysis. In: CSCW 2010. ACM Press, New York (2010)
7. Scholand, A.J., Tausczik, Y.R.: Diagramming Workgroup Interaction via Social Language Network Analysis. In: CSCW 2010 Workshop on The Changing Dynamics of Scientific Collaborations (2010)

Analyzing and Tracking Weblog Communities Using Discriminative Collection Representatives*

Guozhu Dong and Ting Sa

Department of Computer Science and Engineering
Wright State University, Dayton, Ohio 45435, USA

Abstract. Analyzing/tracking weblogs by given communities (ATWC) is increasingly important for sociologists and government agencies, etc. This paper introduces an approach to address the needs of ATWC by using concise discriminative weblog collection representatives (DCRs). DCRs are aimed at helping users to quickly identify the major themes/trends in such collections, and to quickly identify important shifts/differences in major themes and trends of blogs by given communities over time and space. We propose to use the quality of DCR-based classifiers to measure DCRs' quality. We present algorithms for constructing DCRs, report experimental results to evaluate the efficiency of the algorithms and the quality of the DCRs they construct, and provide real-data examples to demonstrate the usefulness of DCRs for ATWC.

1 Introduction

The ability to track and analyze, in a timely manner, weblogs written by given communities becomes an important problem for sociologists, government agencies, and various organizations [1]. In this paper we consider weblog tracking and analysis by extracting concise discriminative collection representatives from large collections of weblogs, in order to help users quickly identify, the major themes/trends in such collections of blogs, and identify shifts/differences in major themes and trends of blogs over time/space.

To illustrate, suppose C is a target collection of weblogs (by bloggers of a given community X in some given time window) to be analyzed. In order to downplay the importance of words that occur frequently in non-C blogs, we gather a background collection C' of weblogs (by bloggers not in X or by bloggers of X but in a different time window or region). A discriminative weblog collection representative (DCR) for C and C' contains two small sets S_1 and S_2, each containing k words for some small $k > 0$. Loosely speaking, a DCR for C and C' describes the main themes of C that do not occur in C' (and vice versa).

Example 1. The weblog data from [10] contains 11 predefined categories[1], including health. Suppose we want to analyze the collection C containing the first

* Part of the work was supported by a WSU/OBR Research Challenge grant.

[1] Although the examples used weblogs with known categories, we emphasize that DCRs are especially useful for suggesting such categories when they are not known and our algorithms for constructing DCRs do not require known weblog categories.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 256–264, 2010.

1000 blogs in the health category. For the contrast collection C', we collect 100 blogs from each of the 10 other categories. The table below lists a DCR for C and C'. Clearly, the words for C are all about health and the words for C' are not. The DCR can help us identify the major themes of C, i.e. health, and estimate what the blogs in C are mainly about.

C: cause health suffer weight	C': movie car music dog

Example 2. For weblog tracking,suppose we want to identify the major changes between the first 1000 blogs (denoted by C), and the next 1000 blogs (denoted by C'), of the health category. (This C is identical to the C of Example 1.) The table below contains a DCR for C and C'. Since C and C' are much more similar to each other than C and C' of Example 1, this DCR is not as informative as that of Example 1 for suggesting the major themes. Importantly, this DCR tells us how C and C' differ from each other, using unique representative theme words from the two collections. Moreover, the four words for C can be viewed as disappearing, and those for C' can be viewed as emerging, theme words.

C: treatment drug popular acid	C': blog week am stroke

Since our aim is to use DCRs to suggest the main themes of two given weblog collections C and C', we propose to evaluate a DCR's quality using the accuracy (measured by FScore) of the simple and natural Naive Bayes classifier, built using the words in the DCR, to classify the weblogs in C and C'. Our rationale is: If the DCR-based classifier has high FScore value, then the words in the DCR have high potential of being able to tell the main themes of the weblog collections and to tell the main change/difference between the two given weblog collections.

Besides introducing DCR for weblog community tracking and analysis, and proposing the DCR-based classification approach to evaluate DCR's quality, this paper makes the following contributions. It presents two algorithms for constructing DCRs: (a) The DCRFS algorithm constructs DCRs by directly using the FScore to incrementally refine DCRs. (b) The DCROD algorithm constructs DCRs by using a more efficiently computable "weaker surrogate" measure to incrementally refine DCRs. The surrogate measure is based on three factors: the discriminativeness of words in DCRs, the collective coverage and the overlap of coverage of blogs by words in DCRs. DCROD is much faster than DCRFS but DCRFS produces higher quality DCRs. An experimental evaluation of our algorithms is reported, concerning computation efficiency, DCR quality, and DCR's usefulness for weblog analysis and tracking. Our experiments also demonstrate that selecting high frequency/ tf-idf words may not lead to high quality DCRs.

1.1 Related Works

Our work is related to weblog analysis and tracking, and to document (collection/clustering) summarization. (a) Previous studies on weblog analysis and tracking include: [2] considered how to identify influential bloggers in a community. [1] presented a tool for tracking and analyzing blogs, which can be used to identify frequent terms, influential bloggers and relationship among bloggers. (b) Document summarization is aimed at generating a small summary of one or

more documents in order to help users to quickly figure out the main ideas of the documents. Such research can be divided into three groups, namely single document summarization [7], single document collection summarization [6,3,9], and multiple document collection summarization [4]. Our work is novel as follows: we use small sets of "theme" words to track/analyze blog collections; we use a summary-based classifier to evaluate a (DCR) summary's quality; we introduce an overlap-adjusted-discriminativeness[2] based measure to efficiently construct DCRs for tracking and analyzing blog collections.

2 DCR and DCR-Classification Based Quality

We view a blog d as a set of words. For a collection C of blogs, let $TS(C) = \cup_{d \in C} d$ denote the set of words of C. For each word t and each collection C of blogs, let $freq_C(t) = \frac{|\{d \in C | t \in d\}|}{|C|}$ denote t's frequency in C.

Definition 1 (DCR). *Given two*[3] *collections C_1 and C_2 of blogs and a positive integer k, a* discriminative collection representative *(DCR) of size k is a pair (S_1, S_2) where $S_i \subseteq TS(C_i)$ is a set of words satisfying $|S_i| = k$ for each i.*

To ensure that a good DCR is indeed a good concise discriminative representation of the given collections, and that the words in the DCR are suggestive of the main themes of the blogs in the collections, we propose to use a DCR-based classification approach to evaluate the quality of a DCR. In this approach, a DCR works as a classifier on the blogs in the collections (viewed as classes). The similarity between the new classes induced by the classifier and the original classes is then used to measure the quality of the DCR; a high similarity indicates a high quality DCR. While many classifiers can be used, in this paper we use the Naive Bayes classifier since it is simple and natural.

Given a DCR $S = (S_1, S_2)$ for C_1 and C_2, let NB_S denote the associated DCR-based Naive Bayes classifier. NB_S requires the probabilities of the classes, namely $P(C_i)$, and the conditional probabilities of each word t given the classes, namely $P(t|C_i)$, where t is a word in $S_1 \cup S_2$ and C_i is a class. NB_S classifies a blog d as belonging to the class C_j (among C_1 and C_2) that maximizes[4] the following probability[5]: $P(C_i|d) = P(C_i) \times \Pi_{t \in d \cap (S_1 \cup S_2)} P(t|C_i)$.

Definition 2 (DCR-based Classification). *Suppose $S = (S_1, S_2)$ is a DCR for two collections C_1 and C_2 of weblogs. The collections induced by the DCR S are C_1' and C_2', where $C_i' = \{d \in C_1 \cup C_2 | P(C_i|d) > P(C_j|d), j \in \{1,2\} - \{i\}\}$.*

Example 3. To illustrate DCR-based classification, consider two collections $C_1 = \{d_{11}, d_{12}\}$ and $C_2 = \{d_{21}, d_{22}\}$, where $d_{11} = \{a, b, e, f\}$, $d_{21} = \{a, e, f\}$, $d_{12} = \{a, c\}$, and $d_{22} = \{d, e\}$ are blogs. Consider the DCR $S = (S_1, S_2)$, where $S_1 =$

[2] This measure helps minimize overlap, which is related to the MMR principle [3].
[3] DCRs can be easily generalized to situations with > 2 collections.
[4] NB_S breaks tie (i.e. $P(C_1|d) = P(C_2|d)$) by assigning d to C_1.
[5] $P(d)^{-1}$ is ignored on the RHS, without any effect on the classification outcome.

$\{a, b\}$ and $S_2 = \{d, e\}$. An example conditional probability of NB_S is $P(a|C_1) = (2 + 1)/(2 + 4) = 0.5$. To classify d_{11}, NB_S computes $P(C_1|d_{11}) = P(C_1) * P(a|C_1)*P(b|C_1)*P(e|C_1) = 0.5*0.5*0.33*0.33 = 0.028$, and $P(C_2|d_{11}) = 0.013$. Since $P(C_1|d_{11}) > P(C_2|d_{11})$, NB_S assigns d_{11} to C_1'. Moreover, the collections induced by S are $C_1' = \{d_{11}, d_{12}, d_{21}\}$ and $C_2' = \{d_{22}\}$.

If the induced collections are very similar to the original ones, the quality of the DCR is considered high. We use the FScore to measure this similarity.

3 Incremental DCR Construction Algorithms

In this section we present two DCR construction algorithms. We first give our FScore-Based Incremental DCR Search (DCRFS) Algorithm. Since DCRFS is often rather slow (see Section 4), we present a more efficient algorithm, called Overlap Adjusted Discriminativeness Based Incremental DCR Search Algorithm (DCROD). Its key idea is to use an efficiently computable surrogate quality measure Ω_{OD} to estimate the FScore value. It trades a slight loss on DCR quality for a big gain on efficiency, making it the algorithm to use when speed is very important.

3.1 The DCRFS Algorithm

DCRFS starts from some seed DCR $S = (S_1, S_2)$, where S_i contains the top k words ranked by frequency ratio. It then incrementally finds the best (measured by FScore) improving replacement to refine the DCR, among all possible single-word replacement DCRs (see Section 3.4).

3.2 Overlap and Discriminativeness for DCR

To design an efficiently computable substitute for the expensive-to-compute FScore measure, we first define two necessary concepts. Let C_1 and C_2 be two collections. (1) A word t is said to be C_i-discriminative if it is much more frequent in C_i than in C_j ($j \neq i$). Let $DIS_-(t, C_i) = freq_{C_i}(t) - freq_{C_j}(t)$, where $j \in \{1, 2\} - \{i\}$; we say $DIS_-(t, C_i)$ is the C_i-discriminativeness of t; we will use $DIS_-(t, C_i)$ to rank words. We also consider the overall C_i-discriminativeness of a set X of words, defined to be $DIS_-(X, C_i) = \sum_{t \in X} DIS_-(t, C_i)$. (2) For each word t, finite set X of words, and C_i, let $cover(t, C_i) = \{d \in C_i | t \in d\}$, and $cover(X, C_i) = \cup_{t \in X} cover(t, C_i)$.

Let $S = (S_1, S_2)$ be a DCR for two weblog collections C_1 and C_2. The rationale behind our surrogate measure formulation is that the FScore quality of S is usually high, when the following are true:

(a) For each i, the words in S_i are highly C_i-discriminative, and the overall collective C_i-discriminativeness of S_i is high.
(b) $|cover(S_i, C_i)|$ is high; i.e., S_i's words collectively cover many blogs of C_i.

Table 1. Impact of Discriminativeness and Coverage on FScore of DCRs

(i) Two Weblog Collections

C_1	C_2
$d_{11}:\{a,c,f\}$	$d_{21}:\{f,e\}$
$d_{12}:\{a,c,h\}$	$d_{22}:\{f,g\}$
$d_{13}:\{a,e\}$	$d_{23}:\{h,e\}$
$d_{14}:\{d,e\}$	$d_{24}:\{h,g\}$
$d_{15}:\{b\}$	$d_{25}:\{e\}$
$d_{16}:\{d\}$	

(ii) Frequency and Discriminativeness of Words

Word t	$freq_{C_1}(t)$	$freq_{C_2}(t)$	$DIS\_(t,C_1)$	$DIS\_(t,C_2)$
a	0.5	0	0.5	-0.5
b	0.17	0	0.17	-0.17
c	0.33	0	0.33	-0.33
d	0.33	0	0.33	-0.33
e	0.33	0.6	-0.27	0.27
f	0.17	0.4	-0.23	0.23
g	0	0.4	-0.4	0.4
h	0.17	0.4	-0.23	0.23

(iii) Discriminativeness, Coverage and FScore of Three DCRs

DCR	C_1	C_2	$\lvert cover(X_{i1},C_1)\rvert$	$\lvert cover(X_{i2},C_2)\rvert$	$DIS\_(X_{i1},C_1)$	$DIS\_(X_{i2},C_2)$	FScore
X_1	$X_{11}=\{a,c\}$	$X_{12}=\{e,h\}$	3	4	0.83	0.5	0.82
X_2	$X_{21}=\{b,d\}$	$X_{22}=\{e,h\}$	3	4	0.5	0.5	0.73
X_3	$X_{31}=\{a,d\}$	$X_{32}=\{e,h\}$	5	4	0.83	0.5	0.91

Example 4. To illustrate the points of (a) and (b), consider the collections C_1 and C_2 in Table 1. The blogs are denoted by d_{ij}.

Consider (a). From Table 1(i), we see that a, b, c, d are preferable candidates for representing C_1, since they are more frequent in C_1 than in C_2. Table 1(ii) shows that a, b, c, d have high C_1-discriminativeness values. Similarly, e, f, g, h are preferable for C_2. Consider the first two DCRs, X_1 and X_2 given in Table 1(iii). Observe that the X_1 row and the X_2 row have identical $|cover|$ and $DIS\_$ values except that $DIS\_(X_{11}, C_1) = 0.83 > 0.5 = DIS\_(X_{21}, C_1)$; moreover, $FScore(X_1) = 0.82 > 0.73 = FScore(X_2)$. This illustrates (a).

For (b), consider the DCRs X_1 and X_3 in Table 1(iii). Observe that the X_1 row and the X_3 row have identical $|cover|$ and $DIS\_$ values except that $|cover(X_{11}, C_1)| = 3 < 5 = |cover(X_{31}, C_1)|$; moreover, $FScore(X_1) = 0.82 < 0.91 = FScore(X_3)$. This illustrates the intuition stated in (b).

The first function assesses how a DCR behaves in relation to (a).

Definition 3. *The discriminativeness of a DCR $S = (S_1, S_2)$ for collections C_1 and C_2 is defined to be $DIS\_(S, C_1, C_2) = \sum_{i=1}^{2} DIS\_(S, C_i)$.*

We now turn to defining our second function, namely "overlap", for assessing how a DCR behaves in relationship to (b). This function is almost the inverse of the cover function. The rationale for choosing this function is: Basic set theory indicates that, for each S_i and C_i, the value of $|cover(S_i, C_i)|$ can be approximated by $\sum_{t \in Si} |cover(t, C_i)| - \sum_{s,t \in S_i, s \neq t} |cover(s, C_i) \cap cover(t, C_i)|$. Since the first sum is already partially reflected by the discriminativeness function, we will now focus on the overlap part (corresponding to the second sum). We need to define overlap on two words, and overlap for a DCR.

Definition 4. *The overlap between two words s and t w.r.t. C_i is defined by $OLP(s, t, C_i) = |cover(s, C_i) \cap cover(t, C_i)|/|C_i|$. The overlap of a DCR $S = (S_1, S_2)$ for C_1, C_2 is defined by $OLP(S, C_1, C_2) = \sum_{i=1}^{2} \sum_{s,t \in S_i, s \neq t} OLP(s, t, C_i)$.*

3.3 The Ω_{OD} Surrogate Quality Measure

The previous section argued that discriminativeness and overlap of a DCR are two factors that influence the FScore value of the DCR. We now formalize the Ω_{OD} surrogate measure, as a weighted difference of the discriminativeness and overlap, with the intention to maximize discriminativeness and minimize overlap.

Definition 5. *Let $S = (S_1, S_2)$ be a DCR for C_1 and C_2. The Ω_{OD} quality value for S is defined by $\Omega_{OD}(S) = a * DIS\_(S, C_1, C_2) - b * OLP(S, C_1, C_2)$, where a, b are certain coefficient values (to be determined). The default values for a and b are both* 1.

3.4 The DCRod Algorithm

To find a desired DCR of size k for two collections C_1 and C_2, the DCRod algorithm starts from an initial DCR $S = (S_1, S_2)$, obtained by extracting certain k words from each C_i (discussed below). For S, all possible single-word replacement[6] DCR candidates S^n are generated. Let S' be the replacement DCR where $\Omega_{OD}(S')$ is the largest among the replacement DCRs S^n. If S' strictly improves S (i.e. $\Omega_{OD}(S') > \Omega_{OD}(S)$ and at least one of $DIS\_$ and OLP improves), then we replace S by S'. The above process is repeated until no improving replacement DCR of S can be found.

Since non-discriminative words are not likely to lead to high quality DCRs, we only consider C_i-discriminative words as candidate replacement words for S_i.

The DCRod Algorithm

Input: Two collections C_1 and C_2 of weblogs and an integer k

Output: A DCR $S = (S_1, S_2)$ of size k for C_1 and C_2 with high Ω_{OD} value

1. Generate an initial DCR $S = (S_1, S_2)$, where S_i contains the top k words of C_i ranked by $DIS\_(t, C_i)$ values; compute and store $OLP(s, t, C_i)$ values for all pairs of distinct C_i-discriminative words s, t for use in Step 2;

2. Repeat the following until no improving replacement can be found:

  2.1 Let S' be the S^n with the highest Ω_{OD} value, among all the possible single-word replacements S^n of S, such that (a) $\Omega_{OD}(S') > \Omega_{OD}(S) + \epsilon$ and (b) either $DIS\_(S') > DIS\_(S)$ or $OLP(S') < OLP(S)$;

  2.2 If S' exists in the previous step, let $S = S'$;

3. Return S.

The default value of ϵ is 0.0001. Roughly speaking, the computation time for DCRod is $O(k^2 N_i N_t + N_t^2 N_d)$, where N_i is the number of iterations over Step 2, k is the desired DCR size, N_t is the average number of candidate replacement words, and N_d is the average number blogs per collection.

[6] For each integer i ($1 \leq i \leq 2$), word $t \in S_i$ and C_i-discriminative word $t' \in TS(C_i) - S_i$, $S^n = (S_1^n, S_2^n)$ is a replacement DCR of S, obtained by replacing t with t', where $S_j^n = S_j$ ($j \in \{1, 2\} - \{i\}$) and $S_i^n = (S_i \cup \{t'\}) - \{t\}$.

3.5 Determining the Coefficients of Ω_{OD} Using Linear Regression

The default a and b values may not produce the best DCRs. Better a and b values can be obtained by collecting a set of training triples $(DIS\_(S), OLP(S), FScore(S))$, where S is a DCR, and then perform linear regression on the set of triples. This process is optional for each given pair of collections C_1 and C_2.

The set of training triples can be obtained from a set of DCRs, generated by running a variant of the DCRod algorithm, with the following modifications: The seed DCR is initialized using the k words having the lowest $DIS\_(t, C_i)$ values for each C_i (to ensure that we get a reasonable number of improving replacement DCRs), and the default $a = 1$ and $b = 1$ values are used in the Ω_{OD} measure. A triple is generated from each replacement DCR that gives (i) improving Ω_{OD} value and (ii) either larger $DIS\_$ value or smaller OLP value.

4 Experimental Evaluation

This section reports our experimental results on (1) the quality and informativeness of DCRs computed by our algorithms for weblog analysis and tracking, (2) the importance of the discriminativeness and overlap factors for constructing good quality DCRs, and (3) the efficiency of our algorithms. We also compare our algorithms against several other potential competing algorithms.

4.1 Data Sets Used

Our experiments were performed on collections extracted from the BlogCatalog weblog data set [10]. Blogs in this data set have pre-defined categories[7] and are written in several languages. We extracted the English blogs from the following 11 popular categories: animals, autos, business, film, food & drink, health, music, political, sports, technology and travel. All blogs used were preprocessed by removing stop-words and stemming, following common procedures in text processing; an offensive four-letter word was also removed since we do not want to print it in a DCR. We also removed short blogs containing < 30 words (including duplicate words), before extracting our data sets.

We constructed four blog data sets, each containing two collections C_1, C_2. D_1's C_1 contains the first 1000 blogs of the sports category, D_2's C_1 contains the first 1000 of music, and D_3's C_1 contains the first 1000 of health. Each D_i's C_2 contains 100 blogs from each of the ten categories not used in D_i's C_1. For D_4, we put the first (second, resp.) 1000 blogs to C_1 (C_2, resp.). (So D_4's C_1 is identical to D_3's C_1.) Each C_i contains around 10000 distinct words.

[7] As noted earlier, the existing categories are not used by our algorithms. We use data with known categories here to demonstrate that DCRs can be used to suggest the categories when the categories/themes are not known.

Dataset	DCR on C_1	DCR on C_2	FScore(C_1)	FScore
D_1	team ride game martial match sports	music recipe car people market dog	77.24%	71.26%
D_2	album music song guitar band dj	people cup car company food dog	78.48%	72.66%
D_3	cause health meditation suffer weight surgery	movie car nice dog music city	75.62%	66.81%

4.2 DCRs Obtained by DCROD, Suggestive Power and Quality

The table above lists the DCRs of size 6 obtained by DCROD, together with the FScores of the DCRs. Linear regression was used to determine a and b of Ω_{OD}. Most importantly, we can see that the words for C_1 are all related to the category where the blogs are extracted from. Hence the DCR words have high suggestive power on the themes of the collections. The FScores are quite high, and the FScores on C_1 are even higher, indicating that the words for C_1 are really discriminative for C_1 and they represent the main themes of C_1.

Remark: In our experiments, we observed that DCRs with $k = 4$ or $k = 6$ are often good enough to suggest the main themes of the collections.

4.3 Using DCROD and FScore for Weblog Tracking

To demonstrate the ability of DCR for weblog community tracking, we used DCROD to extract a DCR of size 6 for D_4. Recall that C_1 and C_2 of D_4 are both from the health category, and hence they contain highly similar blogs. We got the DCR $S = (S_1, S_2)$, where $S_1 = \{treatment, drug, fruit, acid, red, shoulder\}$ and $S_2 = \{week, blog, am, focused, department, stroke\}$, with a FScore of 57.80%.

Remark: When C_1 and C_2 are similar to each other, the generated DCR may have very low FScores. Importantly, the DCR can indicate the subtle representative differences between the collections. This ability can be very useful to track both the degree of change/difference and the main aspects of change/difference between weblogs over different time periods or from two different regions.

4.4 Impact of OD Factors and Linear Regression

Experiments confirmed that discriminativeness and overlap are both important factors for constructing high quality DCRs. For D_1, for example, the relative FScore loss is around 3% when the overlap factor is ignored, and the relative loss is around 9% when the discriminativeness factor is ignored. Linear regression led to relative FScore improvement of 4.35% on average; the improvement depends on the dataset, and comes at a cost of additional computation time.

4.5 Other Performance Issues and Comparisons

(a) Experiments indicate that DCROD can find DCRs fairly quickly, in about 20 seconds (3 minutes and 18 minutes, resp.) for DCR sizes 4 (8 and 16 resp.) (including time for training triple generation and linear regression). The execution

time of DCRFS is about 25 times that of DCROD on average. The experiments were performed on a PC with 2.53GHz CPU and 1024MB of memory, running Windows XP, and the algorithms were coded in java. (b) DCRs obtained by DCRFS have significantly higher FSores (with relative improvement of 15% on average), although we did not notice any significant improvement on their ability to suggest the main themes of the collections. (c) Experiments also show that DCRs containing words with the highest frequency, or with the highest category adjusted tf-idf scores [5], for each C_i have poor suggestive power, and have low FScores.

5 Concluding Remarks

This paper introduced the concept of DCR (discriminative weblog collection representatives) to help suggest the main themes of collections of blogs by given communities, and to track such blog collections over time and space. A DCR can be viewed as a description of the collective behavior of the blogs in given collections. Among the two algorithms proposed, DCRFS can be used if time is not a big concern and DCROD can be used if speed is important. Often, DCRs with very few words (e.g. 6) can suggest the main themes of the given collections.

References

1. Agarwal, N., Kumar, S., Liu, H., Woodward, M.: BlogTrackers: A Tool for Sociologists to Track and Analyze Blogosphere. In: AAAI Conf. on Weblogs and Social Media (2009)
2. Agarwal, N., Liu, H., Tang, L., Yu, P.: Identifying Influential Bloggers in a Community. In: Intl. Conf. on Web Search and Data Mining (2008)
3. Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. SIGIR, 335–336 (1998)
4. Chen, L., Dong, G.: Succint and informative cluster descriptions for document repositories. In: Int'l. Conf. on Web-Age Information Management (2006)
5. Fisher, S., Roark, B.: OGI/OHSU Basline Multilingual Multi-document Summarization System. Mutli-lingual Summarization Evaluation (2005)
6. Lin, C.Y., Hovy, E.: Automated multi-document summarization in neats. In: Proceedings of the Human Language Technology Conference (2002)
7. Metzler, D., Kanungo, T.: Machine learned sentence selection strategies for query-biased summarization. In: SIGIR Learning to Rank Workshop (2008)
8. Mihalcea, R., Tarau, P.: TextRank: bringing order into texts. In: Conference on Empirical Methods in Natural Language Processing (2004)
9. Shen, D., Sun, J.T., Li, H., Yang, Q., Chen, Z.: Document summarization using conditional random fields. IJCAI (2007)
10. Zafarani, R., Liu, H.: Social Computing Data Repository at ASU. In: School of Computing, Informatics and Decision Systems Engineering, Arizona State University (2009), http://socialcomputing.asu.edu

Assortativity Patterns in Multi-dimensional Inter-organizational Networks: A Case Study of the Humanitarian Relief Sector

Kang Zhao, Louis-Marie Ngamassi, John Yen,
Carleen Maitland, and Andrea Tapia

College of Information Sciences and Technology,
The Pennsylvania State University, University Park, PA 16802, USA
kangzhao@psu.edu, {ltchouakeu,jyen,cmaitland,atapia}@ist.psu.edu

Abstract. We use computational tools to study assortativity patterns in multi-dimensional inter-organizational networks on the basis of different node attributes. In the case study of an inter-organizational network in the humanitarian relief sector, we consider not only macro-level topological patterns, but also assortativity on the basis of micro-level organizational attributes. Unlike assortative social networks, this inter-organizational network exhibits disassortative or random patterns on three node attributes. We believe organizations' seek of complementarity is one of the main reasons for the special patterns. Our analysis also provides insights on how to promote collaborations among the humanitarian relief organizations.

Keywords: Assortativity, multi-dimensional inter-organizational network, network analysis, humanitarian relief.

1 Introduction

Networks can represent relationships among entities by connecting related nodes with edges. Many real-world systems can be modeled as networks, such as social networks and the World Wide Web. However, a one-dimensional network can capture only one type of relationship. Take the network among people as an example. A colleague network depicts who work together, while a kinship network is based on family ties. However, people may have multiple types of relationships between them. Thus a person's social network is inherently multi-dimensional and combines multiple one-dimensional networks, such as friendship, kinship, and co-workership. Similarly, an inter-organizational network, in which nodes represent organizations and edges denote relationships, may also have multiple dimensions on the basis of partnership, patronship, sponsorship, and so on.

Assortativity describes the tendency of nodes in a network being connected with similar nodes. For example, sociologists found assortative patterns in social networks. One tends to bond with those who are similar to oneself in demographic characteristics, such as age, gender, race, and education [1]. In other words, assortative patterns emerge from the homophily-based network growth.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 265–272, 2010.

When talking about assortativity for a multi-dimensional network, one needs to be aware of (1) which node attribute and (2) what type of relationship the assortativity is based on. Assortativity depends on node attribute, because assortativity is based on inter-node similarity, which can be measured by various node attributes. Consequently, a network may exhibit different assortative patterns when it is evaluated using different attributes. For example, when we focus on race, a dating network among students in a university may be assortative, which means people tend to date with those who are in the same ethnic group. However, such a dating network may exhibit disassortative patterns when we measure it with students' gender. In addition, even when we measure assortativity with the same node attribute, a multi-dimensional network may have different assortative patterns on different dimensions. For instance, if we only look at the gender of students in a college student social network, then the friendship network may be assortative while the dating network may be disassortative.

The study of assortative patterns in networks can improve our understanding of the network structure and the behaviors of network members. Research in this area can also provide insights on how to improve or attack a network. Network researchers often study assortativity using topological attributes of nodes [2]. Sociologists who study homophily are more interested in individuals' demographic characteristics [1]. Few have considered both the topological and individual attributes of nodes. Also, while some studied homophily of collaborations among companies [3], we find little research on assortativity of multi-dimensional inter-organizational networks. In this research, we study such a multi-dimensional inter-organizational network in the humanitarian relief sector and explore its assortativity using node attributes on both macro and micro levels.

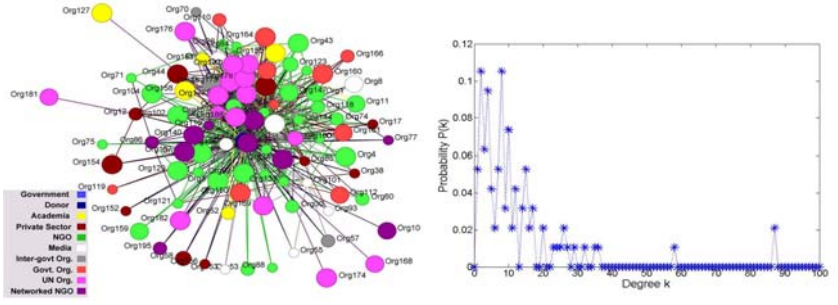
The remainder of the paper proceeds as follows. In Section 2, we introduce our case study–an inter-organizational network in the humanitarian relief sector. Section 3 describes how we analyze assortative patterns on the basis of several node attributes, and illustrates the results. After discussions on the implications of the research, the paper concludes with directions for future work.

2 An Inter-organizational Network in the Humanitarian Relief Sector

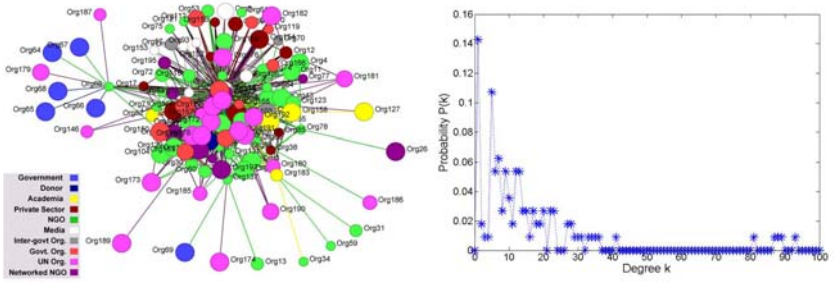
Relief efforts after major natural disasters have highlighted the importance of great levels of inter-organizational coordination. One approach taken by humanitarian organizations has been to organize "coordination bodies" to improve relief efforts through greater coordination among its member organizations [4]. Coordination bodies provide a venue where humanitarian organizations interact with each other and establish further relationships. In this research, we focus on the GlobalSympoNet[1], a major inter-organizational coordination body. Only invited organizations can become members of GlobalSympoNet and attend its meetings.

Through surveys and interviews, we identified a network that consists of about 119 member organizations of GlobalSympoNet. We found 3 dimensions for

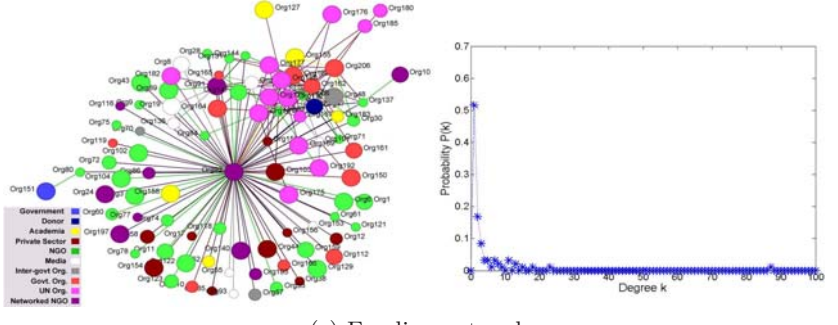
[1] Pseudonyms are used to protect the confidentiality of these organizations.



(a) Advising network



(b) Collaboration network



(c) Funding network

Fig. 1. Multiple dimensions of the inter-organizational network inside GlobalSympoNet

inter-organizational relationships, namely advising, collaboration and funding. An edge in the advising network means the two organizations have exchanged advice concerning policy, technology, data, etc. In the dimension of collaboration, two organizations are connected by an edge if they used to collaborate on humanitarian projects, such as joint training of staff members, coordinated data collection, and shared database. Edges in the funding network denote relations between funding providers and receivers. Figure 1 shows the three one-dimensional networks and their degree distributions.

3 Results

3.1 Calculating Assortativity Coefficients

As we mentioned earlier, the assortativity of a network depends on which node attribute is measured. Nodes in a network basically have two types of attributes: discrete and scalar attributes. A discrete attribute usually describes which category a node belongs to. A person's blood type is such a discrete attribute. By contrast, a scalar attribute is associated with a number and denotes the level, at which a node has certain property. One's age is such a scalar attribute and denotes how old one is. Also, the degree of nodes in a network is a topological scalar attribute. Newman studied degree-based topological assortative patterns [2] and found that human networks are often assortative, such as scholar co-authorship networks. However, technological and biological networks usually feature some disassortativity, such as the Internet and neural networks.

The assortativity for a network is measured by assortativity coefficients [2]. In an undirected network, the coefficient r for a discrete attribute (Equation (1)) is based on the mixing matrix. An element e_{ij} of the matrix denotes the percentage of edges between nodes with attribute value i and value j; $Tr(e) = \sum_j e_{ii}$ is the trace of the matrix; $a_i = \sum_j e_{ij}$ is the fraction of edges between nodes with attribute value i. $r = 1$ means a perfect assortative network and $r = 0$ means a network has no assortative patterns. For a disassortative network, $-1 \leq r \leq 0$.

$$r = [Tr(e) - \sum_i {a_i}^2)]/[1 - \sum_i {a_i}^2)] \tag{1}$$

The assortativity coefficient for a scalar attribute is quite similar. It is essentially the Pearson correlation coefficient of the attribute's value for all pair of nodes connected by edges. In Equation (2), $a_i = \sum_j e_{ij}$; $b_j = \sum_i e_{ij}$; σ_a and σ_b the standard deviations of a_i and b_j respectively. $r = 1$ indicates perfect assortativity and $r = -1$ stands for perfect disassortativity. The expected statistical error σ_r on the value of r can be obtained with the jackknife method [5].

$$r = [\sum_{ij} ij(e_{ij} - a_i b_j)]/\sigma_a \sigma_b \tag{2}$$

Next we will calculate assortativity coefficients for the three-dimensional network in the GlobalSympoNet. Our computational tool for the analysis was based on the Java-based JUNG framework [6], which provides a library of network analysis tools. At macro-level, we use the topological scalar attribute-the node degree. At micro-level, we use two discrete organizational attributes-the size and type of an organization. It is worth noting that organizational sizes can actually be a scalar attributes. However, most of our survey or interview respondents do not know the exact number of employees in their organizations. Thus we had to classify organization sizes into categories. Our respondents were asked to choose from micro (<20 full-time employees), small (21-50), medium (51-100), large (101-500), and very large (>500). As for organization type, United Nations (UN)

provide a classification scheme. Organizations in the GlobalSympoNet are classified into seven types: Academia, Donor, Government, Governmental organization, Inter-Governmental organization, Media, Non-governmental Organization (NGO), Networked NGO, UN organizations, and Private sector.

3.2 Preliminary Results and Analysis

Table 1 summarizes the assortativity coefficients and errors given by our computational analysis. Degree-based assortativity coefficients suggest that the inter-organizational network in the humanitarian sector exhibit disassortative patterns. This is different from the degree-based assortative patterns in a lot of social networks [2]. We believe the degree distributions of the networks may have contributed to the disassortativity. In the degree distributions of the three networks (Figures 1), there are often a large number of nodes with low degrees and a small number of nodes with high degrees. However, few nodes have medium degrees. This type of polarized distribution has led to the core-peripheral structures of the network. This structures suggest that several organizations are very active in this network, connect with many other organizations, and thus serve as the core or hub of the network. Meanwhile, most organizations have low degrees, mainly connect to high-degree nodes, and are relatively peripheral to the network. Among the three dimensions, the polarization in node degrees is especially obvious the funding network, which also has the highest degree-based disassortativity. In the funding network, 80% of the nodes have degrees lower than 5. The highest-degree node has 87 edges, while the second-highest-degree node has merely 25 edges.

The core-peripheral structure could potentially be explained by the nature of the community, in which several general-purpose humanitarian relief organizations, such as Red Cross, interact with many highly specialized organizations. Those specialized organizations include (1) humanitarian relief organizations that focus on a specific humanitarian relief domain, such as providing shelters or protecting children; (2) humanitarian relief organizations that work in a specific geographical area, such as Sub-Saharan Africa or Mideast; or (3) organizations that provide specialized IT services. For example, a humanitarian relief organization may seek advice or help from a partner with expertise in landmine-detection geographic information systems. Further, these specialized organizations are less likely to work with one another in the humanitarian relief sector. For instance, the chance is relatively low for an organization working mainly in East Europe to collaborate with another one that focuses on Latin America. Similarly, a

Table 1. Assortativity coefficients. Correpsonding errors are listed in paretheses.

	Degree-based coefficients	Size-based coefficients	Type-based coefficients
Advising Network	-0.3896 (0.0006)	-0.0196 (0.0005)	0.0009 (0.0003)
Collaboration network	-0.4293 (0.0005)	-0.0960 (0.0004)	-0.0105 (0.0004)
Funding network	-0.4459 (0.0003)	-0.1898 (0.0012)	-0.1176 (0.0006)

software provider may not need to interact with a telecommunication provider for the purpose of humanitarian relief.

The size-based assortativity coefficients also suggest disassortative patterns, which are similar to, but not as strong as, the degree-based disassortative patterns. This pattern means organizations tend to connect with other organizations with different sizes. We explain this pattern by taking the collaboration network as an example. We find that the average degrees of organizations in each size categories in the collaboration network are significantly different at .05 level. Micro, large, and very large organizations have much higher average degrees than small and medium organizations do. It means that these organizations tend to be active in inter-organizational collaborations and their connections have more influence on the assortativity patterns. Then why do they collaborate with organizations with different sizes? On one hand, some smaller organizations have limited resources, because they have relatively small number of full-time staff members. Instead, they rely heavily on contract workers and volunteers. Therefore, they need to reach out and work with those who have complementary resources. On the other hand, larger organizations often have more resources and experience. In addition, several large organizations have highly specialized departments, which may have initially be developed to serve internal needs, but then subsequently begin to offer these services to organizations, often smaller ones, that do not possess such capabilities.

Type-based assortativity of the network is slightly different from degree-based and size-based assortativity. Advising and collaboration networks have near-zero assortative coefficients and are close to random mixings, while the funding network is slightly disassortative. The disassortativity in the funding network is intuitive as the flow of funding often exists between two organizations of different types, such as donors and NGOs.

The lack of type-based disassortative patterns in the advising and collaboration networks is somewhat surprising. One might argue that the collaborative groups generally consist of organizations with different skills or expertise and hence would include heterogeneous types of organizations, combining donors, UN organizations, NGOs, together with for-profit firms in the private sector. However, it may be that more than half of all organizations in this symposium are NGOs or UN organizations. By contrast, there are not as many donors or media in the GlobalSympNet. In addition, the similar goals of NGOs enable them to work with each other more than with organizations of other types. Similarly, organizations from UN often have strong ties with each other and tend to interact with UN organizations. Such a tendency may also have diminished the disassortative patterns. Another possible reason is that types assigned to an organization by UN do not truly reflect the function-based contributions the organization is making either in terms of collaborative projects or advice. For example, a for-profit firm in the private sector may have a humanitarian response team that has little to do with its for-profit status overall. In this case, the role of this firm in the GlobalSympNet is not reflected by its assigned type as a private sector organization.

4 Discussions

Overall, the inter-organizational network we studied exhibits disassortative or random patterns on selected node attributes. That means organizations tend to interact with those that are not similar to themselves on those attributes. This is quite different from social networks, which are often assortative, especially on node degrees. We believe this pattern shows that, in the humanitarian relief sector, inter-organizational relationships are often based on complementarities. Similar to firms looking for complementarity from alliance partners [7], humanitarian relief organizations interact with others, because they are seeking complementary resources or expertise that their own organizations do not possess. Among the three dimensions we studied, the funding network is most disassortative, because funding relationships need the highest level of complementarities. If an organization already possesses enough resources for a specific humanitarian project, the chance that it shares the funding with other similar organizations is relatively low. Among the three node attributes, degree leads to the most disassortative pattern, which implies that nodes with different numbers of edges tend to possess different resources. As we mentioned in Section 3.2, nodes with low degrees are often specialized organizations. Their knowledge and expertise on a specific mission, a specific geographic region, or specific technologies are usually what some high-degree general-purpose organizations need.

Admittedly, the disassortative or random patterns we revealed do not necessarily mean that the inter-organizational network has no assortative patterns at all. We mentioned earlier that assortativity depends on which node attribute is chosen. Our choice of node attributes-degree, size, and type-is mainly based on our interests and data availability. Had we chosen some other attributes, we may have found different patterns. For instance, if we look at the focus region of those organizations, we may find an intuitive assortative pattern: organizations tend to interact with those who focus on similar geographical areas.

In addition, some of the assortativity patterns can be validated by our surveys and interviews. For example, according to our assortativity analysis, the advising network and the collaboration network have similar assortativity patterns. This similarity confirms with our interview outcome–advising relationships often serve as the basis and prerequisites for future collaborations. Another example is the near random type-based assortative patterns, which indicates that the GlobalSympNet is not very diverse in terms of member organization types and needs more donors and media members. This mirrors the result from our survey, in which 40% of the respondents consider “introduce new donors” very important for promoting collaborations.

5 Conclusions and Future Work

In this research, we use computational tools to explore assortativity patterns in multi-dimensional inter-organizational networks. Building assortativity on both topological and organizational attributes, we analyzed the three-dimensional

inter-organizational network inside a major humanitarian relief coordination body. The results suggest that organizations tend to connect with those who are not similar to themselves in terms of degree, size and type. We believe organizations' seeking of complementarity from partners leads to the disassortative and random patterns. This research does not only reveal assortativity patterns, but also improves our understanding of the humanitarian relief sector, such as the relationship between the advising and the collaboration networks. As a result, we are able to provide recommendations on how to improve inter-organizational collaboration, which will eventually benefit disaster victims.

There are several areas that we would like to address in the future. We plan to conduct more statistical analyses to find more evidence for the assortativity patterns revealed by the computational analysis. Collecting more data about organizations and exploring assortativity using other organizational attributes are helpful as well. After all, the assortativity patterns of a network depend on what node attributes are chosen.

Acknowledgments. This research has been supported by the U.S. National Science Foundation grant CMMI-0624219.

References

1. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual Review of Sociology 27(1), 415–444 (2001)
2. Newman, M.E.J.: Mixing patterns in networks. Physical Review E 67(2), 13 (2003)
3. Powell, W.W., White, D.R., Koput, K.W., Owen-Smith, J.: Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. The American Journal of Sociology 110(4), 1132–1205 (2005)
4. Zhao, K., Yen, J., Ngamassi, L.M., Maitland, C., Tapia, A.H.: Modeling emerging coalitions in the context of inter-organizational networks: A case study of humanitarian coordination. International Journal of Intelligent Control and Systems 14(1), 97–103 (2009)
5. Efron, B.: Computers and the theory of statistics: Thinking the unthinkable. Siam Review 21(4), 460–480 (1979)
6. O'Madadhain, J., Fisher, D., Nelson, T., White, S., Boey, Y.B.: The java universal network/graph framework (jung): A brief tour. In: Music-to-Knowledge North American Workshop, University of Illinois (2005)
7. Chung, S., Singh, H., Lee, K.: Complementarity, status similarity and social capital as drivers of alliance formation. Strategic Management Journal 21(1), 1–22 (2000)
8. Krackhardt, D.: Predicting with networks - nonparametric multiple-regression analysis of dyadic data. Social Networks 10(4), 359–381 (1988)

Deconstructing Interaction Dynamics in Knowledge Sharing Communities

Ablimit Aji and Eugene Agichtein

Mathematics and Computer Science Department
Emory University
{aaji,eugene}@mathcs.emory.edu

Abstract. Online knowledge sharing sites have recently exploded in popularity, and have began to play an important role in online information seeking. Unfortunately, many factors that influence the effectiveness of the information exchange in these communities are not well understood. This paper is an attempt to fill this gap by exploring the dynamics of information sharing in such sites - that is, identifying the factors that can explain how people respond to information requests. As a case study, we use Yahoo! Answers, one of the leading knowledge sharing portals on the web with millions of active participants. We follow the progress of thousands of questions, from posting until resolution. We examine contextual factors such as the topical area of the questions, as well as intrinsic factors of question wording, subjectivity, sentiment, and other characteristics that could influence how a community responds to an information request. Our findings could be useful for improving existing collaborative question answering systems, and for designing the next generation of knowledge sharing communities.

Keywords: Social media, Collaborative Question Answering.

1 Introduction

Asking questions and contributing answers can be an effective way of sharing information and expertise. Online, this way of knowledge sharing manifests itself in the form of Collaborative (or Community) Question Answering (CQA) sites, such as Naver, Baidu Knows, Live QnA, and Yahoo! Answers. In the U.S., Yahoo! Answers attracted more than 100 million users, and has a growing archive of more than 400 million answers to questions (2008 estimates). Already, for many information needs, these sites are becoming valuable alternatives to search engines.

Previous studies of CQA focused on the ultimate outcome of the knowledge sharing activity (e.g., the quality of the finally contributed answers). However, as responses and ratings arrive, the perceived quality of an item may change significantly. Our goal is to understand the factors that influence the *dynamics* of the interactions in knowledge sharing communities - that is, to understand *how* this content is generated, and which aspects of the process can affect the resulting content quality. As far as we know, there has been no prior scientific study of how different question characteristics influence the interaction dynamics in CQA. We attempt to fill this gap by systematically exploring the contextual and intrinsic characteristics of the questions posted in CQA, and the effects of these factors. Specifically, we address the following research questions:

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 273–281, 2010.

- *How does the question context influence the community response?*
- *What intrinsic question qualities influence the community response?* For this, we examine how the question aspects such as subjectivity, sentiment, conversational orientation and writing quality affect the dynamics of community responses.
- *How do question context, wording, and response dynamics effect the quality and timeliness of the answers obtained?*

We present, to the best of our knowledge, the first large-scale empirical study of temporal dynamics for collaborative question answering to explore the factors that can explain the CQA interactions.

In addition to better understanding the dynamics and behavior of collaborative question answering communities, our results could have practical applications including better real-time CQA content ranking for search engines, and more accurate and timely filtering for high quality content. More generally, our findings can be useful to improve existing collaborative question answering systems, and for designing the next generation of knowledge sharing communities.

2 CQA Overview and Description

We first briefly review the Community Question Answering setting. A user (asker) selects an appropriate topical category, and posts a question. Newly posted questions appear in the "Open Question" list for each category, reverse ordered by the time they posted. At this point, other users can answer this question, or can rate the already posted answers. If the asker is satisfied with any one of the submitted answers, he or she can choose it as "best" answer. If the asker has not closed the question during the "Open Question" period, the "best" response chosen by votes from other users.

2.1 Data Collection and Statistics

To obtain the data for our study, we repeatedly crawled the Yahoo! Answers using the provided API, to capture the arrival of user contributions nearly real time. For this, we tracked a total of approximately 10,000 questions, sampled from 20 categories, over each of the questions' lifetimes. Specifically, for each category, after a new question appears on the "Open Questions" list, we begin tracking it (up to 200 questions per run per category) every five minutes until the question is closed or moved to the "Undecided Question" status. As a result, we obtained approximately 22 million question-answer-feedback snapshots in total, capturing feedback and contribution dynamics for a representative sample of CQA. We refer to this dataset as *AnswersTemporal*. Ultimately (after filtering and cleaning), the *AnswersTemporal* dataset consists of the interaction data for 9,747 questions and for the 47,780 answers posted for these questions.

2.2 Temporal Dynamics of CQA

In popular CQA sites, most questions stay on the front page of the respective category for only a few minutes. These questions tend to receive many answers initially, but the rate of arrivals of answers and user ratings decreases over time (Fig 1(a)). In contrast,

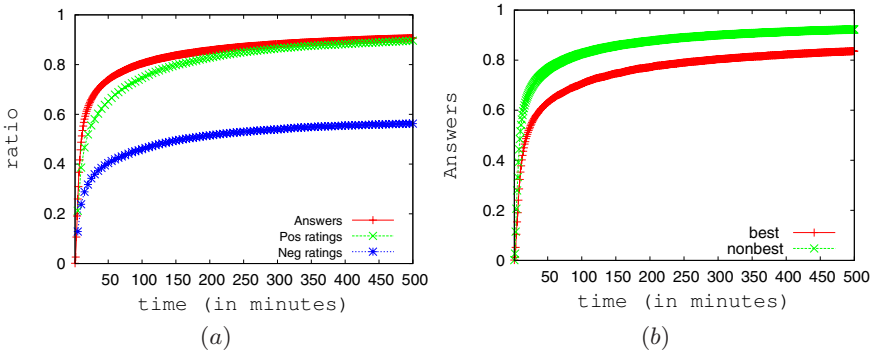


Fig. 1. Answer and rating arrival time, averaged across all categories, for the first 500 minutes of question lifetime (a), and arrival rate for "Best" vs. "Non-Best" answers for the first 500 minutes after posting (b)

user feedback ratings continue to arrive hours or even days after the question and answers have been posted. We conjecture that as the questions and answers are indexed by the web search engines, additional users view the questions and the answers, but tend to rate the content instead of contributing new answers. Interestingly, answers eventually chosen as "Best" by the asker appear to arrive somewhat later than non-best answers (Figure 1 (b)). That is, 56% of the eventual "Best" answers arrive within first 30 minutes, compared to 70% for Non-Best answers (a statistically significant difference over our large sample of questions). Having discussed the overall patterns of answer arrival, we now delve into more detailed analysis of the *content* of the questions.

3 Factors behind CQA Dynamics

In this section we describe the factors that could influence interaction dynamics, namely the question context, the intrinsic characteristics of the questions, and answer quality.

Question Context: Yahoo! Answers has 26 top level categories and more than 1000 lower level subcategories. These categories differs from each other in topical focus and community demographics. Based on previous work [1], we expect question category (and more generally, a forum chosen) to have significant effects on content creation speed and quality.

Intrinsic Question Characteristics: We now define the intrinsic question dimensions, drawn from the literature to be important aspects of textual communication.

– **Subjectivity:** CQA sites increasingly attract users interested in obtaining opinions of other users about social norms, preferences, and popularity. As has been suggested in previous work (e.g., [2] and references therein), question subjectivity can significantly affect the quality, quantity, and the tone of the answers posted by the community. Our hypothesis, is that question subjectivity will also influence interaction

dynamics such as the rate of answer and ratings arrival, and other factors such as agreement of the asker rating ("stars") with community ratings ("thumbs").

- **Conversational Orientation:** Another important dimension, distinct from subjectivity, is whether a question is a fact seeking information or is simply a start of a conversation. Note that conversational questions may in fact be objective (e.g., "What are the problems with [Windows] Vista?"), thus this factor is different from subjectivity [3]. Our hypothesis is that this dimension will also have a measurable effect on interactions and behavior of the community.
- **Sentiment:** An important feature that distinguishes CQA content from other web content is that both questions and answers can carry a sentimental orientation (that is, have a negative or positive connotation). For example, a question ("How stupid must you be to wear shorts and flipflops when it's cold?") makes a clear negative statement. We expect such questions to attract answers at different rates and of different quality than questions that are positive or neutral.
- **Quality:** Previous research showed strong correlation between question quality and answer quality [4]. We hypothesize that question quality would have an effect on the arrival of best answer (and popular answers in general) as questions that are well stated and interesting should attract good answers faster.

Answer Quality: Ultimately, the goal of CQA is to obtain information (answers) for the asker's information needs. Thus, quality and timeliness of answers obtained are perhaps the most important "output" variables of a CQA system. We also examine the effects of the contextual and intrinsic question factors above on the answer quality, as rated by both the asker and the community.

4 Experimental Setup

To analyze interaction dynamics, we first report on the large-scale manual labeling of question characteristics, and then describe the metrics used to analyze the responses.

4.1 Manual Labeling

To obtain human judgements, we utilized the Amazon Mechanical Turk (MTurk) service. Briefly, MTurk provides the infrastructure for "workers" to select Human Intelligence Tasks (HITs) that can be, for example, questions for which we wish to obtain human judgments regarding sentiment. Our typical HIT would include 10 questions, each with the text and the context (category) provided to the rater. Five workers rated each question for the various dimensions described above; the ratings were filtered by using majority opinion (that is, picking the label chosen by at least three out of the five Turk workers). From the 2000 initially labeled questions, 1570 remained after filtering, which we consider reliable human ratings. We refer to this dataset as *QuestionsLabeled*, and report the annotator agreement in Table 1. This dataset can be downloaded, by request, from `http://ir.mathcs.emory.edu/shared/sbp2010/`.

Table 1. Annotator agreement for question and answer labels

	Orientation	*Subjectivity*	*Question Quality*	*Sentiment*	*Answer Quality*
Agreement	0.7765	0.7396	0.701	0.647	0.682

4.2 Asker Satisfaction vs. Popularity

To complement the understanding the dimensions of the questions, it is important to examine the perceived quality of the contributed answers. Intuitively, the most important factor for an answer should be whether the asker considered it to be the best answer for the question. However, we observed that this selection does not always agree with the community ratings. To quantify this discrepancy, we used the traditional measure from information retrieval, Mean Reciprocal Rank (MRR), computed by ranking the answers in order of decreasing "Thumbs up" ratings, and identifying the rank of the actual "best" answer, as selected by the asker. More precisely:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{N}\frac{1}{rank_i}$$

where $rank_i$ is the rank of the best answer among the answers submitted for question i.

5 Results: Effects of Question Factors on CQA Dynamics

This section reports the effects of both contextual and intrinsic question characteristics on content creation, namely answer and rating arrival and answer quality.

5.1 Contextual Factors: Question Category

Question category influences many aspects of CQA dynamics, such as time to first answer, time to closing a question, and answer quality [4]. In particular, as Figure 2 shows, there is a large difference in answer arrival rates across categories. For example, in the top level category "Health," which is relatively popular, nearly 65% of the answers arrived in the first 10 minutes, whereas only 20.4% arrive within the first 10 minutes for the category "Travel".

Interestingly, the agreement between the asker and the community (modeled with the MRR metric defined above) also varies significantly across question categories, as shown in Table 2. In general MRR is higher for categories with shorter average thread length (which would make high MRR more likely), but is not always the case (eg., Beauty & Style vs. Entertainment & Music). This indicates that in some categories community consensus is more difficult to achieve than in other categories, which would also indicate higher rate of subjective or conversational questions.

5.2 Effects of Intrinsic Question Factors

Question Subjectivity: We expect subjectivity of a question to be less dominant than category but still related to other important factors such as average answer arrival and

MRR. In particular, more answers arrive for subjective questions than for objective questions (Figure 3 (a)), and user ratings exhibit a similar skew towards subjective questions (Figure 3 (d)). Interestingly, answer ratings for subjective question arrive faster than ratings to user answers to objective questions. This is more noticeable for positive user ratings.

Conversational Orientation: Answer arrival and rating arrival for questions with different conversational vs. informational orientations are reported in Figure 3 (b) and (e),

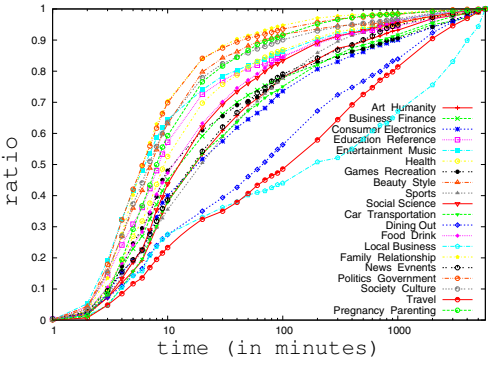


Fig. 2. Answer arrival patterns for categories

Table 2. Average MRR values (asker vs. community agreement) for selected categories

Category Name	MRR	Answer
Local Business	0.46	2.96
Yahoo Products	0.46	4.28
Travel	0.43	4.86
Science and Mathematics	0.41	4.27
Computer and Internet	0.40	4.11
Consumer Electronics	0.39	3.84
Entertainment and Music	0.33	9.74
Family and Relationship	0.30	8.32
Beauty and Style	0.30	6.67
Pregnancy and Parenting	0.29	9.06
Average	0.37	6.4

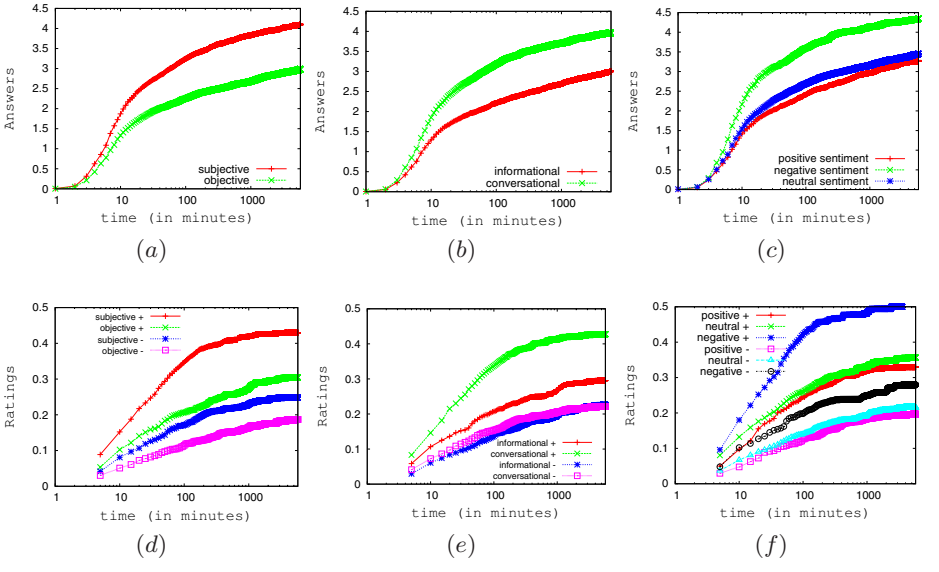


Fig. 3. Effects of Question Characteristics on Answer and Vote Dynamics: answer arrival (a) and answer ratings (d) grouped by subjectivity; answer arrival (b) and answer ratings (e) grouped by informational orientation; and answer arrival (c) and answer ratings (f) grouped by question sentiment

respectively. We observe that first answers for conversational questions arrive faster than for informational questions. This is not surprising since conversational questions are more common and users are willing to answer conversational questions compared to more difficult factual or informational ones. However, user ratings for informational questions suggest a different dynamic. Positive user ratings for conversational questions arrived faster than for informational questions, while *negative* user ratings for both conversational and informational questions arrive at the same rate.

Question Sentiment: Recall that we hypothesized that question's sentiment would affect the answer arrival rate as well as the ratings. Figure 3 reports answer arrival dynamics (c) and rating arrival dynamics (f) for these different question types. Answers to negative questions arrive significantly faster than answers to positive and neutral questions; interestingly, positive *ratings* arrive much faster to negative questions, whereas positive and negative ratings arrive roughly at the same rate for positive and neutral questions. Based on manual examination of the examples, we conjecture that this effect is caused by the selection bias of the raters participating in negative question threads, who tend to support answers that strongly agree (or strongly disagree) with the asker.

Answer Quality: We now consider the effects of all the question factors on answer quality. As reported in previous work [4], question quality indeed moderately correlates with answer quality (Pearson $p = 0.23$), and answer length correlates more strongly with answer quality (Pearson $p = 0.45$). However, we were surprised to find that there is weak or no correlation between question subjectivity, orientation, or sentiment and answer quality. That is, while conversational questions tend to elicit more participation from others, the overall quality of the contributed content does not exhibit significant differences (we omit the detailed results for lack of space). Interestingly, CQA participants respond differently to high quality vs. low quality answers (Figure 4). Surprisingly, there is no difference between the number of positive "thumbs up" ratings for high vs. low quality answers, but there *is* a significant difference in the number of

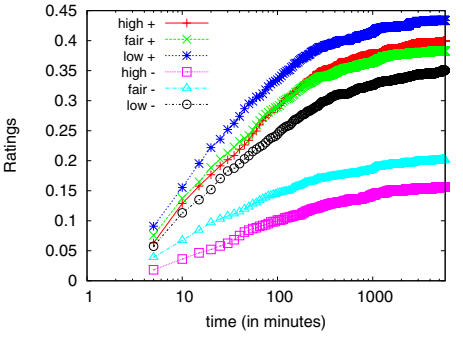


Fig. 4. Positive (+) and Negative (-) rating arrival for high, fair, and low-quality answers

Table 3. Agreement(MRR) of asker's choice of best answer with the community for different question types

Type	MRR	Answer
Informational	0.43	3.90
Conversational	0.39	4.80
Subjective	0.38	4.91
Objective	0.44	3.85
High Quality	0.40	4.37
Low Quality	0.42	4.52
Positive	0.41	4.10
Neutral	0.41	4.40
Negative	0.38	5.00

the negative "thumbs down" ratings: in other words, it is the negative ratings (or lack thereof) that can more reliably separate high quality answers from low-quality answers.

Asker vs. Community Agreement: Finally, we consider the Asker and community agreement for different types of questions (Table 3). The most noticeable differences confirm our hypotheses: askers of informational questions tend to agree with the most popular answers more often than conversational (MRR of 0.43 vs. 0.39), and askers of objective questions tend to agree with community more than askers of subjective questions (MRR 0.44 vs. MRR 0.38). Interestingly, question quality and question sentiment do not appear to have significant influence on asker's and community agreement.

6 Related Work

One of the main goals of CQA is to enable the exchange of high-quality, relevant information between community participants. Finding such quality information, where in QA communities quality varies significantly, provides a unique challenge, which recently has been addressed in references [4,5]. This previous work treated CQA content as static and no attempt was made to classify content while it is still being updated/rated.

References [6,7] introduced more fine-grained models of individual user actions generating content in blogs and other social media. Our study is also related to recent work by Harper et al. [3] which consider some of the similar features for automatic classification of questions in CQA into informational or conversational, but in a static, off-line setting. In contrast to these efforts, our work attempts to recover the underlying factors for content generation where we attempt to exploit factors such as question subjectivity. Other related work focuses on the temporal evolution of the social media or web graph structures, such as [8] that analyzes the temporal evolution of the wikipedia graph; [9] predicts controversy of Slashdot posts based on social network and discussion structure. Closest to our work, Adamic et al. [10] examined the category-centric variations in link structure in the Yahoo! Answers community. Additionally, Leskovec et al. [11] developed models for microscopic evolution, including one for Yahoo! Answers. In contrast, our work focuses on *understanding* the content generation dynamics, that appear to be more influenced by question characteristics than structural properties.

7 Conclusions

We presented the first, to our knowledge, large-scale study of the underlying factors influencing the dynamics of participation in knowledge sharing communities. In addition to confirming previous findings on a new dataset (e.g., the effects of each community or forum category), other more subtle characteristics such as the question sentiment, subjectivity, and informational orientation all influence the arrival of answers, ratings, and agreement between the "popular" answers and the ones chosen by the asker. These findings could be useful both for improving existing CQA systems and for designing the next generation of collaborative information sharing environments.

References

1. Liu, Y., Bian, J., Agichtein, E.: Predicting information seeker satisfaction in community question answering. In: SIGIR (2008)
2. Li, B., Liu, Y., Agichtein, E.: CoCQA: Co-Training over questions and answers with an application to predicting question subjectivity orientation. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNL (2008)
3. Harper, F.M., Moy, D., Konstan, J.A.: Facts or friends?: distinguishing informational and conversational questions in social qa sites. In: CHI 2009: Proceedings of the 27th international conference on Human factors in computing systems, pp. 759–768 (2009)
4. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high quality content in social media. In: Proc. of WSDM (2008)
5. Jeon, J., Croft, W., Lee, J., Park, S.: A framework to predict the quality of answers with non-textual features. In: Proc. of SIGIR (2006)
6. Hogg, T., Szabo, G.: Diversity of user activity and content quality in online communities. In: Proc. of ICWSM (2009)
7. Goetz, M., Leskovec, J., Mcglohon, M., Faloutsos, C.: Modeling blog dynamics. In: Proc. of ICWSM (2009)
8. Buriol, L.S., Castillo, C., Donato, D., Leonardi, S., Millozzi, S.: Temporal analysis of the wikigraph. In: Proceedings of the International Conference on Web Intelligence, WI (2006)
9. Gómez, V., Kaltenbrunner, A., López, V.: Statistical analysis of the social network and discussion threads in slashdot. In: Proc. of WWW (2008)
10. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge sharing and yahoo answers: everyone knows something. In: WWW 2008: Proceeding of the 17th international conference on World Wide Web, pp. 665–674 (2008)
11. Leskovec, J., Backstrom, L., Kumar, R., Tomkins, A.: Microscopic evolution of social networks. In: Proceeding of KDD (2008)

Workings of Collective Intelligence within Open Source Communities

Everett Stiles and Xiaohui Cui

Electrical Engineering and Computer Science Department,
University of Tennessee,
Knoxville, TN 37996
estiles@utk.edu
Computational Sciences and Engineering Division,
Oak Ridge National Laboratory,
Oak Ridge, TN 37831-6085
cuix@ornl.gov

Abstract. Open source communities have been of great interest for researchers recently, yet little can be agreed upon when it comes to developers motives. While it has been shown that participants are mostly driven to contribute based on work related needs, it has also been shown that they contribute to fulfill an ideological purpose. We believe that the majority of participants contribute to satisfy their own personal goals. We reveal how developers function as a collective intelligence by modeling the open source community as a disjoint group of contributors. We show that most developers contribute to only one project and only to a small portion of its source code. We demonstrated that useful functionality of most OSS software is an emergent phenomenon created by a collection of developers with different motivations and personal goals.

Keywords: open source software, developers, collective intelligence, sourceforge, concurrent versions system.

1 Introduction

It has been proposed that open source software is heavily reliant on the community that developed it [2]. Without the collective effort, social interactions, and group influences of open source communities, the development and adoption of said software would not occur [2]. The view of developers laboring together to achieve a common purpose is an attractive, if unproven, assumption. Previous research has revealed that the population of open source developers is disjoint with the largest cluster consisting of only 25% of the total number of developers [12]. Even though their conclusions will be shown as incorrect, they did present evidence that suggested that the open source community does not function as a whole. We wish to address this inaccurate portrayal of the open source community directly and demonstrate, using the popular open source development website SourceForge, that virtually all developers within the open source community work independent of one another.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 282–289, 2010.

Our hypothesis is the open source community and their achievements can be viewed as a collective intelligence model. Developers spend their time attempting to achieve their own goals, while their contributions collect and emerge as a related collection of useful functionality that can compete with peripheral software. A few developers, which we have named long term contributors, bind this functionality into a marketable application and do their best to maintain it. The open source software community is the collective workings of individual developers with different motivations and personal goals. In this paper, we presented our research results to demonstrate that this collective intelligence model is a reasonably accurate portrayal of the open source community thereby, correcting the held assumption that it is the open source communitys intention to labor together to achieve a common ideological goal.

2 Related Work

Open source communities have been researched greatly over the last couple of years, but not much research on developers motives has been agreed upon. Views that explain participation have ranged from the free software ideology [15] to enjoyment and creativity [4] [5] [16]. Other views that have been propagated include career concerns, learning, and reputation [7] [11]; affiliation and identity [8]; and the developers desire to fulfill personal needs [3] [9] [10] [13]. Although survey evidence exists to support each of these theories [6] [8], we feel that the majority of developers would be contributing in an attempt to achieve their own objectives.

In the article titled Motivation, Governance, and the Viability of Hybrid Forms in Open Source Software Development, Shah attempts to address these conflicting opinions by conducting interviews with participants from the open source community [14]. Virtually all interviewees reported that their initial contribution was motivated by a need brought on by a work related purpose. Open source software was used to fulfill their work related purpose due to the relative ease of licensing as opposed to contract negotiated closed source software. Their reason for contribution was usually caused by the absence of their required feature. Shah states, Because a need exists, participants generally do not wait for others to solve the problem. Developers cooperate with other individuals in order to avoid conflicts within projects, but tended to focus on their own needs. Pertaining to the survey evidence from previous research, Shah found that a small subset of original contributors continue contributing long after their initial needs are met. These long term contributors tended to be highly skilled software developers who held managerial positions in their companies of employment. Open source software development provided them with a venue to solve challenging problems while taking advantage of their creativity. Interviewees indicated that these long term contributors are the ones who commit code, rewrite sections, fix bugs, and design software releases. These contributors themselves reported that they monitor mailing lists, bug reports, and code contributions in order to receive feedback and improve the project's code.

The items that long term contributors are tasked with completing are the only visible signs that the open source community exists: committing code, rewriting sections, fixing bugs, designing software releases, maintaining mailing lists, reading bug reports and improving existing code. If only a small portion of the open source community consists of long term contributors and long term contributors produce the visible signs that a community exists then the majority of developers are short term contributors and do not partake in the communities objective.

3 Experiment

Our aim is to show that the open source community consists of individuals carrying out their own agendas. If shown, it would follow that another model, other than a common ideology, would best represent the motives and behavior of developers within the community. We highlight details that reveal most developers contribute to a single project, and only a small portion of a project is impacted by any developers contributions. The methods and results are outlined below.

4 Data and Methods

Our data collection focuses around SourceForge, the largest open source development website that hosts a large variety and number of open source projects. Our aim is to demonstrate that open source developers contributions are limited to a narrow scope within the open source community. We will address this in two stages. In the first stage we will use an external database, containing information collected from SourceForge over the last six years, to analyze developer participation at a community level. The second stage will encompass the use of SourceForge's Concurrent Versions System (CVS). Using this record we will examine developers contributions at a project level. Both stages of observation are discussed in more detail below.

The University of Notre Dame hosts an open source database (`http://www.nd.edu/~oss/Data/data.html`) which maintains a vast amount of data collected from SourceForge. The data collection started with a dump of SourceForge's content in January of 2003. Starting in February of 2005, this dump became monthly and consists of endless resources for research on the open source community. We will use this wealth of knowledge to demonstrate that most participants, within the open source community, only participate in a single open source project. We believe this would be a fundamental blow to the idea of a well connected social community and strengthen the proposal that participants our fulfilling their own independent needs. The databases information however is limited, and it cannot be used to accurately determine a single developers contributions to a given project. We use Concurrent Versions System (CVS) data to bridge this gap and determine the knowledge developers have of a project.

The Concurrent Versions System, also referred to as CVS, maintains the history of all changes within a project. This system, an open source project itself,

is used widely throughout the SourceForge community. Its purpose is to track changes within source code and record appropriate information that can used to revert back to previous versions. This information includes details on the change along with the author who committed it. By using the CVS historical record, we are able to get each individual developers contribution to a project at any given time by counting the lines of codes he/she had modified or developed at that time. We use these details to show that the knowledge individual developers have of a project is greatly limited. If this can be shown then it can also be assumed that developers lack the necessary knowledge to contribute towards a unified goal.

In our project, the CVS data was collected directly from SourceForge through the servers that they provide. Using their system and a standard CVS client, the entire source code repository for a number of selected projects was downloaded. Through further use of the CVS client, the projects history was extracted using a standard command named log. This history was used to create our results.

5 Results

5.1 Particpant's Community Contributions

Our first hypothesis is: most participants within the open source community contribute to a single project. In order to demonstrate this hypothesis, we collected data that reflected participation within open source projects. For each month that data was available, we calculated the percentage of developers assigned to one project, two projects, and so on. The data for each month was an accumulation from all previous months. We have presented these findings in Fig. 1.

An almost unchanging trend that persists from 2003 to present can be observed. Data points that reflect participation during the months of July to September 2007 are corrupted due to a privacy issue SourceForge had with the database that was used (`http://zerlot.cse.nd.edu/mediawiki/index.php?title=User_group`). The data demonstrates that over 75% of all developers in the history of SourceForge have taken part in no more than one project. It shows that out of the 246,114 developers reported to be participating in the month of July 2009, only 55,873 are participating in more than one project. However these results do include projects that have been inactive for some time. To receive results that more accurately describe the month being analyzed, the data was filtered each month to include only active projects. Active projects were defined as projects that had issued a release of their software within two years prior of the month being evaluated.

Fig. 2 contains the filtered project participation results for active projects. Similarly to the unfiltered data, the months of July to September 2007 are corrupted, but differences are clearly shown. Approximately 85% of developers on active projects contribute to one project with less than 4% of developers contributing to more than three. A linear trend can be extracted to show that the percentage of developers working on a single project is rising. Likewise, linear trends can show that participation in multiple projects is declining.

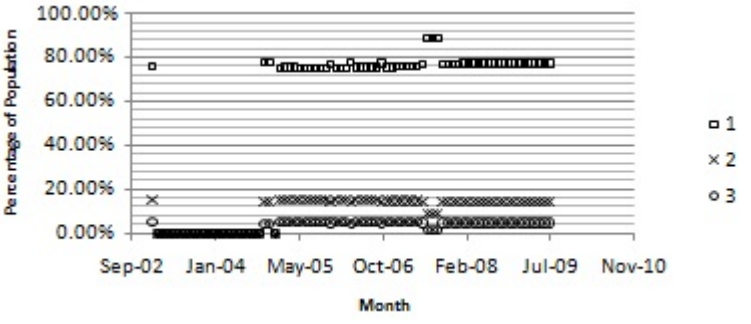


Fig. 1. Developer Participation on SourceForge

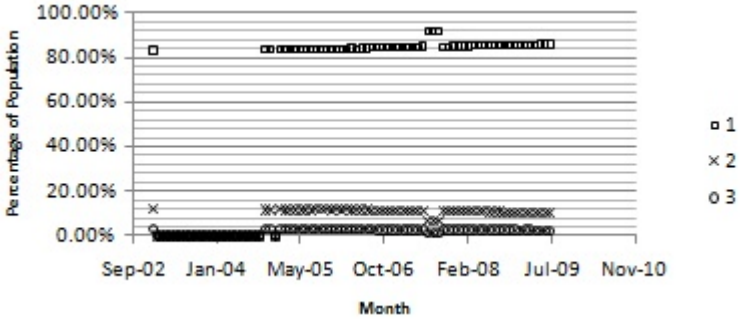


Fig. 2. Developer Participation within Active Projects on SourceForge

5.2 Particpant's Project Contributions

Our second hypothesis is: virtually all developers do not have knowledge of the entire project. This hypothesis can be proofed by examining each developers participation levels within the projects. To observe these levels, we measured the amount of knowledge participants had pertaining to projects in which they participated. SourceForge's Concurrent Versions System contained a repository of information that was able to perform this measurement. We analyzed projects with the highest amounts of participation during the month of July 2009. Out of the top 20 projects identified, 10 of these projects, including the top 4, were available to be analyzed using CVS repository. We measured the percentage of files that developers had made contributions to within a single project. A developer had to edit at least one line of a file to be counted as a contributor. Measurements are listed in Table 1. The number of participants for each project and the percentage of files developers contributed to are indicated.

Table 1 indicates that developers have a narrow scope of their projects. In some of these projects, the majority of developers have never contributed. Since there are relatively no project members participating in all coding efforts within a project, we can conclude that virtually all developers do not have knowledge of the entire project.

Table 1. Developer's Scope within Projects

Number of Participants	0%	0%-1%	1%-10%	10%-20%	20%-50%	50%-100%
84	61	16	4	2	0	1
93	0	70	20	0	3	0
94	56	22	10	3	1	0
103	34	40	20	3	6	0
112	28	57	20	7	0	0
129	12	83	28	4	2	0
151	77	53	20	1	0	0
162	73	64	24	1	0	0
249	68	136	44	1	0	0
428	204	184	32	4	3	1

6 Discussion

Past research [14] revealed virtually all participants in the open source community started contributing due to a work related need. This combined with the observation that nearly all participants discontinue their contributions within one year [14], we can assume that the majority of participants are fueled to contribute based on those needs. Our research results demonstrated that contributions from each developer were very narrow. Contributions from a single developer were mainly funneled into a single project and only impacted a small percentage of that project. This reinforces the notion that developers are driven to achieve their own objectives while contributing to the open source community.

Long term contributors, who likely started contributing through their own needs, continue to exercise authority over open source projects. Through code revisions and software releases, they guide the progress and future of open source projects. In our findings, we discovered some developers within the open source community that worked on a great number of projects and others that had contributed to a large portion of the project in which they were participating in. Determining if these ambitious developers are the same as the long term contributors, that maintain the community, should be an objective of future research.

With most participants working to achieve their own goals and long term contributors maintaining the projects, functioning software is produced. This conclusion proofed our hypothesis that the open source community software development phenomenon can be explained by modeling the community as a collective intelligence. All developers goals within a single project can be assumed to be related in some form, since they are all tied to the same project. As developers goals are fulfilled, the project will start to reach a functioning state. The long term contributors are the only ones needed to maintain that functioning state as the project grows with expanding functionality.

7 Conclusion

Our research has reinforced the idea that developers within the open source community are driven to contribute based on their personal needs. In this paper, we developed a collective intelligence model in order to explain how useful open source software can be created by developers who only contribute to satisfy personal needs. While this view does not hold for every contributor, it represents a large portion of the open source community. Based on our research presented in this paper, we have proofed that most participants within the open source community contribute to a single project and virtually all developers do not have knowledge of the entire project. The developers make contributions based on their personal needs, while their contributions collect and emerge as a related collection of useful functionality.

The scope of our data was limited to SourceForge, and it is still unknown if this community is the best sample to represent the open source community. Our analysis of project contributions could only be done on projects that made use of SourceForges CVS. Our sample did include fifty percent of the top 20 projects participated in, but this may be significant. The research presented in this paper did not demonstrate or highlight any differences among short term and long term contributors. Our aim was to give insight into how the majority of the open source community operates. Future research will aim at addressing these differences. Research on the open source community is still in its prime. We feel there are still benefits to be had even though we have learned a great deal. A self governing body, that produces useful society constructs, should be researched until we can recreate what nature has already done for us.

Acknowledgements

This research was done at Oak Ridge National Laboratory as part of the Department of Energys Internship program. Oak Ridge National Laboratory is managed by UT-Battelle LLC for the US Department of Energy under contract number DEAC05-00OR22725. This work was supported in part by the Lockheed Martin Corporation Shared Vision fund and Oak Ridge National Laboratory Seed Money fund (3210-2276). 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 Oak Ridge National Laboratory, the Department of Energy or the United States government. This manuscript has been authored by UT-Battelle, LLC, under contract DEAC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

1. Notre Dame OSS Database (2009), http://www.nd.edu/~oss/Data/data.html
2. Bagozzi, R.P., Dholakia, U.M.: Open source software user communities: A study of participation in Linux user groups. Management Science 52(7), 1099–1115 (2006)
3. Franke, N., von Hippel, E.: Satisfying heterogeneous user needs via innovation tool kits: The case of apache security software. Res. Policy 32, 1199–1215 (2003)
4. Gelernter, D.: Machine Beauty. Basic Books, New York (1998)
5. Ghosh, R.A.: First Monday interview with Linus Torvalds: What motivates free-software developers. First Monday 3(3) (1998), http://www.rstmonday.org/issues/issue3_3/torvalds/
6. Ghosh, R.A., Glott, R., Krieger, B., Robles, G.: Free/libre and open source software: Survey and study. Report, International Institute of Infonomics, University of Maastricht, Maastricht, The Netherlands (2002)
7. Hann, I.H., Roberts, J., Slaughter, S., Fielding, R.: Delayed returns to open source participation: An empirical analysis of the Apache HTTP Server Project. Working paper. Carnegie Mellon University, Pittsburgh, PA (2002)
8. Hertel, G., Niedner, S., Hermann, S.: Motivation of software developers in open source projects: An Internet based survey of contributors to the linux kernel. Res. Policy 32, 1159–1177 (2003)
9. Kuan, J.: Open source software as consumer integration into production. Working paper, SSRN (2001), http://ssrn.com/
10. Lakhani, K., von Hippel, E.: How open source software works: Free user to user assistance. Res. Policy 32(6), 923–943 (2003)
11. Lerner, J., Tirole, J.: The simple economics of open source. J. Indust. Econom. 52(6), 197–234 (2002)
12. Madey, G., Freeh, V., Tynan, R.: The Open Source Software Development Phenomenon: An Analysis based on Social Network Theory. In: Proceedings of the Americas Conference on Information Systems (AMCIS 2002), Dallas, TX, pp. 1806–1813 (2002)
13. Raymond, E.: The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. OReilly & Associates, Sebastopol (1999)
14. Shah, S.K.: Motivation, governance, and the viability of hybrid forms in open sources software development. Management Science 52, 1000–1014 (2006)
15. Stallman, R.: Philosophy of the GNU project (2001), http://www.gnu.org/philosophy/
16. Weizenbaum, J.: Computer Power and Human Reason: From Judgment to Calculation. W. H. Freeman, San Francisco (1976)

Manipulation as a Security Mechanism in Sensor Networks*

Ruiyi Zhang[1], Johnson P. Thomas[1], Qiurui Zhu[1], and Mathews Thomas[2]

[1] Department of Computer Science, Oklahoma State University, USA
[2] IBM Global Solution Center, Coppell, Texas, USA

Abstract. Manipulation is used as a behavior mechanism to achieve a certain goal. In this paper we propose a framework for manipulation in sensor networks where an attacker is manipulated with a long term objective in view. In particular an approach for selecting manipulation nodes is presented. Results show that the proposed scheme improves the security of the network by allowing more of the network to operate normally even when under attack. A manipulation scheme for a Denial of Service attack is proposed. A prediction based scheme is used to drive the manipulation.

Keywords: Manipulation, sensor networks.

1 Introduction

Research on computer and information security has focused on protective mechanisms such as encryption and key management for authentication or secure routing as well as anonymity. Although such protection is essential, security research must embrace a broader scope for a variety of reasons. For example what happens if the protective mechanism is broken, is there a last line of defense? Moreover, protective security mechanisms will not help in tracing an attacker or in learning attacker behavior and so on. Manipulation may be used as an added layer of security to complement the existing techniques. Manipulation is used as a behavior mechanism to achieve a certain goal. In the social context, people manipulate others by using deception, emotional manipulation, offering short term incentives to achieve a longer term goal, exerting control and so on. In this paper we propose manipulation as a mechanism for improving the security of a networked information system. In particular we propose a manipulation framework for a sensor network that is under attack. Manipulation can be highly effective against certain kind of attacks because information systems are generally thought to be honest. Although manipulation introduces more complexity, the benefits make it appealing. Manipulation provides an opportunity for the defender to actively manipulate the attacker's behavior; It also gives one more level of protection. If the protective mechanisms have been broken, manipulation provides a last line of defense. Even if the protective mechanisms have not been broken, manipulation can be used to achieve a desired goal. Manipulation

* This work is supported in part by United States DoD Army Research Office grant No. # W911 NF 051 0285 and the State of Oklahoma.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 290–297, 2010.

possibly increases the attacker's uncertainty. It increases the sophistication required for attack and may even thwart an attack and it can exhaust attacker resources;

In this paper we assume the attacker is intelligent. A lot of research has been done in cryptography systems including 3DES (Triple DES), RC5, AES [1]. Eschenauer and Gligor propose a key pre-distribution scheme [2] that relies on probabilistic key sharing among nodes within the sensor network. Further enhancements have been proposed in [3][4]. The LEAP protocol [5] is based on the observation that no single security requirement accurately suites all types of communication in a wireless sensor network. Chan and Perrig [6] describe a mechanism for establishing a key between two sensor nodes that is based on the common trust of a third node somewhere within the sensor network. Secure mechanisms for broadcasting and multicasting have also been proposed [7]. [8] presents a hierarchical anonymous communication protocol that hides the location of nodes and obscures the correlation between event zones and data flow from snooping adversaries. As far as we are aware, no-one has proposed manipulation as a security mechanism in networks. The closest model to our approach is the Honeynet model [9] in which a fake but attractive network is built in order to mislead attackers as well as waste attackers' time and resources. After outlining the concepts behind sensor networks, we present the proposed manipulation framework.

1.1 Wireless Sensor Networks

Applications based on wireless sensor networks (WSNs) are finding their way into numerous important domains including transportation systems, precision agriculture, battlefield monitoring, disaster zone management, homeland security and healthcare to name a few. Types of devices included in WSNs are sensor nodes, base stations, and optionally cluster heads. Sensor nodes have extremely basic functionalities in terms of their interfaces and components. They usually consist of a processing unit with limited computational power and limited memory, sensing components, communication components (usually radio transceivers), and a power source usually in the form of a battery. Base stations possess more computational, energy and communication resources. They act as a gateway between sensor nodes and the end users. The cluster heads are generally more powerful than sensor nodes but weaker than the base stations. They usually perform data aggregation operations, security related functions, and many other functionalities which are not suitable to put into sensor nodes.

Fig. 1. Crossbow Mica 2 sensor node

2 Manipulation Framework

Manipulation is carried out by nodes that neighbour the attacker. Non-neighbouring nodes carry on with normal communications. A subset of the neighbouring nodes,

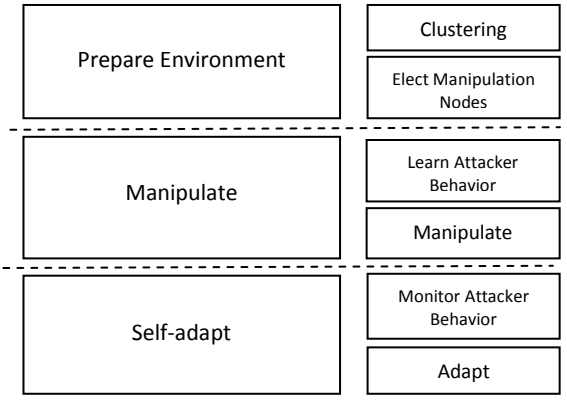


Fig. 2. Logical Framework of Manipulation Framework

called manipulation nodes focus exclusively on manipulating the attacker and perform no other function. Attacking activities are collected and analyzed at a central location. After the pattern of the attack is extracted, manipulation takes place.

If the manipulation matches a known pattern, the target is likely to follow the expectations of that pattern. The manipulation framework tries to meet the expectation of the target by mimicking a well known pattern by using minimal resources to meet the objective of the defender. For example, if the objective is to prevent the attacker moving to a different part of the network, the attacker is given what he expects (although it may not be correct information). This may induce the attacker to stay longer in the vicinity. There is a cost to the manipulation in terms of resources being diverted to manipulate as well as computational and communications overheads to learn the attacker behaviour and formulate an appropriate manipulating response. However, in the long run, the defender benefits as his objectives are met. We assume that an intrusion detection system will detect the type of attack. A lot of work has been done on intrusion detection [10] and is not discussed further here.

The attack we look at specifically is a denial of service attack. The attacker broadcasts requests to sensor nodes in the network. Each sensor can only communicate with its immediate neighbours. All the nodes including those that are multiple hops away from the attacker receive a request, respond to each request, thereby tying up the network to responding to the attacker's requests. The attacker is able to determine the success of his attacks by monitoring the rate of response to the request. Manipulation will be conducted in three steps (Fig. 2):

- Initiate environment: form clusters, elect manipulation nodes
- Respond to adversary: learn attacking pattern, generate responses
- Self adaptation: monitor behaviour of adversary, adapt to behavioural changes

2.1 Design and Architecture

The primary goal of the architecture is to provide an infrastructure that is energy efficient and enables the manipulation framework to respond to attacks in a timely fashion. Distributed Manipulation Agents (DMAs) act as cluster heads which are

specifically designed for the purpose of deceiving attackers. DMAs have stronger computational capability, larger storage capacity, longer communication range, and hardened security than ordinary nodes. DMAs are the localized center place for local data processing and decision making. A layered communication hierarchy is formed. One layer is the communication of sensor nodes and DMAs. The other layer is the DMAs and base station(s). The sensor field is divided into sub-zones. Each sub-zone has at least one DMA. The manipulation actions are executed by manipulation nodes. DMAs obtain an overview of sub-zones by surveying digests from individual nodes. A digest is simply a series of average values over a certain period of time which we call time windows. Each node keeps two digests of its interaction with an adversary. Digest $Digest_{in;i}$ records inbound requests coming from the adversary to node i, and the other digest $Digest_{iout;i}$ records outbound responses sent from node i to the adversary. Whether one or both digests are used by DMAs depends on requirements of the manipulation algorithm on DMAs.

2.2 Selection of Manipulation Nodes

Candidate manipulation nodes are nodes within transmission range of the attacker and can also communicate with a DMA. Manipulation nodes are the nodes selected from the candidate manipulation nodes. However, DMAs must not be able to communicate directly with the attacker, because each DMA has the responsibility of identifying manipulation nodes and this communication must be hidden from attackers. After selecting the manipulation nodes, the rest of the nodes in the network excluding the nodes in the attacking area will operate as normal. Other candidate manipulation nodes will be set to sleep. The DMAs decide how many manipulation nodes are needed to carry out the manipulation. They try to minimize the number of such nodes while ensuring that the chosen nodes are capable of fulfilling the manipulation mission. We assume the DMAs in conjunction with the base is able a priori to determine the appropriate manipulation approach. Hence the DMA is aware of the upper bound of the number of responses (or manipulation messages) an attacker expects in an attack (given by specific manipulation algorithm). As the DMA knows the manipulation scheme, it knows the number of manipulations nodes m needed.

3 Manipulation Approach - Pattern Learning and Prediction

We have a simple Denial of Servce (DoS) attack model here for each request by an attacker, each node sends a response. We assume the attacker request can only be transmitted k-hops before the request is dropped from the network. In our simple DoS model, the attacker sends requests and the defender replies with manipulation responses in a 1:1 mapping where each node sends a response for each request it receives. A multiple nodes are responding to a single request, the infected part of the network will become congested and a DoS attack takes place. The objective of manipulation in this instance is to keep the user attacking, that is, give the attacker what he expects. We assume the attacker is intelligent and has been active in the network for sometime and has therefore some expectation of the number of responses he should be getting. Given the density of the network is d, and depending on the

routing protocol, the total number of messages expected by the attacker for each request he sends is $\mu\pi r^2 d$ where μ is a constant. If each hop is a distance of *h,* then *r=hk.* The number of manipulation nodes *m* is therefore determined as $\frac{\mu\pi r^2 d}{c}$, where *n* is the number of responses (or manipulation packets) the attacker expects in response to a single request and *c* is the maximum number of packets a manipulation node can transmit in the same time period. Function *ReqToResp*( ) :

$$ReqToResp(\quad) = \frac{\mu\pi r^2 d}{c} \tag{1}$$

defines the manipulation response to an attack. Function *RespToReq*( ) is the inverse of function *ReqToResp*( ),that is, *RespToReq*( ) is the function to capture the effect of manipulation. If the requests from the attacker as a result of the manipulation response from the defender do not satisfy this function, then the manipulation is not successful In our simple case *RespToReq*( ) = 1, that is, the defender expects a request from the attacker over the next time period. A response may be doctored or fake data.

Although we have considered only a single request above, we consider the information received by a DMA is a streaming time series input. We model the manipulation system as a time series and integrate a time series prediction technology into our framework. Data flows in the network are modeled as a collection of time series $T = \{S_{ij}\}$ $(i, j \in \{1,\ldots,n\})$ where $\{1,\ldots,n\}$ denotes sensor nodes (including DMAs). Data flows received by DMAs are also modeled as a collection of time series $D = \{X_k\}$ where *k* denotes DMA *k*. *D* is a subset of *T*. Hence the time series of an individual DMA is $X_k = \sum_l S_{lk}$ where *l* denotes neighbors of the DMA.

The application's specific payloads (e.g. temperature, vibrations, etc.) carried by the data flow can be modeled as a time series. This information (or logical information) represents application domain knowledge. Similarly, physical characteristics (e.g. transmission rate) of the network can also be modeled as time series. They are independent to the application domain and closely related to the underlying network structures. We name them physical information. Whether logical information or physical information are collected depends on the needs of the specific manipulation. As a proof of concept, we only consider manipulations in which only one property of the network, be it logical information or physical information, is involved. A manipulation system with multiple properties can be done by using a collective time series or multiple independent time series.

In our validation case only one DMA will be operating in a sub-zone in which an attacker is present. Analyzing the operation of a single DMA suffices to demonstrate the capability of our manipulation approach. The time series on a single DMA is: $X = \{x_1, x_2, \ldots, x_n, x_{n+1}, \ldots\}$ where *X* is an infinite time series. Subscripts $\{1,\ldots, n,\ldots\}$ denote the time point of each value. Considering the network as a black box, we have a relation between T_{req} and T_{resp}, where T_{req} and T_{resp} are collective time series of requests and responses respectively. Therefore: $ReqToResp(T_{req}) = T_{resp}$. Here for each point in T_{req} in the time series, the number of points in the time series T_{resp} is as in eq. (2). Similarly $RespToReq(T_{resp})$ is the inverse function of $ReqToResp(T_{req})$. A prediction of future input from the attacker is:

$$Predict(T_{req}) = Predict\left(RespToReq(T_{resp})\right) = RespToReq(Predict(T_{resp})) \quad (2)$$

Hence, by predicting or planning appropriate future manipulation responses, the future requests or manipulated behavior of the attacker are well reflected or predicted.

3.1 Validation

A training set is a set of time series generated by the same physical system in different time periods, that is, training set $TS = \{X_1,\ldots,X_N\}$ where $X_i = \{x_{it}, x_{i(t+1)},\ldots, x_{i(t+l)}\}$. x_t is the value of the time series at time t, and l is the length of the time series X_i. A query sequence is provided to be predicted upon with. Therefore, the problem of predicting future manipulation responses becomes: given a training set TS and a query sequence $Q = \{q_1, q_2,\ldots, q_n\}$, predict $q_{n+1}, q_{n+2},\ldots$ and so on. A simple form of this prediction problem called *Predict – 1* problem is that there is only one training sequence in the training set, which is also the query sequence.

3.1.1 Fractal-Based Delay Coordinate Embedding Prediction

We expect the ideal prediction algorithm to be able to take an infinite streaming time series input and produce a streaming output. Due to limited resources and the large input data set, the parameter learning algorithm has to be a one pass algorithm. During the initialization, the algorithm finds parameters by itself without manual setup. A time series prediction approach using a forecasting method called F4 (Fractal FOREcasting) [11] provides automatic methods to do parameter induction, without human intervention. Hence, given any time series the optimal parameters can be automatically determined to build a prediction system.

The attacking traffic flow is simulated as a time series. We simplified the translation function to $Y_t = X_t$ where X_t represents the requesting time series and Y_t is the responding time series. This means there is only one node in the network (excluding the attacker), and therefore only one responding or manipulation node. However, eq.(2) can be applied to cater for more than one node in the network. In this particular case multiple nodes attacked can be modeled by simply scaling as defined by eq.(2). The latency caused by asynchronized communications between nodes introduces a shift of value among time points. Thus the resulting time series exhibits a noise-like property. Two types of attacking traffic are simulated. A constant rate attacking traffic flow represents a fixed strength of attack where the attacker finds an equilibrium. The other type of traffic is periodical, specifically a sine function against time points. The simulation is written in C, Matlab, and glued up with Perl scripts.

Fig. 3 shows the result under constant rate attacking traffic flow. The attacking traffic flow is simply a time series of $X_t = 10$ (Fig 3(a)). The first 1000 time points of responses (Fig. 3(b)) are used as a training set to train the manipulation algorithm. Fig. 3(c) shows the observed responses of size 1000 following the first 1000 time points. Fig. 3(d) shows the predicted responses produced by the manipulation algorithm. The equation used to generate the attacking periodic traffic flow is $X_t = 10 + sin(t/30)$. The manipulation process is similar to that of the constant rate attack. Table 1, shows the Normalized Root Mean Square Deviation (NRMSD) and Mean difference of observed sequence and predicted sequence. The mean difference gives a sense of how good the predicted sequence compares with the observed sequence

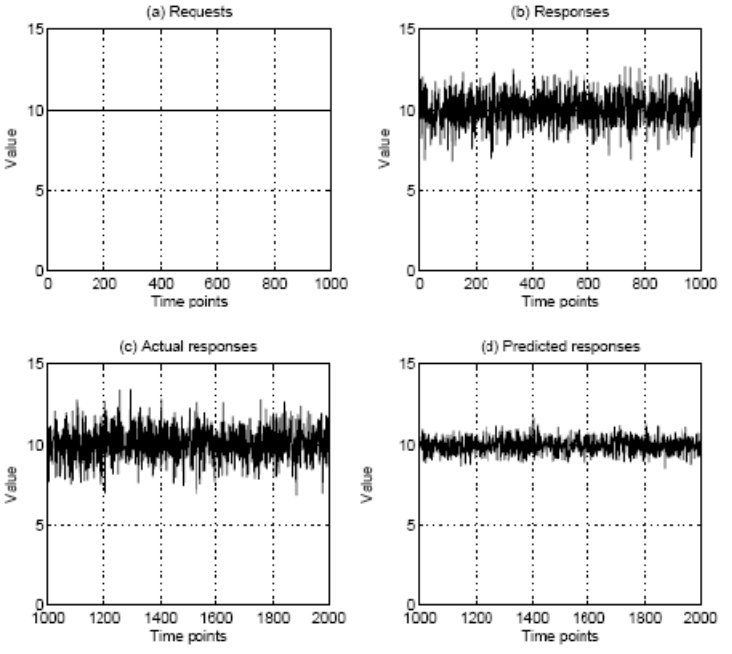


Fig. 3. Results of Deceptive Responses

Table 1. Quality of manipulation responses

	NRMSD	Mean (Observed)	Mean (Predicted)	Mean(Difference)
CONST	0.1682	10.0272	9.9416	0.0856
SINE	0.1799	9.9342	9.8906	0.0436

quantitatively. The NRMSD shows the quality of prediction in terms of motif details. The difference of means is in the range of sub 0.1. As mentioned earlier, if there are multiple nodes attacked, then the response predictions have to be scaled up, resulting in a more complex response time series. The scaled response predictions are then used to predict the attacker requests as defined by eq.(2) and thereby measure the success of manipulation. Prediction algorithms such as F4 or simpler ones can be used to predict both attacker requests and defender manipulation responses.

4 Conclusions

In this paper we propose a framework for manipulation in sensor networks where the attacker is manipulated with a long term objective in view. This is very preliminary work and future work will focus on using games of manipulation to drive the manipulation. Manipulation schemes for different attacks need to be devised. The prediction approach needs to be simplified for sensor networks. Measuring the success of manipulation remains to be determined. Other components of the

manipulation system, such as self-adaptation also need further research. Limitations of the work such as the required intelligence levels of the attacker or capabilities of the intrusion detection system are also avenues for further investigation.

References

1. Schneier, B.: Applied Cryptography, 2nd edn. John Wiley and Sons, Chichester (1996)
2. Eschenauer, L., Gligor, V.D.: A key-management scheme for distributed sensor networks. In: Proceedings of the 9th ACM conference on Computer and communications security, pp. 41–47. ACM Press, New York (2002)
3. Du, W., Deng, J., Han, Y.S., Varshney, P.K.: A pairwise key pre-distribution scheme for wireless sensor networks. In: CCS 2003: Proceedings of the 10th ACM conference on Computer and communications security, pp. 42–51. ACM, New York (2003)
4. Liu, D., Ning, P., Li, R.: Establishing pairwise keys in distributed sensor networks. ACM Trans. Inf. Syst. Secur. 8(1), 41–77 (2005)
5. Zhu, S., Setia, S., Jajodia, S.: Leap: Efficient security mechanisms for large-scale distributed sensor networks. ACM Trans. Sen. Netw. 2(4), 500–528 (2006)
6. Haowen., A., Perrig, C.: Pike: peer intermediaries for key establishment in sensor networks. In: Proceedings IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 1, pp. 524–535 (2005)
7. Lazos, L., Poovendran, R.: Secure broadcast in energy-aware wireless sensor networks. In: Proceedings IEEE International Symposium on Advances in Wireless Communications ISWC 2002 (2002)
8. Durresi, A., Paruchuri, V., Durresi, M., Barolli, L.: A Hierarchical Anonymous Communication Protocol for Sensor Networks. In: Yang, L.T., Amamiya, M., Liu, Z., Guo, M., Rammig, F.J. (eds.) EUC 2005. LNCS, vol. 3824, pp. 1123–1132. Springer, Heidelberg (2005)
9. Rowe, N.: Designing good deceptions in defense of information systems. In: 20th Annual Computer Security Applications Conference (2004)
10. Hai, T.H., Huh, E.-N., Jo, M.: A lightweight intrusion detection framework for wireless sensor networks. In: Wireless Communications and Mobile Computing. Wiley, Chichester (2009)
11. Chakrabarti, D., Faloutsos, C.: F4: large-scale automated forecasting using fractals. In: CIKM 2002: Proceedings of the eleventh international conference on Information and knowledge management, pp. 2–9 (2002)

Modeling the Impact of Motivation, Personality, and Emotion on Social Behavior

Lynn C. Miller[1], Stephen J. Read[2], Wayne Zachary[3], and Andrew Rosoff[3]

[1] Annenberg School for Communication and Journalism, University of Southern California, Los Angeles, CA 90089-0281, USA
lmiller@usc.edu
[2] Dept. of Psychology, University of Southern California, Los Angeles, California 90089, USA
read@usc.edu
[3] CHI Systems, Inc., Suite 300, 1035 Virginia Drive, Ft. Washington, PA 19034, USA
{wzachary, arosoff}@chisystems.com

Abstract. Models seeking to predict human social behavior must contend with multiple sources of individual and group variability that underlie social behavior. One set of interrelated factors that strongly contribute to that variability -- motivations, personality, and emotions -- has been only minimally incorporated in previous computational models of social behavior. The Personality, Affect, Culture (PAC) framework is a theory-based computational model that addresses this gap. PAC is used to simulate social agents whose social behavior varies according to their personalities and emotions, which, in turn, vary according to their motivations and underlying motive control parameters. Examples involving disease spread and counter-insurgency operations show how PAC can be used to study behavioral variability in different social contexts.

1 Introduction

The fundamental problem underlying the prediction of human behavior is its variability and the complexity of that variability. Social and cognitive sciences have slowly made progress on decomposing the multiple, interacting factors that contribute to human behavioral variability in social contexts. Some sources of variability have been addressed in prior social behavior modeling and prediction (SBP), including:

- *forms of social organization* -- long-term forms such as societies, organizational institutions, and kin-based units constrain behavior in differing and interacting ways, as do more situational units such as crowds, fluid social networks, and purpose-based groups, (including political-military groups such as terrorist cells);
- *cultural constructs* – configurations of knowledge, reasoning patterns, beliefs and attitudes influence behavior in both surface- and deep-structural ways. These constructs range from broad ideational systems such as religion, political/economic ideology, and frameworks for attributing causality, to more constrained but often more visible constructs such as conventions for face-to-face interaction;
- *individual cognition* -- elements of individual experience, expertise and skills shape individual differences in behavior in a given social context.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 298–305, 2010.

Prior SBP modeling has often focused on one of these factors at a time, for example modeling behavior of crowds, but without consideration of cultural or individual factors. This ignores the role of individual differences *within* social contexts. Here, we explore one important set of factors, motives, personality traits, and emotions that can explain and predict such individual behavioral differences in social contexts. We first summarize the effects of these factors on social behavior and how those effects are being modeled in a specific SBP framework called PAC. We then use case examples from the spread of disease and counter-insurgency operations to show how changes in the distribution of patterns of motivation might affect social behavior.

2 Motivation, Emotion and Personality in Social Behavior

PAC is built on an explicit theory-based representation of motives, personality, and emotion drawn from personality, social psychology, and neuroscience:

1. **Motivational Systems of Social Individuals.** Social animals have specifically *social* motives, such as those for affiliation, cooperation, status, and avoiding rejection. Human social behavior, both at the individual and group level, cannot be understood without such motives. For example, behavior in epidemics depends on fear of death and concern for one's family and friends. Behavior in teams is influenced by need for status, desire to cooperate, and fear of rejection. As we argue in the case examples below, the behavior of individuals in the local population during counter-insurgency operations is complicated by conflicting motives for self-/family-preservation and economic/social gain. Similarly, the risky sexual behavior of many men who have sex with men (MSM) is probably strongly influenced by such motives as avoidance of intimacy and fear of social rejection. These examples highlight that the *dynamics* of the *interaction* of motives is key to understanding, modeling, and predicting social behavior.
2. **Personality and Individual Differences.** Personality can be defined as the persistent tendency to behave in a similar way, and personality traits – aggressiveness, shyness –label these tendencies. Social interaction, whether at the level of the dyad, or larger social groups, is influenced by the personality of individuals in the group. For example, a work group composed solely of dominant individuals would jockey for status and control, whereas a group of highly agreeable individuals might socialize. Neither would be productive. Optimal group behavior depends on the appropriate group composition for the task. Or at a higher level of scale, for example, the collective behavior of a group in response to an external threat, such as an epidemic, would probably depend on the proportion of individuals who are highly anxious and sensitive to physical threat. Extensive research by Miller and Read (1991; Read et al., 2010) has demonstrated how personality traits can be modeled computationally in terms of the interactions of persistent motives and motive control mechanisms in an individual. These motive control mechanisms influence relative sensitivities to potentially rewarding and punishing outcomes, and how many motives influence behavior (Read et al., 2010).
3. **Emotion.** Emotion is critical, both as an indicator of what is happening to motives and as a signaling device to others about aspects of the social and physical environment. Emotions and motivations are closely linked. First, emotions result from

what is happening (or might happen) to our motives and signal the state of our motives. Second, emotions often have motivational force. For example, anger provides a motivation to remove an obstacle. Third, emotions provide signals about other's emotional and motivational states, helping coordinate social interaction.

3 The PAC Framework

We have been developing a computational model of human social behavior called PAC (Personality, Affect and Culture), which can be used to create and simulate social agents whose social behavior varies according to their personalities and emotions, which, in turn, vary according to their individual motivations and motive control system characteristics. In creating and evolving PAC, we drew on literatures in motivation, personality, and the neurobiology of personality. To complement this framework and its dynamics, we defined a cognitively-based processing architecture that simulates the ways these general processes (i.e., motivation, personality, and emotion) dynamically operate in the context of a specific social situation, and generate the resultant social behavior of each individual. Architecturally, dynamics of human behavior are modeled by hierarchically structured motivational systems (Read et al, 2010), which are activated and moderated by two general levels of motivation control systems – an approach/avoidance level of motivation, plus a higher level control system that affects the number of motives that are active. The Approach motive control mechanism governs response to rewarding stimuli and parallels the broad trait of Extraversion, whereas the Avoidance motive control mechanism governs response to aversive stimuli and parallels the broad trait of Neuroticism, particularly anxiety and fearfulness. Each of these two broad motive control mechanisms influences the activation levels of more specific motives, such as affiliation, dominance, avoiding social rejection and avoiding physical harm. The behavior of specific motives is a joint function of general characteristics of the broad motivational system of which it is a part and its own specific parameters. The higher-level motive control mechanism operates on and moderates the activity of the Approach/Avoidance mechanisms through inhibitory processes, and has been characterized as a Disinhibition / Constraint system (Clark & Watson, 2008).

In developing our model we used both lexical analyses of trait terms (e.g., John, Naumann, & Soto, 2008) and work on the structure of different trait inventories (e.g., McCrae & Costa, 1999; Tellegen, & Waller, 2008). This provided information on the nature and structure of personality, especially on what is called the Big Five: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness to Experience. In addition, we drew on work on the potential neurobiological bases of human personality (e.g., Clark & Watson, 2008; Gray & McNaughton, 2000) that has identified major dimensions of temperament (e.g., Neuroticism, Extraversion) and possible biological bases. These dimensions have close parallels with four major dimensions of the Big Five: Extraversion, Neuroticism, Agreeableness, and Conscientiousness.

To model how specific traits depended on more specific motives we drew on goal-based models of traits (Miller & Read, 1991; Mischel & Shoda, 1995) that described how specific traits could be related to specific motives. We also relied on evolutionary analyses (e.g., Bugental, 2000) of the problems that all humans must solve, and

recent work on goal taxonomies (e.g., Chulef, Read, & Walsh, 2001) to help identify specific motivational systems that underlie traits, such as: mating, nurturance of young, affiliation, establishing dominance hierarchies, and avoiding social rejection.

A PAC social agent operates by perceiving and understanding its ongoing social situation, and using that understanding to derive its (local) action strategies based on its motive and personality, as well as to develop its emotional responses. The ongoing social understanding process recognizes situational affordances to pursue specific motivations, and is driven by the representation of social knowledge. In PAC, social knowledge is represented in an extensible set of story structures, which contain a representation of the affordances that different parts of a story structure provide for pursuit of specific motives. Miller and Read (1991) have long argued that story structures are fundamental to the representation of traits, as well as being central to how people understand social interaction.

A PAC agent can perceive and parse the physical and social environment in a limited way on its own, but the PAC framework is designed to allow the PAC agent's social processing and behavioral choices to 'piggy-back' onto a more conventional behavioral agent that can provide more detailed capabilities for environmental perception and action. Further, PAC is designed to interface with any of the widely used cognitive architectures such as SOAR, ACT-R, or COGNET/iGEN, if one needs greater cognitive sophistication. On the perceptual side, these more detailed behavioral agent capabilities capture the information that drives the situational understanding process for PAC. On the action side, the behavioral agent's action capabilities translate the action strategy of the PAC agent into specific behaviors that implement that strategy in the (simulated) environment.

The PAC framework supports the creation and execution of intelligent agents that can interact with other agents in virtual environments, thus both modeling and predicting aspects of social behavior. Further, PAC agents can operate on their own or be integrated as social/emotional processing components of other behavioral agents. Figure 1 depicts the PAC macro-architecture. More detailed aspects are discussed below (but see also Read et al (2006) and Zachary et al (2005) for more discussion).

Personality model (symbolic). The personality model in a PAC agent integrates its situational understanding with the baseline activations of its individual characteristics. These include its personality, defined by its baseline motives and its parameters for the Approach, Avoidance, and Disinhibition / Constraint motive control systems. Together, this personality gives the individual agent a persistent tendency to behave socially in a specific way. The baseline activation of a motive moderates the impact of an affordance on the activation of the corresponding motive. For example, an agent with a low baseline for pursuing dominance is less likely to recognize (or react to) situations that afford an opportunity to increase social status. The unfolding social situation and the person's response also result in changes to activations of the motivations. Changes in the motive activations are a central input to PAC framework's model of emotion (see below).

Story Processing. A story in PAC is represented as a series of interconnected Action Structures that specify such elements as the character/agent (WHO), the act-type (DOES-WHAT), the modality of action (HOW), and the setting (WHERE/WHEN). It also specifies the opportunities that different possible evolutions of the story afford

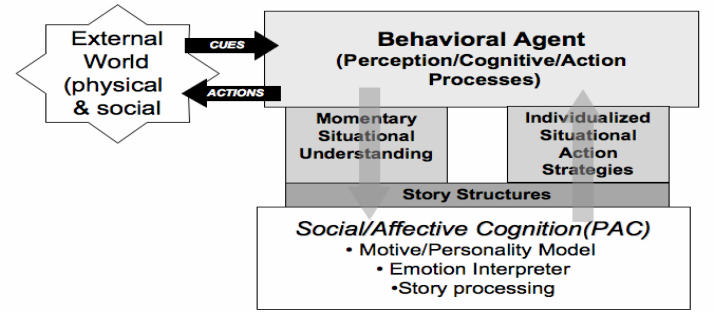


Fig. 1. PAC Macro-Architecture

for application of the motives in the PAC personality model. A PAC agent constantly monitors its understanding of the situation and tries to match possible other-agent behaviors with behaviors from action structures in its knowledge base. For example, if the agent has a high activation for the motive of seeking status and the situation affords the pursuit of status, than an action structure that facilitates increasing status might be favored over, say, another action structure that would afford an opportunity to help others. Each Action Structure contains information about its possible effects on the activation of specific motives. These motive implications are processed by the Motive Interpreter, which updates the activation of all motives to reflect their state at the end of the execution of this Action Structure.

Subsymbolic Personality Model. The story structures afford opportunities for a PAC agent to exhibit specific traits. For example, one part of the interaction may afford the opportunity for an assertive person to exert leadership, but simultaneously afford opportunity for an insecure person to be led. The personality mechanism within the PAC agent controls the process by which personality traits of the agent are exhibited.

PAC calculates the activations for each motive as each Action Structure is processed during an interaction. It operates on three types of data: (1) motive implications from the current action structure –indicating the relevance of an action structure to a specific motive. (2) individual motive baseline activations -- Each motive in PAC has a baseline activation representing the tendency of the individual to pursue that motive. (3) sensitivity levels associated with the Approach, Avoidance, and Disinhibition/Constraint systems. PAC selects the action or interpretation with the combination of motive implications that is most consistent with the current motive activations.

Model of Emotion. The PAC model of emotion is based on Roseman's (2001) appraisal model. In appraisal models, emotions result from appraisals of what happens to an individual's motives and goals, and why. Roseman makes two key distinctions: one between appetitive (Approach) and aversive (Avoidance) motives (i.e., what we want versus what we want to avoid) and one concerning what happens to those motives, (i.e., success or failure). Four possible combinations of motive relevant events can occur: getting what you want (appetitive, success), not getting what you want (appetitive, failure), not getting what you don't want (aversive, success) and getting what you don't want (aversive, failure). In PAC, the specific emotion experienced by

an agent is also a function of the: (1) unexpectedness of the event; (2) probability of the motive relevant event occurring; (3) agency: what or who caused the event? and (4) control potential: Can one do something about the motive relevant outcome? Emotions result from combinations of these factors. For example, sadness occurs when one has an appetitive motive, there is a motive inconsistent outcome, the outcome is certain and there is low control potential to change the outcome.

A PAC agent builds and maintains a set of action structure graphs, each of which represents a specific on-going story. The graph captures the current story state's action option, specifying the actor/agent, the act-type, the possible behavioral outcomes, and the opportunities that the action affords for activation or application of the individual's motivations. This graph also captures the path of past actions and interactions taken through the story structure, as well as the future possible evolutions of the story. Based on the action graph and agent's personality characteristics, PAC calculates continuous values of each appraisal dimension as well as the current activation level of the agent's motives, using those values to generate emotion levels continuously. New emotion levels are typically triggered by a change in the appraisal dimensions, and their value decays over time.

4 Two Examples Applying PAC to Social Behavior Predictions

Counter-insurgency. US and coalition forces engaged in counter-insurgency operations must engage the local populace and gain their support. This can be problematic because of differences in motivations and cultures of the two groups, as well as the personality differences within each group. Several PAC simulations were developed to model the factors involved in face-to-face transactions between the two groups in a typical context. The agents modeled were members of an Iraqi Arabic community and members of a (US) military patrol trying to gain information on a recent insurgent act of sabotage. The community residents– typically in shops or offices – had to deal with the patrol team based on their expectation about how such interactions should proceed, as well on their individual personalities and those of the members of the patrol team. For example, a shopkeeper highly motivated by fear for personal safety would react differently to an information request from an armed soldier entering his shop than would a shopkeeper highly motivated by the potential for material gain. On the other side, patrol leaders acting aggressively should engender more hostile responses than patrol leaders acting patiently. PAC agents with each of these motive structures were modeled and placed in simulated information-seeking interactions. One feature of these models was that the soldiers (Westerners) and the shopkeepers (Arabs) had different cultural variations of the story structures involved in this type of social transaction. The Arabs, for example, relied on story structures in which politeness and respect had to be established via preliminary interactions, while the Western soldiers, particularly Americans, relied on story structures based on directness.

A series of simulations reported in Zachary et al (2005) and Read et al. (2006) showed that within this interactional context, a wide range of possible behaviors, could be generated by varying the personality and emotional presets for the shopkeeper and soldier agents. The agents are currently being integrated as characters in a cultural learning simulation intended to help troops learn, pre-deployment, how to

deal with social situations involving local inhabitants and how to adapt to the different personality types and emotional reactions they are likely to encounter.

Social Behavior Contributors to Disease Spread. HIV/AIDS disproportionately threatens men who have sex with men (MSM): Although less than 7% of men, MSM account for 70% of all HIV transmissions among male adolescents and adults (CDC, 2005). Attachment Theory (Bowlby, 1979) provides a comprehensive framework for understanding the complex personality and interpersonal dynamics that underlie sexual decision-making. Bowlby argued that early interactions with caregivers (that did or did not reduce anxiety associated with threat) affected individuals' working models of self (loveable), others (trustworthy) and environments (threatening). This in turn biased baseline activation and sensitivity of attachment relevant motives. For example, if a caretaker is often inconsistent or rejecting the individual may develop a heightened fear of rejection and sensitivity to cues that signal the possibility of rejection. In fact, MSM who are more fearful take more behavioral risks associated with contracting HIV (e.g., higher rates of homelessness, daily substance abuse, participating in sex work) (Gwadz, et al., 2004). Miller et al. (2005) argue that for non-secure MSM, affectional pair-bond formation and maintenance is problematic, perpetuating a cycle of new sexual partners and increasing risk. Attachment models contain attachment relevant motives and goals (e.g., fear of rejection, desire for emotional closeness), and strategies for managing attachment related events. These aspects of the attachment system, as well as other critical components (e.g., emotion and beliefs) can be modeled in the PAC system. PAC multi-agent modeling could model the impact of population changes in rejection sensitivity on critical outcomes (e.g., partner number, type of sexual risk). This would help delineate what parameters might impact motives and emotions that might reduce the spread of HIV among MSM.

5 Conclusions and Future Directions

Prior applications of PAC have empirically demonstrated that PAC agents with identical domain and task knowledge but different motive and personality characteristics will behave differently in identical social contexts. Ultimately, we seek to create the capability to instantiate social agents with specific personality traits and motive structures so that the effects of these factors can systematically and effectively be predicted in social simulation models. Currently, research and application of PAC is ongoing on multiple fronts. Research directions range from expanding the theoretical framework to deal with motive and emotion dynamics in different time scales, to generalizing the PAC agent communications interfaces to support web-based simulation and game uses. Multiple applications are currently underway to use PAC as a realistic yet computationally light-weight social intelligence 'plug-in' to non-player characters in virtual environments for inter-personal skills (e.g., clinical communications, cultural familiarization) training and for mission rehearsal in military operations (e.g., for Stability, Security, Transition, and Reconstruction operations in counter-insurgency and peace-keeping missions).

References

Bowlby, J.: The Making and Breaking of Affectional Bonds. Tavistock, London (1979)

Bugental, D.B.: Acquisition of the algorithms of social life: A domain-based approach. Psychological Bulletin 126, 187–219 (2000)

Cassidy, J., Shaver, P.R. (eds.): Handbook of attachment: Theory, research, and clinical applications. Guilford Press, New York (1999)

Chulef, A., Read, S.J., Walsh, D.A.: A Hierarchical Taxonomy of Human Goals. Motivation and Emotion 25, 191–232 (2001)

Clark, L.A., Watson, D.: Temperament: An organizing paradigm for trait psychology. In: John, O.P., Robins, R.W., Pervin, L.A. (eds.) Handbook of Personality: Theory and Research, 3rd edn., pp. 265–286. Guilford Press, New York (2008)

Gray, J.A., McNaughton, N.: The neuropsychology of anxiety: An Enquiry into the functions of the septo-hippocampal system, 2nd edn. Oxford University Press, New York (2000)

Gwadz, M., Clatts, M., Leonard, N., Goldsamt, L.: Attachment style, childhood adversity, and behavioral risk among young men who have sex with men. Jrnl. of Adol. Hlth. 34, 402–413 (2004)

John, O.P., Naumann, L.P., Soto, C.J.: Paradigm shift to the integrative big five trait taxonomy. In: John, O.P., Robins, R.W., Pervin, L.A. (eds.) Handbook of Personality: Theory and Research, 3rd edn., pp. 114–158. Guilford Press, New York (2008)

McCrae, R.R., Costa, P.T.: A Five-Factor Theory of Personality. In: Pervin, L., John, O. (eds.) Hndkbk. of Personality: Theory and Research, 2nd edn., pp. 139–153. Guilford, NY (1999)

Miller, L.C., Read, S.J.: On the coherence of mental models of persons and relationships: A knowledge structure approach. In: Fletcher, G.J.O., Fincham, F. (eds.) Cognition in Close Relationships, pp. 69–99. Erlbaum, Hillsdale (1991)

Miller, L., Pedersen, W., Putcha-Bhagavatula, A.: Promiscuity in an evolved pair-bonding system: Mating within and outside the Pleistocene box. Behavioral & Brain Sci. 28, 290–291 (2005)

Mischel, W., Shoda, Y.: A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. Psych. Review 102, 246–268 (1995)

Read, S.J., Miller, L.C., Rosoff, A., Eilbert, J., Iordanov, V., Le Mentec, J.-C., Zachary, W.: Integrating Emotional Dynamics into the PAC Cognitive Architecture. In: Proceedings of the Annual Conference on Behavioral Representation in Modeling and Simulation, Baltimore, MD (2006)

Read, S., Monroe, B., Brownstein, A., Yang, Y., Chopra, G., Miller, L.C.: A Neural Network Model of the Structure and Dynamics of Human Personality. Psych. Review (2010) (in press)

Roseman, I.J.: A model of appraisal in the emotion system: Integrating theory, research, and applications. In: Scherer, K.R., Schorr, A., Johnstone, T. (eds.) Appraisal processes in emotion: Theory, methods, research, pp. 68–91. Oxford University Press, Oxford (2001)

Tellegen, A., Waller, N.G.: Exploring personality through test construction: Development of the Multi-dimensional Personality Questionnaire. In: Boyle, G.J., Matthews, G., Saklofske, D.H. (eds.) Handbook of personality theory and testing. Personality measurement and assessment, vol. II. Sage, London (2008)

Zachary, W., Le Mentec, J.-C., Miller, L.C., Read, S.J., Thomas-Meyers, G.: Human behavioral representations with realistic personality and cultural characteristics. In: Proc. Tenth Intnl. Command and Control Research and Technology Symposium. DoD CCRP, Wash. DC (2005)

Expressing Effects-Based Outcomes from Patterns of Emergent Population Behaviors

Colleen L. Phillips and Norman D. Geddes

Applied Systems Intelligence, Inc., Alpharetta, GA
(cphillips, ngeddes)@asinc.com

Abstract. There are basic social and cultural structures and processes that influence effects-based outcomes. When populations experience a miss-match of sociocultural factors that influence expectations and intentions, this can trigger an affective reaction – strong enough to measure and possibly forecast. This paper provides a framework for modeling the emergent behaviors between populations with conflicting values and intentions.

Keywords: Population Modeling, Stratagemical Behavior Patterns, Sociocultural Factors, Group Dynamics.

1 Introduction

There are basic social and cultural structures and processes that influence effects-based outcomes (http://obssr.od.nih.gov/pdf/OBSSR_Prospectus.pdf). When populations experience a miss-match of sociocultural factors that influence expectations and intentions, this can trigger an affective reaction – strong enough to measure and possibly forecast [1]. Sociocultural factors also play a role in shaping population perceptions of and responses to interventions, courses of action, or business plans and the impact of cultural values, beliefs, and intentions on populations' actions, reactions and well-being. In addition, sociocultural factors contribute to understanding societal and population processes such as current and changing rates of influencers that drive those decisive actions. There are opportunities for improving influence over a particular population through a better understanding of mechanisms linking the social and cultural environment to specific effects-based outcomes. To realize these opportunities, population modeling research related to sociocultural factors must be further developed and ultimately integrated into interdisciplinary, multi-level studies of groups of interest. Linking research from the macro-societal levels, through behavioral and psychological levels, to the science of population influence will provide the integrative sociocultural research necessary to fully understand the impact of human values, beliefs and intentions on decisions and actions [2].

Because the existence of racial/ethnic, social class, and rural-urban sociocultural disparities are influenced by behavioral and social factors, modeling and simulation environments for small group analysis may provide better knowledge of their specific causes and give insights to finding solutions [3]. Low sociocultural awareness results in analyst's inadequate engagement in decisions regarding population influencers and

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 306–313, 2010.

can hinder their ability to realize the benefits of effects-based outcomes. The most challenging problems in modeling small groups are so because they are complex in nature. These group dynamics problems have typically been approached using correlation based analytic methods (e.g., regression), which are useful for identifying linear relationships but are by themselves, insufficient, because of their inability to set up and test a web of causal relationships. Systems science methodologies provide a way to address complex problems by identifying and testing causal structures and theories, while taking into account the "big picture" and context of such problems.

The focus of this paper is to describe a methodology for expressing effects-based outcomes from patterns of emergent population behaviors. Section 2 describes a means for assessing population values, beliefs and intentions. Section 3 details the risk factors for which our approach would enhance understanding and decision making and contribute knowledge that will based on the prioritization of policies, interventions, and programs, especially where resources are limited and only a limited number of programs/policies/interventions can be implemented. Section 4 explains the methodology for meso-modeling of populations using the SOPATH Architecture for population modeling and Section 5 looks at validation issues and implications.

2 Meso-modeling and Sociopolitical/Ideological Reasoning

Meso-modeling falls between individual and societal models of human populations and into the small group category. The sociocultural factors influencing small groups can be seen to encompass three levels. The three levels of interdependent factors (Figure 1) - Individual Factors: (e.g. individual wants, wishes, preferences, values); Immediate environment: (e.g., family, friends, co-workers, rules); and Society and Culture: (e.g. religion, beliefs, social laws, legal laws, norms).

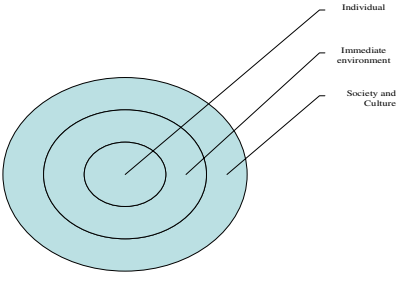


Fig. 1. The three interdependent sociocultural Factors

Sociocultural Factors. According to the Army Field Manuel, there are six main sociocultural factors ingrained in societal values and affecting the populace's opinion of operational effects as listed below [4]:

1. Society - The operational environment contains a populace under the same political authority with a common culture, region, and sense of identity.
2. Social Structure - The subunits or groups that have people distributed within those groups based on the role they play and their relations with each other in fulfilling their group's goals and the arrangement and rearrangement of these groups within the operational environment.
3. Culture - There exists a set of ideas, norms, rituals, codes of behavior that are shared, or coping mechanisms for the individual's world and for each other.

4. Language - Any set of symbols that people use to communicate.
5. Power and Authority - The ability to carry out one's own will even when up against resistance from the group, or outside the group.
6. Interests - The core beliefs and values that motivate our behaviors.

The relationship between these factors and various populations' values, beliefs, and intentions is forged by strong socio-political ideologies about health, wealth, and equity that each population holds and which takes many years to change [5]. Values have been strongly mapped to sociocultural behaviors observed within populations [6].

Population Beliefs. The sociocultural factors described above were used as multiple indicators of overall perceived levels of a group's belief about an observed event and under what conditions those beliefs hold true, how they change, and influence population's actions (Figure 2). Beliefs can be demographically-related, value-related, psychologically-related, and/or culturally-related. Bayesian Belief Nets were used to model and compute a group's belief factor and its corresponding classification level of that belief in relation to defined thresholds for a given population related to population attitudes, effects and perceived outcomes [7].

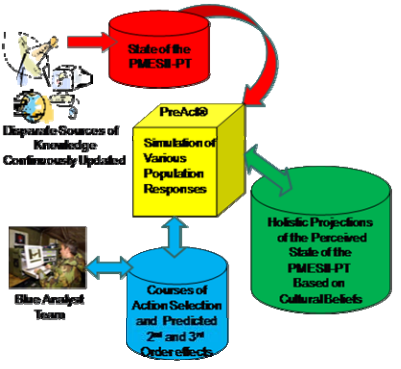


Fig. 2. How the influence of beliefs drive the courses of action taken

Reactive Activity Planning. In Figure 3, one can trace the path of beliefs, values, and goals from observation of various signals in the operational environment, through changing states of the world, to notifications of what has just happened and population reactions, to plans for simulation and execution. When this is done for multiple populations consecutively that have miss-matched expectations and perceptions, various emergent behaviors evolve. These behaviors are manifested into plans, goals, and scripts representing reports and actions of the populations.

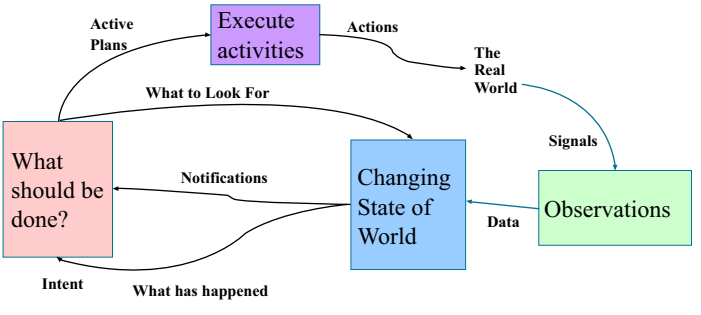


Fig. 3. A Reactive Planning loop for modeling populations that are observed, changes noticed, and activities planned and simulated for desired effects-based outcomes

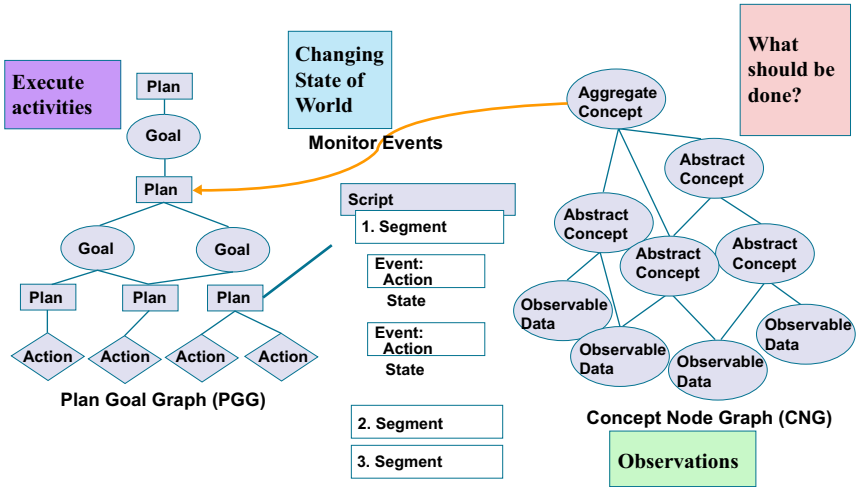


Fig. 4. PreAct® Intelligent Software Suite of concept, plan, and goal nodes with scripts

The representation of this population model has been embedded into PreAct® Intelligent Software Suite [8] depicted in Figure 4. Population observations and reports come into the system from the bottom right-hand corner. Embedded Bayesian Belief Nets abstract and aggregate the data until thresholds or conditions are met that signal changes in the state of the world which then decompose specific goals and plans into scripts which are actions and reports. These, in turn, feed back into the system as observations which may or may not change the state of the world (depends on everyone else's actions and beliefs). The figure shows the system for one population. There would be a different set of concepts, plans, and goals for each population being modeled. The entire architecture of simulation services, knowledge and databases, and interfaces are discussed in Section 4.

3 Patterns of Population Behaviors

Since most of the arrows in Figures 2 and 3 are bi-directional, the models are used iteratively to simulate and rank the results of various courses of actions based on the predicted 2nd and 3rd order cultural effects. The goal is to eliminate or reduce within acceptable measures the 2nd and 3rd order Cultural Effects (e.g. 1st order – raid village home and mistake child wielding stick for terrorist pointing gun and child was attacked; 2nd order – village elder hears of incident and holds US responsible for child's injuries and wages revenge; and 3rd order – populace now skeptical of US presence. Interventions proposed as courses of action include arranging a meeting with village elder to explain mistake and make amends (treat child's injuries with US medics and supplies) and propose a media campaign to calm local populace. The following six principles of influence can be seen as the basis for the population actions/reactions to the various courses of action or intervention used [9]:

1. Reciprocation – people have an inherent desire to return favors.
2. Commitment and Consistency – people's past decisions guide their future decisions.
3. Social Proof – people look to others and follow what they are doing.
4. Liking – people have less resistance to those who they like and are more easily persuaded.
5. Authority – people are more persuaded by those who possess authority.
6. Scarcity – people are more persuaded when the resource at hand is perceived to be limited.

Stratagemical Behavior Patterns (http://dictionary.reference.com/browse/stratagemical), defined as any artifice, ruse, or trick devised or used to attain a goal or to gain an advantage over an adversary or competitor, are formed when the various sociocultural factors are manipulated to control the operational environment or influence the population according to the appropriate combination of the influence principles. A group's intentions can be seen through the set of actions that are taken in response to a course of action or intervention that is taken to persuade or influence the group. This set of actions or use of stratagems is called stratagemical behavior. The set of behaviors to affect an influence are called stratagemical behavior patterns.

4 The SOPATH Architecture for Population Modeling

The SOPATH (service-oriented population analyzer for testing hypotheses) Architecture is comprised of the following components 1) component services, 2) a persistence database that holds data from the constructive simulation, 3) a graphical user interface for the analyst to interact with the simulation tool, and 4) the various Population Group Models (PGMs) of interest and can be seen in Figure 5.

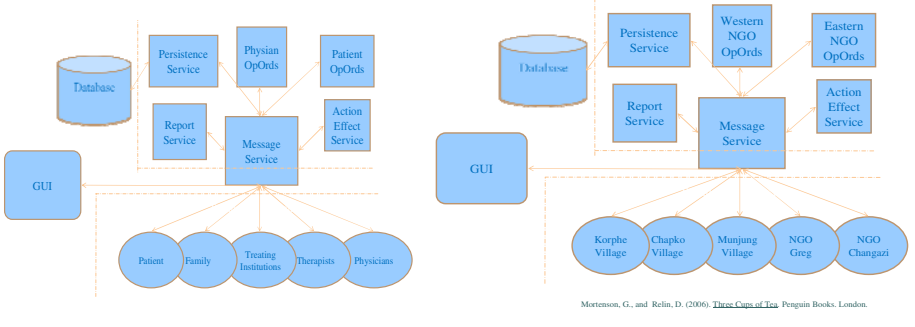


Fig. 5. a (left) SOPATH Architecture for expressing health-related and b (right) humanitarian-related effects-based outcomes based on various population beliefs, values, and goals

SOPATH Architecture provides the following major capabilities:

*A means to select and input a file containing the decision-maker's OPORD (operational orders) in temporal order and composed of plans, goals, and intentions based on values and beliefs.

*A means to create and initialize a suite of PreAct meso-models in which each instance of PreAct represents an instance of a group (Population Group Model, or PGM). Each PGM instance has a complete set of PreAct functionality including situation assessment, planning, intent interpretation and activity performance.

*A means to step through the decision-makers OPORD files in discrete time steps from days to weeks or months. At each time step, the OPORD elements for that time step are added to the PGM instances and processed. This results in script activations, the input of data that updates the PGM instances, and output of actions from PGM scripts.

*A means to execute each of the other PGM instances that are populations of interest but may have no decision input, so that they read report data, update their beliefs, update their plans, goals and intentions, activate scripts and perform actions as appropriate for the time period selected.

*An Operational Environment State Engine that receives actions from the PGMs, determines the effects of the actions, updates internal state tables and generates reports that are used by each of the PGM instances to update their beliefs and plans.

*A GUI that allows a user to step through the OPORD files one time step at a time or to allow it to run to completion. The User can specify the persistence interval in terms of time steps. At the end of each save interval, the following data is persisted, labeled with the time step: operational environment state of grounded truths, and the plans, goals, beliefs, and script instances for each PGM.

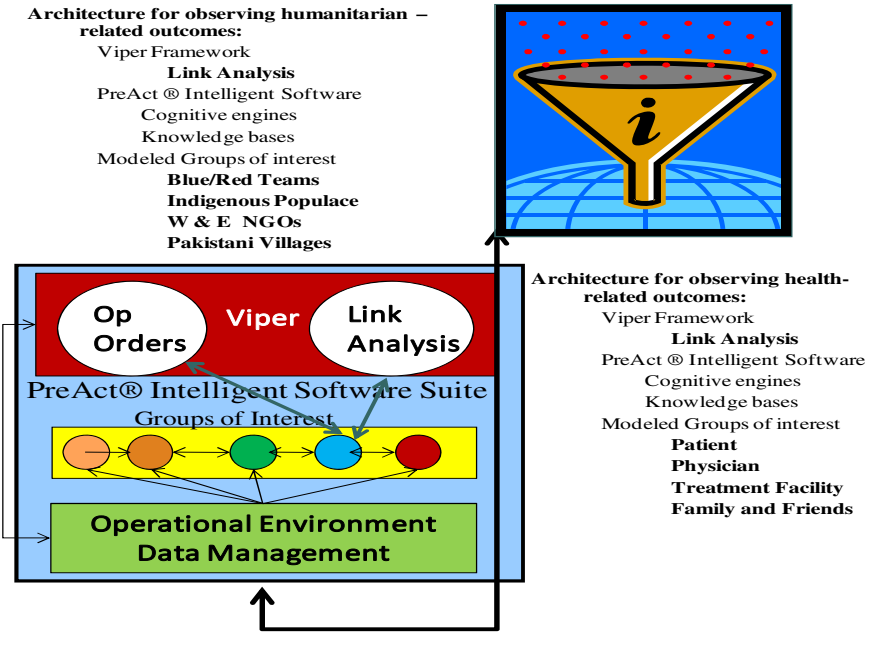


Fig. 6. Schematic of the SOPATH Architecture used for analyzing population models

*The Operational Environment State Engine (OESE) is a simulator for the ground truth state of the environment. It does not store beliefs or intentions, but consists of simple look-up relationships between input actions and output reports. Internal is a series of tables that store the state of the OE elements that the actions act upon. At Time Zero, the OESE internal tables shall be initialized by reading a file. Following initialization, each Population Group Model (PGM) will be executed based on the static data within the OESE. Actions by the PGMs are queued up. Time then steps to Time One. The action queue is then processed to update the state of the OE internally and to generate reports to the output queue. Each of the PGMs, starting with the decision-making PGM, reads the report queue and updates its beliefs, plans, goals, and scripts to generate output actions into the action queue. Once all of the PGMs have been executed, the time steps forward to Time Two, and so forth. At each Save Interval, the internal state of the OESE (and contents of the PGMs) shall be persisted to a database, labeled with the time of the save interval so that the state data may be inspected by the user after the system has completed its scenario execution. Two use cases are represented in Figure 6, which shows the components for analyzing emergent behaviors when modeling the interventions and their effects on children who are obese and for courses of action taken by a non-government official (NGO) performing humanitarian work within Pakistani Villages.

5 The Population Model Validation Process and Implications

Solid validation techniques for population modeling are still in their infancy. An attempt to validate the military model's predictions is currently underway and is being addressed by considering a time-split historical case study [10]. The only way of testing the probability distributions is to repeat the same test a number of times and count the number of times each outcome occurred. It will be important to modify the scenario to exhaust a significant number of variations, to confirm that the outcomes are sufficiently affected by the input. In order to validate the population models, data from a trial program will be used as input reports, and surveys about various population beliefs and values will need to be conducted in order to obtain the outputs of the program in terms of effects-based outcomes.

There are basic social and cultural structures and processes that influence effects-based outcomes. When populations experience a miss-match of sociocultural factors that influence expectations and intentions, this can trigger an affective reaction – strong enough to measure and possibly forecast. This paper provides a framework for modeling the emergent behaviors between populations with conflicting values and intentions. The models presented with corresponding conceptual, architectural models for forecasting outcomes can be used when interventions, courses of actions, government policies, or business plans are simulated within constructive simulations.

References

1. Hogg, M.A., Tindale, R.S. (eds.): Blackwell Handbook of Social Psychology: Group Processes. Blackwell Publishing, London (2003)
2. Lennox, M.G.R.: Modeling and Simulation in an Era of Persistent Conflict. Headquarters, Depart-ment of the Army G-3/5 Strategic Plans, Concepts, & Doctrine Division (2007)

3. Phillips, C.L., Geddes, N., Kanareykin, S.: A Balanced Approach for LLOs Using Group Dynamics for COIN Efficacy. In: Proceedings of the 2nd International Applied Human Factors and Ergonomics Conference, Las Vegas NV, July 14-17 (2008)
4. Army Field Manual, Counterinsurgency. FM 3-24/MCWP 3-33.5 (2006)
5. Geddes, N.D., Atkinson, M.L.: An approach to modeling group behaviors and beliefs in con-flict situations. In: Liu, H., Salerno, J.J., Young, M.J. (eds.) Social Computing, Behavioral Modeling, and Prediction, pp. 46–56. Springer, New York (2008)
6. MacNulty, C.A.R.: Truth, Perception & Consequences. Proteus Monograph Series, Army War College 1(1) (November 2007)
7. Phillips, C.L., Geddes, N., Crosscope, J.: Bayesian Modeling using Belief Nets of Perceived Threat Levels Affected by Stratagemical Behavior Pattern. In: Proceedings of the 2nd International Conference on Cultural Computational Dynamics, Washington, D.C., September 15-16 (2008)
8. Geddes, N.D.: A model for intent interpretation for multiple agents with conflicts. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, SMC 1994, San Antonio, TX (1994)
9. Cialdini, R.: Influence: The Psychology of Persuasion. Collins Pub., NY (1998)
10. Phillips, C., Sokoloff, S., Crosscope, J., Geddes, N.: A Validation Process for Predicting Stratagemical Behavior Patterns of Powerful Leaders in Conflict. In: Proceedings of the 2nd International Workshop on Social Computing, Behavior Modeling, and Prediction, Phoenix, AZ, March 31-April 1 (2009)

PGT: A Statistical Approach to Prediction and Mechanism Design

David H. Wolpert[1,⋆] and James W. Bono[2]

[1] NASA Ames Research Center
MailStop 269-1
Moffett Field, CA 94035-1000
david.h.wolpert@nasa.gov
[2] Department of Economics
American University
Washington, D.C. 20016
bono@american.edu

Abstract. One of the biggest challenges facing behavioral economics is the lack of a single theoretical framework that is capable of directly utilizing all types of behavioral data. One of the biggest challenges of game theory is the lack of a framework for making predictions and designing markets in a manner that is consistent with the axioms of decision theory. An approach in which solution concepts are distribution-valued rather than set-valued (i.e. equilibrium theory) has both capabilities. We call this approach Predictive Game Theory (or PGT). This paper outlines a general Bayesian approach to PGT. It also presents one simple example to illustrate the way in which this approach differs from equilibrium approaches in both prediction and mechanism design settings.

1 Introduction

Often we can quantify the preferences of the players in a non-cooperative game in terms of Von Neumann-Morgenstern utility functions. One can then predict that the mixed strategy profile of the players is a Nash Equilibrium (NE) (or some appropriate refinement of the NE) of that game [1].

A difficulty with this approach to predicting the mixed strategy profile is that it conflicts with the extensive experimental data in behavioral game theory that has established that human beings are not fully rational and do not play NE. In addition, this approach typically predicts that out of the uncountably infinite possible mixed strategy profiles that could be chosen by the players, there are very few with non-zero probability (sometimes just one, in fact). All other strategy profiles are deemed to be physically impossible. Moreover, the NE does not provide relative probabilities of the strategy profiles it deems to have non-zero probability.

Other non-Nash equilibrium concepts, like the Quantal Response Equilibrium (QRE) [2], level-k thinking [3], etc., can be viewed as modeling players that are

⋆ We would like to thank George Judge and Julian Jamison.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 314–322, 2010.

not fully rational, in contrast to the NE. But they too typically predict that out of the uncountably infinite possible mixed strategy profiles chosen by the players, the set of all profiles that have non-zero probability has measure zero, in manifest contradiction of physical truth. Also like the NE, they do not provide relative probabilities of the strategy profiles deemed to have non-zero probability. We will use the term Equilibrium Concept Approach (ECA) to refer to any approach, like the NE, QRE, etc., that maps an arbitrary game specification $\mathscr{I}$ to an associated set of mixed strategy profiles, $\mathscr{E}(\mathscr{I})$, without providing any information concerning relative probabilities.

Predictive Game Theory (PGT) is a new approach to the positive problem of predicting the mixed strategy profile. In this approach one views specification of the game as statistical data, stochastically coupled to the player behavior that we wish to predict. This approach replaces the ECA issue, of how to specify a function $\mathscr{E}$ mapping any game $\mathscr{I}$ to a set of equilibrium strategy profiles, with the issue of how to specify a density function over all possible mixed strategy profiles of that game, e.g., i.e., of how to specify a Bayesian posterior $P(q \mid \mathscr{I})$. Therefore, this new approach should be viewed as an alternative to the ECA in general, rather than as an alternative to any particular instance of the ECA, e.g. Nash equilibrium.

We use the term PGT to refer to any such application of statistical inference *to* games, in contrast to the use of statistical inference by some players *within* a game. The premise is that the game theoretician outside a game, making predictions about player behavior, should use the same Bayesian kind of reasoning used by all players inside the game who are rational. By adopting this premise PGT avoids the problems of the ECA. Indeed, because PGT assigns relative probabilities to all q's, it provides capabilities beyond those of the ECA. In particular, PGT allows external regulators to determine optimal regulations for situations that are beyond the ability of ECA to address.

To illustrate how PGT differs from the ECA, note that a PGT distribution $P(q \mid \mathscr{I})$ is far more informative than a single "best prediction". However if desired we can synopsize the distribution with a single prediction. One way to do that is to use the mode of the distribution as the prediction. When the distribution is a Bayesian posterior probability, $P(q \mid \mathscr{I})$, this mode is called the Maximum A Posterior (MAP) prediction. Alternatively, say there is a real-valued loss function, $L(q, q')$ that quantifies the penalty we will incur if we predict q' and the true value is q. Then Bayesian decision theory counsels us to predict the "Bayes optimal" value, which is the q' that minimizes the posterior expected loss, $\int dq L(q, q') P(q \mid \mathscr{I})$ [4,5,6,7,8].[1]

Using decision theory this way, to map any game to a single associated Bayes optimal strategy profile, provides a PGT "equilibrium concept". Note that this equilibrium concept depends on the loss function of the game theoretician who is making the prediction, and therefore is *not* specified in the game. Accordingly,

[1] Here we will write integrals with the measure implicit. So if the set being "integrated over" is countable, we implicitly mean a point measure, in which the integral is equivalent to a sum.

this equilibrium concept will vary with the external game theoretician who is making the prediction. This contrasts with the ECA, which ignores the concerns of the external game theoretician when telling that game theoretician what prediction to make, and therefore forces that game theoretician to violate Savage's axioms when making her prediction.

Another contrast with the ECA is that the PGT equilibrium concept typically produces a single strategy profile, without any need for a refinement. Yet another contrast is that typically in PGT the pure strategies of the players are not statistically independent. This is true even though the support of the posterior density function $P(q \mid \mathscr{I})$ is restricted to mixed strategy profiles q under which the players *are* statistically independent. Formally, the distribution over pure strategy profiles is given by

$$\begin{aligned} P(x_i, x_{-i}) &= \int dq\ P(x_i, x_{-i} \mid q, \mathscr{I}) P(q \mid \mathscr{I}) \\ &= \int dq\ q(x_i, x_{-i}) P(q \mid \mathscr{I}) \\ &= \int dq_i dq_{-i}\ q_i(x_i) q_{-i}(x_{-i}) P(q_i, q_{-i} \mid \mathscr{I}) \end{aligned} \tag{1}$$

and in general this differs from the product of the distributions over pure strategies,

$$P(x_i)P(x_{-i}) = \left[\int dq_i\ q_i(x_i) P(q_i \mid \mathscr{I}) \right] \left[\int dq_{-i}\ q_{-i}(x_{-i}) P(q_{-i} \mid \mathscr{I}) \right] \tag{2}$$

Furthermore, often under the Bayes optimal strategy profile no player's strategy is a best response to the strategies of the other players. Assuming there is more than one NE of the game, this is true even if the players are all fully rational, i.e., if the support of the density over strategy profiles is restricted to the NE. In this sense, "predictive" bounded rationality is automatic under PGT, in contrast to the case when using the ECA.

Another important difference between PGT and ECA's involves the types of numerical techniques that may be needed to evaluate their predictions. In the ECA, typically such techniques arise to solve sets of simultaneous nonlinear equations. In contrast, under PGT numerical techniques typically arise to solve constrained maximization problems (e.g., if one wishes to find $\text{argmax}_q P(q \mid \mathscr{I})$) or to solve integrals (e.g., if the loss function is quadratic, so that the Bayes optimal prediction is the average $\int dq\ qP(q \mid \mathscr{I})$). Especially in large problems, the computational burdens of the numerical techniques used in PGT can be far smaller than those under the ECA.

The following are two additional benefits of PGT over the ECA:

1. Another advantage of PGT is that, being a fully statistical model, it can combine multiple types of information / data into an associated posterior. This ability is necessary to properly express the uncertainty the game theoretician still has about the strategy profile after incorporating all that information.

As an example, say the game theoretician is uncertain about the utility functions, so that $\mathscr{I}$ is a distribution over possible utility functions. (Note that the game theoretician may have such uncertainty about the players' utility functions even for a complete information game, where the players have no such uncertainty about one anothers' utility functions.) Then the proper way for the game theoretician to express her associated uncertainty over mixed strategy profiles is by averaging over that distribution.

As a simple illustration, suppose m is the probability that the utility functions are $\mathscr{I}'$, and $1-m$ the probability that they are instead $\mathscr{I}''$. Then PGT says we must average over those two sets of utility information to properly express game theoretician uncertainty. Formally, we write $\mathscr{I} = \{\mathscr{I}', \mathscr{I}''\}$ and break the posterior into two terms:

$$P(q \mid \mathscr{I}) = mP(q \mid \mathscr{I}') + (1-m)P(q \mid \mathscr{I}'').$$

In contrast, in the ECA, trying to address uncertainty about the utility functions in a similar fashion would entail averaging over the associated equilibrium sets somehow. It is not at all clear that the axiomatic foundations of the ECA provide a principled way of doing such averaging.

2. Perhaps the most important benefit of PGT's statistical approach is that it not only allows us to address point prediction in a principled, decision theoretic manner (as described above), but also to address mechanism design problems this way [9,1,10]. In fact, PGT allows us to extend the scope of "mechanism design" far beyond its usual domain, into a full-fledged theory of "game control".

   More precisely, say we have a controller who can set a parameter λ specifying some aspect of a game played by a set of (perhaps bounded rational) players, whose mixed strategy profile is q, as usual. As an example, λ might specify the form of an auction, or any similar choice of a mechanism in a mechanism design problem. More generally, λ can be any choice that someone external to the N-player game can make that will modify that game before it is played.

   Let $G(q, \lambda)$ be the "social welfare" function of the controller, and indicate the game specified by λ as Γ_λ. Let $\mathscr{I}$ be some other information that the controller has concerning the game and/or player behavior, in addition to the value λ that she will choose. Then the standard approach of optimal control (i.e., Bayesian decision theory) says that the controller should set λ to

$$\text{argmax}_\lambda \left[\mathbb{E}(G \mid \mathscr{I}, \lambda) \right] = \text{argmax}_\lambda \left[\int dq \; G(q, \lambda) P(q \mid \mathscr{I}, \lambda) \right] \tag{3}$$

   So for example, if the controller's utility function only depends on the pure strategy profile of the players, we can write $G(q, \lambda) = \int dx \; q(x) W(x)$ for some function W. In this case the controller should set λ to

$$\text{argmax}_\lambda \left[\int dq \; G(q, \lambda) P(q \mid \mathscr{I}, \lambda) \right] = \text{argmax}_\lambda \left[\int dq dx \, W(x) q(x) P(q \mid \mathscr{I}, \lambda) \right] \tag{4}$$

To contrast this with the ECA, consider the case where the N-player game has multiple equilibria for every value of λ. Let G^j_λ be the expected social welfare of the j'th equilibrium for value λ. Consider the case there are pairs $\lambda, \lambda' \neq \lambda$ such that the intervals $[\min_j(G^j_\lambda), \max_j(G^j_\lambda)]$ and $[\min_j(G^j_{\lambda'}), \max_j(G^j_{\lambda'})]$ overlap. Then in contrast to PGT, the ECA can provide no advice whatsoever on whether the controller should choose λ or λ'.

2 PGT Posterior with Perfectly Rational Agents

The first thing we know about the players is that under their joint mixed strategy their moves are statistically independent (since we are restricting attention to normal form games). Beyond that, all of the insights of behavioral game theory, psychology, and human modeling [11,12,13,14,15] could be brought to bear on the task of determining the likelihood.

Here though our goal is far more modest, and we will not try to formalize those insights. Rather we will simply construct a likelihood function that is a plausible model of real world behavior. We offer an illustrative example of a simple likelihood for NE in which a regulator is charged with choosing among two games. This example is not offered as a plausible model of real world behavior. It is intended to distinguish between equilibrium concepts and statistical prediction. This example highlights the importance of taking a decision-theoretic approach to mechanism design.

Example 1: Say that a market regulator can choose to have a pair of players play the game g_1, or instead play the game g_2. Each is given by its payoff function below.

$$g_1 = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \qquad g_2 = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix} \tag{5}$$

Say we know that the players are fully rational. This knowledge, combined with the game specification, is $\mathscr{I}$. So $\mathscr{I}$ consists of two parts: the specification of the utility functions of the game, $\mathscr{I}_u$, and the information that q must be a NE, $\mathscr{I}_{NE}$. This means that the likelihood is given by

$$\begin{aligned} P(\mathscr{I} \mid q) &= P(\mathscr{I}_u, \mathscr{I}_{NE} \mid q) \\ &= P(\mathscr{I}_{NE} \mid \mathscr{I}_u, q) P(\mathscr{I}_u \mid q). \end{aligned} \tag{6}$$

The first term, $P(\mathscr{I}_{NE} \mid \mathscr{I}_u, q)$, equals 0 for any q that is not a NE, and it has the same value for all NE q's. So we know that $P(\mathscr{I} \mid q) = 0$ for non-NE q.

To complete the specification of $P(\mathscr{I} \mid q)$, we must specify the value of the remaining $P(\mathscr{I}_u \mid q)$ term in Eq. 6. For simplicity, here we assume this term is the same constant for all q that we are interested in, the ones that are NE for the joint utility u. This assumption allows us to equate $P(\mathscr{I} \mid q)$ with $P(\mathscr{I}_{NE} \mid \mathscr{I}_u, q)$, up to an irrelevant overall multiplicative constant. (The constant divides out when

we calculate the normalization term in the posterior.) This means that we can set $P(\mathscr{I} \mid q)$ for any NE q equal to 1 (again, up to an overall constant) while it equals 0 for all other q.

The prior we use is the entropic prior, given by $P(q) \propto e^{-\alpha S(q)}$, where α is an associated hyperparameter and $S(q) = -\sum_y q(y)\ln[\frac{q(y)}{\mu(y)}]$ is the Shannon entropy of q. This prior gives more weight to q's that contain less information about the moves of the players.

Let the social welfare function that the regulator wants to optimize be

$$\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \tag{7}$$

Should the regulator choose to have the players play game g_1 or g_2?

Note that even under the assumption that the players are perfectly rational, *this question cannot be addressed using ECA*. Conventional game theory is moot on whether the regulator should choose g_1 or g_2, because there are multiple equilibria for both of those games.

To answer the question using PGT, we simply calculate the expected social welfare under choices g_1 and g_2. This means we must evaluate the posterior. To do this for g_1, first note that it has three NE: (A, A), (B, B), and $(1/3, 1/3)$. The first two of those q have entropy 0 (they are delta functions). The associated value of the entropic prior, $exp(\alpha S(q))/Z(\alpha)$, is just $[Z(\alpha)]^{-1}$. The last NE has entropy $2ln[3] - (4/3)ln[2]$.

If we define $w(\alpha) \triangleq exp(\alpha\{2\ln[3] - (4/3)\ln[2]\})$, then the prior probability of the first two (pure strategy) NE are $1/[2 + w(\alpha)]$, and the prior probability of the last (mixed strategy) NE is $w(\alpha)/[2 + w(\alpha)]$. Since all three equilibria have the same likelihood (namely, 1), these prior probabilities of the equilibria are also their posterior probabilities, $P(q \mid \mathscr{I})$. All other q have zero posterior probability. We can similarly calculate the posterior for g_2.

Evaluating expected social welfare yields $36 + 19w(\alpha)$ for g_1 and $54 + 16w(\alpha)$ for g_2. The choice of α — which reflects the regulator's prior belief about the behavior of the players — directly determines the optimal action of the regulator.

Next consider a modification of this scenario where the regulator is risk averse. Then it is not just the expectation values $\mathbb{E}(\text{social welfare} \mid \mathscr{I}, g_i)$ that matter; the regulator's utility function is not linear in social welfare. So higher moments of the social welfare matter as well. Since it provides a full distribution over all q's for any game g_i, PGT can tell the regulator how to behave in these situations as well. Again, this contrasts with ECA, which only provides a set of q's, not a distribution over them.

Now return to the case of a risk-neutral regulator. However say that rather than a social welfare function that depends on the joint pure strategy x, the social welfare function depends on the mixed strategy profile, q. For example, the social welfare function might be $G(q) = -S(q)$, reflecting a preference by the regulator that the players randomize as little as possible, instead acting in a consistent, predictable manner. So the expected welfare would be $-\int dq\, P(q \mid \mathscr{I})S(q)$. Say

that the regulator can choose between the game g_1 and the game g_3 which has common payoff function

$$\begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \tag{8}$$

Which game should the regulator choose?

Under game g_1, the players choose the mixed strategy NE, which has entropy $s \equiv 2\ln[3] - (4/3)\ln[2]$, with probability $w(\alpha)/[2+w(\alpha)]$, where $w(\alpha) = \exp(\alpha s)$. (The other two NE have zero entropy, and therefore social welfare of 0, and therefore have social welfare of 0.) So expected social welfare is

$$\mathbb{E}(G \mid \mathscr{I}, g_1) = -s\frac{\exp(\alpha s)}{2 + \exp(\alpha s)} \tag{9}$$

In contrast, under game g_3, the entropy of the mixed NE is $s' \equiv 4\ln[2] - (3/2)\ln[3]$, and it occurs with probability $w'(\alpha)/[2 + w'(\alpha)]$, where $w'(\alpha) = \exp(\alpha s')$, resulting in

$$\mathbb{E}(G \mid \mathscr{I}, g_3) = -s'\frac{\exp(\alpha s')}{2 + \exp(\alpha s')} \tag{10}$$

So the entropy of the mixed strategy profile is larger in game g_1, and it occurs with higher probability in that game as well. Accordingly, PGT tells us that the regulator should choose game g_3, no matter what α is. Again, this analysis contrast with ECA, which cannot provide any advice whatsoever to help the regulator make their decision.

For this entropy-based social welfare function, the regulator is concerned with the distribution over q's, not just the resultant projection to a distribution over x's. There are many other social welfare functions that have this character. An example is where the regulator can set a sales tax level κ, and we model the effect of the choice of κ by having it be a parameter in a game being played by a set of N people. More concretely, say that after that tax level is announced, the players all set their mixed strategies to an associated NE, and then sample those mixed strategies repeatedly to get joint moves. So $P(q \mid \kappa)$ is sampled once, producing some strategy profile q^* that is a NE for the game defined by κ. After this, q^* is IID sampled many times, getting a set $D \equiv \{x^j : j = 1, M\}$ of many pure strategy profiles x^j. All that sampling of q^* will provide a set of empirical values $\{\hat{u}_i(D)\}$ that can be viewed as unbiased, low-variance estimates of the expected utilities of each player under q^*.

In some situations the players will only care about those the sum of those empirical values, and reflecting this our regulator may only care about the sum of those empirical values, $\hat{U} \equiv \sum_i \hat{u}_i(D)$. (For pedagogical clarity, we here take all discounting factors to equal 1.) When this is the case, at the stage where she is setting κ, the regulator will only care about the κ-dependence of the distribution over values of $\hat{U}$,

$$P(\hat{U} = v \mid \kappa) = \int dD dq^* \; P(D \mid q^*) P(q^* \mid \kappa) \delta\left(\sum_i \hat{u}_i(D) - v \right) \tag{11}$$

Now for any q^* and player i, for large M, it is unlikely to have a D such that $\hat{u}_i(D)$ differs significantly from the associated value $\mathbb{E}(u_i \mid q^*)$. In such cases we can approximate $\sum_i \hat{u}_i(D)$ in the integrand with $\sum_i \mathbb{E}(u_i \mid q^*)$. This means that we can approximate

$$P(\hat{U} = v \mid \kappa) \simeq \int dq^* \; P(q^* \mid \kappa) \delta\left(\sum_i \mathbb{E}(u_i \mid q^*) - v \right) \tag{12}$$

$$= P([q^* : \sum_i \mathbb{E}(u_i \mid q^*) = v] \mid \kappa). \tag{13}$$

In other words, because q^* will be sampled many times, the regulator cares only about expected utility under q^*, and is not sensitive to the variations in total utility that occur from one sample of q^* to the next. By inspection of Eq. (13), this means that the regulator is concerned with the full distribution $P(q \mid \kappa)$, not the marginalization $P(x \mid \kappa)$.

ECA does not say anything about the relative probabilities of the q^* associated with any given κ. Accordingly, ECA cannot provide any advice on what the regulator should do in these kinds of situations.

References

1. Myerson, R.B.: Game theory: Analysis of Conflict. Harvard University Press, Cambridge (1991)
2. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. Games and Economic Behavior 10, 6–38 (1995)
3. Crawford, V., Iriberri, N.: Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? Econometrica 75, 1721–1770 (2007)
4. Jaynes, E.T., Bretthorst, G.L.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
5. Berger, J.M.: Statistical Decision theory and Bayesian Analysis. Springer, Heidelberg (1985)
6. Zellner, A.: Some aspects of the history of bayesian information processing. Journal of Econometrics (2004)
7. Paris, J.B.: The Uncertain Reasoner's Companion: A Mathematical Perspective. Cambridge University Press, Cambridge (1994)
8. Horn, K.S.V.: Constructing a logic of plausible inference: a guide to cox's theorem. International Journal of Approximate Reasoning 34, 3–24 (2003)
9. Fudenberg, D., Tirole, J.: Game Theory. MIT Press, Cambridge (1991)
10. Nisan, N., Ronen, A.: Algorithmic mechanism design. Games and Economic Behavior 35, 166–196 (2001)
11. Camerer, C.: Behavioral Game theory: experiments in strategic interaction. Princeton University Press, Princeton (2003)

12. Starmer, C.: Developments in non-expected utility theory: the hunt for a descriptive theory of choice under risk. Journal of Economic Literature 38, 332–382 (2000)
13. Allais, M.: Econometrica 21, 503–546 (1953)
14. List, J.A., Haigh, M.S.: A simple test of expected utility theory using professional traders. Proceedings of the National Academy of Sciences 102, 945–948 (2005)
15. Kurzban, R., Houser, D.: Experiments investigating cooperative types in humans. Proceedings of the National Academy of Sciences 102, 1803–1807 (2005)

Developing Cognitive Models for Social Simulation from Survey Data

Jonathan K. Alt and Stephen Lieberman

Modeling, Virtual Environments and Simulation (MOVES) Institute
Naval Postgraduate School
Monterey, California 93943

Abstract. The representation of human behavior and cognition continues to challenge the modeling and simulation community. The use of survey and polling instruments to inform belief states, issue stances and action choice models provides a compelling means of developing models and simulations with empirical data. Using these types of data to population social simulations can greatly enhance the feasibility of validation efforts, the reusability of social and behavioral modeling frameworks, and the testable reliability of simulations. We provide a case study demonstrating these effects, document the use of survey data to develop cognitive models, and suggest future paths forward for social and behavioral modeling.

Keywords: Social simulation, cognitive models, agents.

1 Introduction

Populating simulations of society for use in analysis, training, and experimentation requires the identification of reliable data sources from which to derive authentic representations of human behavior and cognition. Survey and polling data present a compelling source of information from which to draw inferences about a population's social structure, and the underlying beliefs, values and interests (BVI) of individuals. These types of data present a traceable means of informing simulation effort from empirical data despite inherent limitations of self-reported data [1]. The ability to gain insight into how individuals in a population are likely to behave in a given situation makes social simulation an attractive tool for decision makers. Here we explore the use of existing survey data to populate a multi-agent social simulation with both representative entity cognition and social network structures, and discuss considerations in producing survey instruments tailored for this use case.

The use of appropriate survey and polling data to populate social simulations provides a direct and traceable means of model development. This paper will describe the process of populating entity cognitive models from survey data for use in the CG model. The paper will first discuss.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 323–329, 2010.

2 Models, Methods, and Tools

We describe the process of populating a social simulation, the CG model, with data from existing open source surveys to depict the general concepts of the theory under study. This section will provide an overview of the CG model, the World Value Survey (WVS), and discuss the development of cognitive models from the WVS.

2.1 CG Model

The Cultural Geography (CG) model serves here as an example of a multi-agent social system [2]. It is a government owned, open source, data driven simulation for the representation of societies. It is built upon open architecture and a re-usable "plug and play" modular framework that allows researchers and decision-makers to easily experiment with different social theories and data sets to allow simulations to be completely tailored to the area and time period under study, facilitating the implementation of appropriate social theories for the each society to be examined [3]. The data driven tool, continuously under development, serves as a platform for experimentation and analysis of potential futures based on actions taken by actors in the multi-agent system [4]. The data development process closely follows the counter-insurgency intelligence preparation of the battlefield model outlined by Mansoor, with procedural roots in anthropology and demography [5].

Each model instantiation in CG is the result of a data development and collection efforts to 1) populate this framework with heterogeneous agents that are representative of the society under consideration, and 2) authentically represent the social structure of the population in terms of the likelihood or frequency of agent interaction, communication, and influence [2]. The main modules related to the representation of the population within the simulation are the entity cognition module and the social structure module. The entity cognition module contains implementations of social theories that describe the way each agent perceives and responds to events within the simulation, and manages the internal BVI states of each agent. Conversely, the social structure module contains implementations of social theories that describe the distributions of attributes in a population, and manages the interactions and communications between agents in the system.

The identification of relevant issues is the first step toward the instantiation of a scenario within the simulation. Once the top issues have been identified they are treated as agent responses in the simulation. For each population issue, or response, to be examined within the simulation, factor selection is conducted over the remaining responses for each socio-demographic subtype of interest. Using the most relevant contributors to the response a Bayesian network is generated for each socio-demographic subtype from the source data, the survey results. Multiple techniques exist for conducting feature selection and for generating Bayesian networks from source data [6-8]. In this case the commercial data mining software, Clementine[1], was used for both steps. Future data generation efforts will seek to streamline the process through the use of Weka[2], an open source data mining package implemented in Java.

[1] http://www.spss.com/software/modeling/modeler/
[2] http://www.cs.waikato.ac.nz/ml/weka/

2.1.1 Entity Cognition

The cognitive module implementation used in this example relies on the concept of the narrative paradigm which states that each human possess a unique identity based on their culture and their life experiences [9]. This "narrative identity" forms the lens through which an individual views the world and interprets events, and is implemented as a Bayesian belief network (BBN) for each agent within the model [10]. The beliefs, values and interests relevant to a particular issue of interest to both the population and the policy maker are identified and used to populate a BBN. This serves to guide the entity's stance on issues such as, "Are you willing to fight in war for your country?" The strength of these foundational beliefs, treated as Bayesian priors, controls the ease with which new information can change an entities stance on an issue [11].

Action choice in the implementation used for this case study is guided by the theory of planned behavior [12]. The theory of planned behavior (TPB) describes the manner in which people develop intentions to carry out some behavior. The theory states that individuals will form an intention to execute a behavior based on: 1) their attitude toward the behavior, 2) their perception of the group norms associated with that behavior, and 3) their perceived level of behavioral control in regard to that behavior [12]. The TPB is again implemented as a BBN, which serves to guide entities formation of the intention to act based on its unique interpretation of events within the model. This paper will focus on BBN's related issue stance within the model.

The Bayesian approach is used extensively in this cognitive model implementation. It is not the only approach that could be implemented within the framework, but it is appealing for a number of reasons [13]. Literature from cognitive psychology points to Bayesian networks as a general approach to understanding how the human mind executes induction about the world based on noisy observations provided by world experience [14]. As Griffith points out, philosophers and mathematicians have used probability theory to describe human cognition for over 200 years and the transfer of cultural knowledge from generation to generation can be viewed in a Bayesian framework [11].

2.1.2 Social Structure

Social structure refers to the distributions of social factors in the population under study. Social factors influence interaction, communication and affiliation among agents in the simulation. The social structure is implemented as a social network with the link between nodes reflecting the similarity between nodes based on relevant social factors. This degree of likeness, or homophily, regulates the likelihood of communications between entities within the model [14-21]. The development of a homophily network from survey data can be accomplished using the full set of responses from an individual or a subset of the responses, unlike the case in the generation of cognitive models when it is preferable to use a subset of the available data. The use of survey data in the development of social network structure for use in social simulation will be discussed in a separate paper.

2.2 Survey Data Considerations

Despite the inherent limitations in the use of survey and polling data, it presents a ready and well accepted means of gaining insight into large samples of a population.

We use the CG model to illustrate the process of populating a social simulation from existing survey or polling data. In the initial use of social simulations this will most often be the case that the researcher or analyst will find themselves in when examining a new geographic area with social simulations. The initial models will need to be informed by existing data sources and information requirements specific to the population of the social simulation of choice will need to be planned into ongoing and future data collection efforts using tailored collection instruments.

The World Value Survey presents a useful source of data for developing initial models of populations for areas of interest where focused data collection tailored for use in social simulation is not ongoing. This paper presents a case study development of entity cognitive models using data from the World Value Survey[3] (WVS) for 2005 for Indonesia. The WVS is carried out by members of the World Values Survey Association to gain insight into changing worldviews and changing beliefs, values and interests around the world. A full description of the survey and the data itself is available at the website in the footnote below.

2.3 Populating the Entity Cognition Module

This case study will develop BBN for two representative entity stereotypes for the country of Indonesia. The issue for consideration is the level of agreement with the statement, "I am willing to fight in war for my country." The current implementation relies on BBN's to determine each entities stance on issues such as the statement above. Using the WVS as a data source for each country this leaves over 300 responses as potential contributors to an entity's level of agreement with this statement.

In this example a representative stereotype developed from the relevant socio-demographic dimensions for the area is developed. In developing models of these stereotypes from survey data the first step is to partition the data set and determine if data is available to support the representation of the people group selected. The desired stereotype is representative of a male manual laborer, with low education levels, in the youngest age bracket. The levels selected for the partitioning must be informed by the meaning of the categories as documented in the survey documentation, as well as the practical matter of data availability. This is particularly the case when working from existing data.

```
  ('Highest educational level attained' <= 5)    and
  ('Nature of tasks: manual vs. Cognitive' <= 5) and
(Sex = 1) and ('Age of respondent -   3 intervals' = 1)
```

The first step in the process used in this case is the application of feature selection techniques to identify the relative importance of the beliefs represented by survey responses to the stereotype under study's stance on the issue. Numerous techniques exist for feature selection [15]. Given the categorical nature of the WVS data, feature selection using Cramer's V was chosen as a reasonable means of identifying the most relevant beliefs to include in the development of BBN's for the issue. Cramer's V determines the nominal level of association between predictors and a target based on Pearson's chi-square statistic, x^2 in the equation below [16].

[3] http://margauz.grandvinum.se/Seb/Test/wvs

$$CV(f_k) = \sqrt{\frac{x^2(f_k)}{n * \min(r-1, c-1)}}$$

where $x^2(f_k)$ is the chi-square value of variable f_k
r is the number of values of the variable,
c is the number of classes of the target variable,
n is the total number of records.

While chi-square determines if there is a significant relationship between variables, it does not give insight into the relative level of significance or importance to the target [16]. Cramer's V provides this information in the form of a score between 0 and 1, with 1 indicating a strong association between variables and a null or 0 relationship equivalent to statistical independence between the target and response being scored [16]. Using this technique reduces the potential contributors from 315 to 5 for the first example. Note that the use of simpler BBN's is driven by the desire to facilitate subject matter expert input and face validation of the constructed models and further sensitivity analysis needs to be conducted to better characterizes the advantages and disadvantages of larger networks.

Table 1. Results of application of Cramer's V during feature selection for Indonesian entity stereotype 1 left and stereotype 2 right

Item	Value
Child qualities: tolerance and respect for other people	1
Neighbours: Militant minority	0.982
Child qualities: determination perseverance	0.98
Neighbours: People who speak a different language	0.98
Neighbours: Immigrants/foreign workers	0.96

Next the same test was applied to the factors selected for inclusion in the model, treating each as a target for the other 4 to check for covariance. This revealed a strong association between "Neighbours: Immigrants/foreign workers" and "Neighbours: Militant minority" which was subsequently removed from the model.

The remaining factors were then used as the basis for a BBN. In this case, tree-augmented naive (TAN) Bayes techniques are used to instantiate a BBN using maximum likelihood criteria for learning the conditional relationships among the predictors and the target [17]. TAN Bayesian networks are distinguished from naïve Bayesian networks in that networks generated from TAN techniques allow for relationships among the predictors using a restricted form of correlation edges [6]. These attributes allow TAN Bayesian networks to maintain the general applicability of naïve Bayesian networks while improving accuracy or the naïve approach [6].

Once the belief space is reduced to a feasible number of dimensions for implementation in the chosen social simulation, a case file data set is created for import into the simulation. The feasible number of dimensions will vary based on the constraints of the chosen simulation platform. Note that data and the needs of the analysis drives the level of granularity feasible within the model. In this case only 30 respondents from Indonesia fit the criteria to be included in this stereotype. This case file represents a sample from the population described by the BBN.

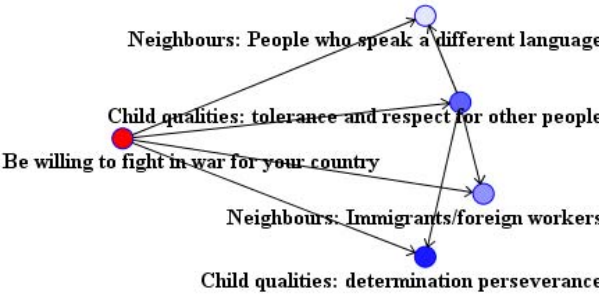


Fig. 1. BBN for Indonesian entity stereotype 1 produced using TAN Bayes

Table 2. Partial case file for Indonesian entity stereotype 1for use in populating entity cognition models in social simulation

Child qualities: tolerance and respect for other people	Child qualities: determination perseverance	Neighbours: Immigrants/foreign workers	Neighbours: People who speak a different language	Be willing to fight in war for your country
2	1	2	2	1
1	1	2	2	1
2	2	2	2	1

Cognitive models for use in social simulations developed from survey data allow traceability of data from the population to the model. The use of longitudinal survey data provides a potential means to assist with efforts to validate models derived in this manner, reducing a hurdle in most cases, the source of the data.

3 Conclusions and Future Work

Social simulations populated by survey or polling data provide a viable means of supplying decision makers with insights into the potential futures generated by their policy actions. Decision makers are faced with complex adaptive systems, societies, that they must gain insight into in order to make informed decisions with limited resources to most effectively address relevant issues. Social simulations provide a potential means of conducting analysis of potential futures, but must present challenges in the development of data from traceable sources. This paper presents one method to leverage survey data, with its inherent limitations, to populate cognitive models of issue stance for social simulations. Further work is required to characterize the multiple approaches to this problem domain and strengths and weaknesses of each. Additional work is also required to document a methodology for the use of survey data in the population of action choice models, such as the theory of planned behavior. A further discussion of populating social network structures using the Indonesia case study will be provided in a separate, but complementary paper.

References

[1] National Research Council. In: Behavioral Modeling and Simulation: From Individuals to Societies. National Academies Press, Washington (2008)

[2] Jackson, L.: Narrative Paradigm in Cultural Geography Modeling. In: Human Socio-Cultural Behavioral Focus 2010 (2009)

[3] Michel, F., Ferber, J., Gutknecht, O.: Generic simulation tools based on mas organization. In: 10th European Workshop on Modelling Autonomous Agents in a Multi Agent World MAMAAW (2001)

[4] Ferber, J., Gutknecht, O., Michel, F.: From agents to organizations: an organizational view of multi-agent systems. In: Giorgini, P., Müller, J.P., Odell, J.J. (eds.) AOSE 2003. LNCS, vol. 2935, pp. 214–230. Springer, Heidelberg (2004)

[5] Mansoor, P.: Mershon Center for International Security Studies. The Ohio State University

[6] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning 29, 131–163 (1997)

[7] Zhao, Z., Liu, H.: Searching for interacting features. In: Proceedings of the 20th International Joint Conference on AI, IJCAI 2007 (2007)

[8] Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182 (2003)

[9] Fisher, W.: Clarifying the Narrative Paradigm. Communication Monographs 56, 55–58 (1989)

[10] Smith, K., Kalish, M.L., Griffiths, T.L., Lewandowsky, S.: Introduction. Cultural transmission and the evolution of human behaviour. Philosophical Transactions B 363, 3469 (2008)

[11] Beppu, A., Griffiths, T.L.: Iterated Learning and the Cultural Ratchet

[12] Ajzen, I.: The theory of planned behavior. Organizational behavior and human decision processes 50, 179–211 (1991)

[13] Tenenbaum, J., Griffiths, T., Kemp, C.: Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences 10, 309–318 (2006)

[14] Griffiths, T.L., Tenenbaum, J.B.: Randomness and coincidences: Reconciling intuition and probability theory. In: Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society. Human Communication Research Centre, p. 370. University of Edinburgh, Edinburgh (August 1-4, 2001)

[15] C. DATA, Exploratory Data Analysis with Categorical Variables: An Improved Rank-by-Feature Framework and a Case Study

[16] Wang, H., Parrish, A., Smith, R.K., Vrbsky, S.: Variable selection and ranking for analyzing automobile traffic accident data. In: Proceedings of the 2005 ACM symposium on Applied computing, pp. 32–37 (2005)

[17] Friedman, N., Goldszmidt, M.: Building classifiers using Bayesian networks. In: Proceedings of the National Conference on Artificial Intelligence, pp. 1277–1284 (1996)

Dynamic Creation of Social Networks for Syndromic Surveillance Using Information Fusion

Jared Holsopple[1], Shanchieh Yang[2], Moises Sudit[1], and Adam Stotz[1]

[1] CUBRC, Inc., Information Exploitation, Buffalo, NY
[2] Rochester Institute of Technology, Department of Computer Engineering, Rochester, NY
{holsopple,sudit,stotz}@cubrc.org, jay.yang@rit.edu

Abstract. To enhance the effectiveness of health care, many medical institutions have started transitioning to electronic health and medical records and sharing these records between institutions. The large amount of complex and diverse data makes it difficult to identify and track relationships and trends, such as disease outbreaks, from the data points. INFERD: Information Fusion Engine for Real-Time Decision-Making is an information fusion tool that dynamically correlates and tracks event progressions. This paper presents a methodology that utilizes the efficient and flexible structure of INFERD to create social networks representing progressions of disease outbreaks. Individual symptoms are treated as features allowing multiple hypothesis being tracked and analyzed for effective and comprehensive syndromic surveillance.

Keywords: Information Fusion, Social Networks, Disease Outbreak Prevention, Syndromic Surveillance.

1 Introduction

This paper considers the problem of syndromic surveillance for the early detection of disease outbreaks through the use of social network modeling. Clusters within a theoretical social network will be dynamically created by INFERD: Information Fusion Engine for Real-Time Decision-Making. These clusters, called tracks, are intended to follow the onset of symptoms over time that may be indicative of a potential spread of a disease. They contain not only relevant relationships, but also a *confidence* that a given patient has symptoms corresponding to a given disease. The tracks represent the presence of known disease symptoms and indicators in an array of patients related based on criteria such as geographical residence, family relationships, previous medical conditions, etc.

Syndromic Surveillance is defined as "surveillance using health-related data that precede diagnosis and signal a sufficient probability of a case or an outbreak to warrant further public health response [4]." As is widely known, many medical institutions are beginning to transition to Electronic Medical Records (EMRs) and Electronic Health Records (EHRs). There are emerging standards (such as Health Level 7 [2] or NHIN Connect [3]) that attempt to standardize the storage and transmission of these data structures. These electronic records offer numerous advantages over traditional paper records: greatly reduced storage space, immediate

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 330–337, 2010.

access to records, electronic back-up, and electronic record-sharing between institutions. EHRs and EMRs provide a wealth of data about patients, such as: demographics, medical history, family relations, current symptoms and diagnoses, prescriptions, and lab test results. These health records, among others, can be mined to determine any trends that may be associated with a potential disease outbreak.

Many diseases (such as the flu or common cold) can be transmitted through physical contact. Other diseases (such as HIV) can only be transmitted through certain types of physical contact (such as sexual intercourse or intravenous drug use). Many of these types of relationships can be modeled as social network, which is a structure of persons and entities that are related in some way. While social networks have been applied to different applications, Eubank, et al. explored the modeling of disease outbreaks through the use of social networks for an urban environment [5]. This paper will explore a way to dynamically generate similar social networks using electronic health and medical records. Rather than creating a complete social network, INFERD dynamically instantiates multiple clusters, or tracks, based on available patient information. The hypothesized tracks can be analyzed and cross-referenced based on a variety of criteria and measurements for syndromic surveillance analysis, thus providing a comprehensive situation assessment of potential disease outbreaks.

2 INFERD Overview

INFERD (Information Fusion for Real-Time Decision-Making) is a general information fusion tool that allows for the tracking of related uncertain events of interest [1]. Using data from heterogeneous sensors or databases, INFERD dynamically instantiates *tracks* of related events and observables from a *model* catered to the specific application of interest. INFERD is implemented in Java using a service-oriented messaging architecture, and receives sensor data from different types of sensors and databases.

INFERD first receives data from various databases and then must align that data so it can be properly interpreted with INFERD. Once the data has been aligned, INFERD uses a *model* to give connotation, or meaning, to the raw data. The model pre-defines certain relationships that should be tracked over time. INFERD was originally implemented for computer security, so that model tracked the actions of hackers through a given target network. For disease outbreak, we will track the onset of disease symptoms through people based on time, location, family relations, etc. Once meaning has been applied to the data, INFERD then tries to associate this data with an existing *track* or create a new one. A *track* illustrates how different concepts are related based on the input data. Once the new data has been associated with a track, the tracks are updated and additional processing can occur that allows various measures (such as breadth and density) of the track to be analyzed for any strong indicators of trends important to a given application.

3 INFERD for Syndromic Surveillance

In this section we will illustrate how INFERD can be applied to the early detection of disease outbreak by applying its usage to the tracking of H1N1 virus symptoms and

showing how multiple tracks can be analyzed to capture important characteristics that may be indicative of a potential disease outbreak. While INFERD could be applied towards many different diseases, we chose H1N1 due not only to its recent interest in the media and medical fields, but also because symptoms are similar, but different from, other strains of influenza.

3.1 Data Alignment

For this application, we will assume that the data is being sent from various medical databases containing EHRs, EMRs, and other pertinent records using the Health Level 7 [2] standards. It should be noted that for the sake of simplicity in this paper, we will limit our inputs to these records. However, INFERD is designed to handle input from multiple heterogeneous sources provided that the data from each source is formally structured, so other sources could also be added as input to the system to enhance the accuracy. When data is received by INFERD, it first passes through the Data Alignment process, which converts the incoming data to a format that INFERD is able to process, such as an XML document. The fields within the XML document can be referenced by the *model*, which is used to dynamically create tracks.

3.2 Connotation Elicitation: Modeling Disease Symptoms and Indicators

INFERD uses a data structure called a *feature tree* to represent a *concept*. The leaves of the feature tree are called *feature nodes* that map observables to the tree. Each feature node contains a *confidence* that corresponds to the confidence that the observable is helpful in the inference. Other nodes in the feature tree are *aggregation nodes* that perform a specific logical or mathematical aggregation function such as *and*, *or*, *max*, *min*, *at least N*, etc. The feature tree can then be evaluated to determine the confidence that the given concept is present.

At the core of a disease outbreak, many people start exhibiting similar symptoms such as diarrhea, vomiting, nasal problems, etc. These symptoms can not only be indicated by records from a doctor's visit, but they can also be indicated by the prescription of various drugs. For example, Amoxicillin is commonly prescribed for nasal problems such as a sinus infection. Since there are different policies at different institutions and pharmacies for reporting data, the database(s) of information INFERD receives its data from may not contain every record created. Since not all data may be available, INFERD must be able to model and infer the presence of symptoms based on incomplete data. The presence of a symptom in INFERD is represented by a feature tree that provides the mapping from the XML document in the data alignment process to the model. Observables are mapped to feature nodes through the definition of *constraints* which must be satisfied for the mapping to exist.

Consider Fig. 1 which shows a sample feature tree to capture indicators for the onset of nasal symptoms in a patient. The three oval nodes are feature nodes and contain constraints indicated by the boxes underneath them. When an observable is sent to INFERD, the dynamic constraints on *instantiated* feature nodes in tracks are checked first. If none of those are satisfied, then the *static* constraints in the model are checked. Static constraints are represented by the shaded boxes and static constraints

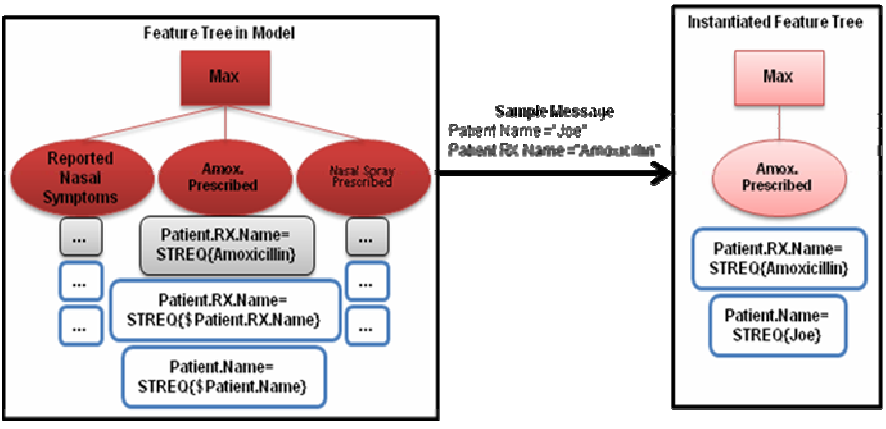


Fig. 1. An example feature tree to represent indicators of nasal symptoms. It should be noted that this feature tree has intentionally been reduced to a small number of indicators for the sake of brevity.

are represented by the non-shaded boxes. The feature node "Amox.Prescribed" contains the static constraint "Patient.RX.Name=STREQ{Amoxicillin}", which tells INFERD to map a given observable to that feature node if the Patient.RX.Name field is equal to (by the STREQ function) "Amoxicillin". INFERD supports other constraint functions corresponding to basic logic and comparison operators (such as and, or, greater than or equal to, etc.) as well as custom user-defined functions. INFERD also allows for the optimization of constraint checking, so that the number of nodes to check the constraints of is minimized through the use of lookup tables and other functions. When the static constraint are satisfied by a record, INFERD instantiates an instance of that feature tree and creates the dynamic constraints. Not shown in Fig. 1 is that each feature node contains a weight that corresponds to the confidence that the given observable does indicate the presence of a given symptom.

3.3 Connotation Elicitation: Modeling a Disease

Recall that a feature tree represents a *concept*. We have already shown that a feature tree can be used to represent the concept of a symptom, but it can also be used to represent the concept of a disease. It should be noted that the modeling of a disease is *not* intended for automatic diagnosis, but it is intended to determine a level of confidence in the disease being present given a *potentially incomplete* set of data.

INFERD models a disease by the logical combination of various symptoms and indicators. Consider the H1N1 virus that exhibits very similar symptoms to the common flu (fever, sore throat, nasal problems, cough, respiratory problems, body aches), but the presence of diarrhea and/or vomiting in an adult is a strong indicator that it is the H1N1 strain. Also, there is a lab test that can confirm the presence of the disease. We can model the H1N1 virus for a patient using the feature tree in Fig. 2. The ">=5" and ">=1" aggregator functions are the "at least N" aggregator function that calculates the minimum of the highest 5 confidences of each child. The "WA"

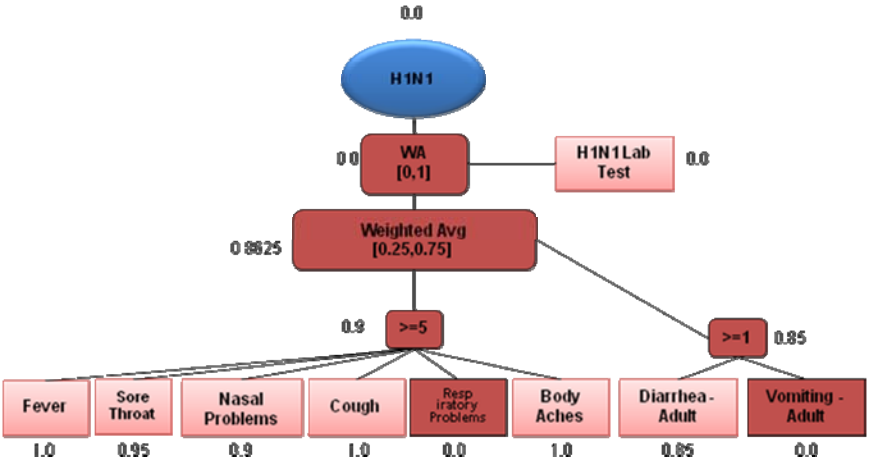


Fig. 2. Instantiated feature tree after receiving a negative lab test

aggregator function is the weighted average function where the weights are defined in the brackets for each child from left to right. Note that when an instance of this tree is created (in the same manner as symptoms are instantiated) the weights are normalized to only the instantiated children. We will illustrate weight normalization and feature tree instantiation through an example.

Suppose that INFERD receives multiple records corresponding to the presence of H1N1 symptoms in a patient. This would trigger the instantiation of the each observed symptom as well as all parent nodes of that symptom, as shown in Fig. 2. Since all required disease symptoms are inferred to be present, the confidence for the presence of H1N1 was calculated to be high at the "Weighted Average Node" (0.8625). However, this example also shows a negative lab test, so the confidence that the given patient has H1N1 is zero since the lab test overrides the symptoms (indicated by the round H1N1 node).

3.4 Data Association: Creating a Model for Disease Outbreak Prevention

Now that we have identified how INFERD can calculate the confidence that a patient has a given disease (or set of symptoms) with incomplete data, we can define a complete INFERD model that is able to dynamically create attributed graphs equivalent to social networks, called tracks, that can then be analyzed through social network analysis.

INFERD uses a *model* to relate concepts to each other. For the sake of disease outbreak prevention, we consider a node to represent the presence of a disease in a person. Each edge will represent some type of relationship between two people, such as location, race, demographic, family relations, etc. For this paper, we will consider location and family relations (which are both available in a complete EHR), but the model can easily be extended to capture other relationships if they can be provided by the incoming records.

Each *concept* node in the model represents a different disease. These concept nodes each contain a feature tree to define the symptoms and indicators for the disease, which was described in the previous section. This model contains multiple arcs that define whether two possible instances of the same disease are present. Like feature nodes, these arcs contain dynamic constraints that must be satisfied for two patients to be linked together in the same track. The INFERD model contains multiple types of arcs that capture relationships involving location, family, time of observed symptoms, etc.

When a concept node is instantiated (due to its feature tree being instantiated), its emanating arcs are also instantiated. When this concept is instantiated, the dynamic constraints (defined in the same way as feature nodes) are created. A generated track is equivalent to a social network in that the nodes have edges between them that represent a given relationship. INFERD also allows for a given observable to be correlated to multiple disease nodes and/or tracks. Therefore, there may be one or more tracks for a given disease. Multiple tracks for a given disease each represent a segment, or cluster, of a complete social network.

Recall that a disease node is instantiated when *any* symptoms are present in a given person, so using the example, the H1N1 disease is instantiated anytime Amoxicillin is prescribed, or any other flu symptoms are present. However, we can take advantage of INFERD's confidence assessments to filter out any nodes below a certain threshold. When SNA is performed on this network, we can choose to a node because it is below a certain threshold. Suppose, though, that future medical records are received that increase the confidence of an instantiated concept. In this case, the confidence may increase above the threshold and the node would now be considered for SNA.

3.5 Reporting: Multi-track Example

In this section, we will consider an example that shows multiple attack tracks and how these INFERD tracks could be analyzed for the detection of disease outbreaks. Consider the example tracks in Fig. 3 which were all generated to monitor the same disease.

Let us first consider how we can analyze tracks 1 and 2 which were generated from the location-based model that links patients together if they are from the same or neighboring cities and their symptoms were exhibited within two days of each other. Let us assume that each track, t, contains N_t nodes and has c_t unique cities represented by all the nodes. We can define $breadth_t = c_t$ to be an indicator of how wide-spread the symptoms are. We can then define $density_t = N_t/breadth_t$ to show how dense the symptoms are within a given region. Both of these measures can be helpful in identifying a potential disease outbreak. For example, track 1 has $breadth_1 = 3$ *cities* and $density_1 = 5/3 = 1.67$ *patients/city*. Track 2 has a low breadth measure ($breadth_2 = 1$), but a high density ($density_2 = 37/1 = 37$ *patients/city*). Based on these measures, we can conclude that the disease is concentrated to a much higher area in Boston than in Western New York. However, the breadth and density measurements ignore these confidences. Realistically, one would want to only consider nodes above a certain confidence level in their calculations. Such filtering is illustrated in track 3, which is

the same as track 2, but with a filter of 0.7 applied to it. Note that the track has now been partitioned into three separate fragments. This indicates three potentially independent outbreaks of the same disease within the same city. This also reduces the density measure since nodes have been removed ($density_3$ = *31 patients/city*). These measures can be compared to historical data to indicate acceptable levels of each measure.

Since each track is evolving and node confidences can change as more information is input to INFERD, we can used time-based measurements to determine the $breadth\_rate_t$ and $density\_rate_t$ for a given track that indicate how quickly $breadth_t$ and $density_t$, respectively, are changing over time. High breadth or density rates can also be indicative of a possible disease outbreak.

While INFERD does not directly generate a complete social network, it does identify multiple clusters of nodes related in different ways. Tracks 4-8 illustrate five tracks that were generated using a model that links patients together using any immediate family relationship. In this example, only track 4 identified any patients to be related. Interestingly, both Joe and Josh are in different location-based tracks. So this information could provide a potential link between tracks 1 and 2. It should be noted that a similar analysis can be done through various clustering algorithms, too. However, INFERD distinguishes itself from such algorithms because clustering algorithms are typically also complex and used as an offline algorithm that do not support continuous tracking.

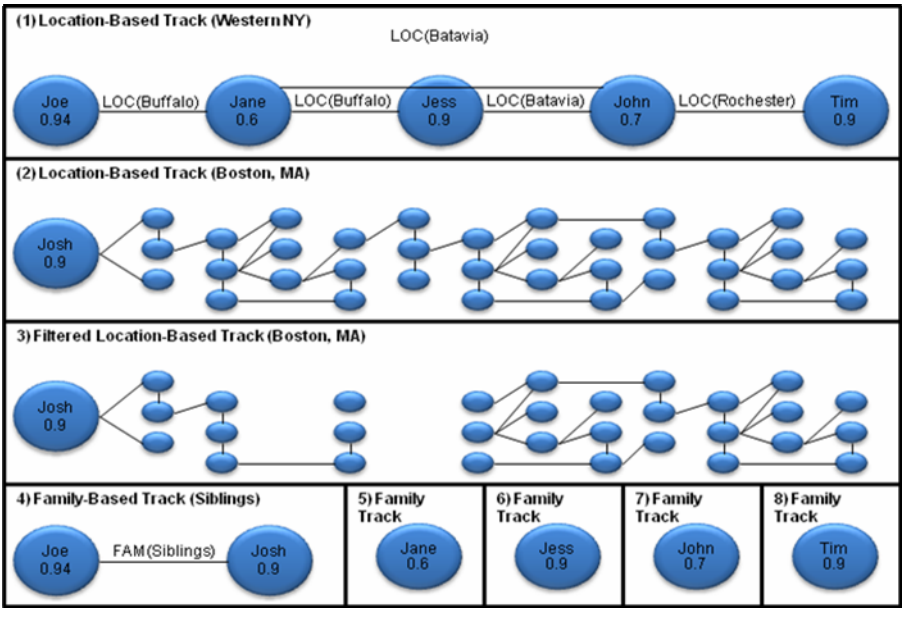


Fig. 3. Eight example tracks generated by INFERD. Buffalo, Batavia, and Rochester are all cities that border each other from west to east. Each node is labeled by the name of the patient and INFERD's associated confidence.

4 Conclusion

In this paper we have shown how an information fusion tool, INFERD, can be applied to dynamically create social networks for early disease outbreak detection using potentially incomplete electronic medical and health records. Using raw data provided from EMRs and EHRs we can dynamically create clusters of nodes that track the possible spread of a disease. This tracking is accomplished through the a priori definition of important relationships to the spread of a disease. Since raw data is available, the symptoms and other indicators exhibited by each patient allow INFERD to calculate a confidence of each association that can be used to filter tracks to contain only relevant nodes. These tracks can then be processed and cross-referenced with each other to extract meaningful information and identify potential disease outbreaks before they become a problem.

References

1. Sudit, M., Stotz, A., Holender, M.: Situational awareness of a coordinated cyber attack. In: Proceedings of International Data Fusion Conference, Quebec City, Quebec, CA (July 2007)
2. Health Level 7, http://www.hl7.org/ (accessed 11/4/2009)
3. NHIN Connect, http://www.connectopensource.org (accessed 11/4/2009)
4. Syndromic Surveillance: an Applied Approach to Outbreak Detection, http://www.cdc.gov/ncphi/disss/nndss/syndromic.htm (accessed 11/4/2009)
5. Eubank, S., Guclu, H., Kumar, V.S., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N.: Modelling disease outbreaks in realistic urban social networks. Nature 429(6988), 180–184 (2004)
6. Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A 311(3), 590–614 (2002)
7. Newman, M.E.J.: Analysis of weighted networks. Phys. Rev. E. 70(5) (November 2004)
8. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 7821 (2002)
9. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99(12), 7821–7826 (2002)
10. Wilson, A.G., Wilson, G.D., Olwell, D.H.: Evaluating Statistical Methods for Syndromic Surveillance, pp. 141–172. Springer, New York (2006)

Calibrating Bayesian Network Representations of Social-Behavioral Models

Paul Whitney and Stephen Walsh

Pacific Northwest National Laboratory
902 Battelle Boulevard, Richland, WA, 99352, USA
{paul.whitney, stephen.walsh}@pnl.gov
http://www.pnl.gov/computing

Abstract. While human behavior has long been studied, recent and ongoing advances in computational modeling present opportunities for recasting research outcomes in human behavior. In this paper we describe how Bayesian networks can represent outcomes of human behavior research. We demonstrate a Bayesian network that represents political radicalization research – and show a corresponding visual representation of aspects of this research outcome. Since Bayesian networks can be quantitatively compared with external observations, the representation can also be used for empirical assessments of the research which the network summarizes. For a political radicalization model based on published research, we show this empirical comparison with data taken from the Minorities at Risk Organizational Behaviors database.

Keywords: Computational social science, calibration.

1 Introduction

The patterns and underlying mechanisms which govern social-behavior are ongoing research in the political and social sciences. The theories governing these mechanisms are constructed based on empirical observations and experiments. The resulting theories and patterns of behavior are often documented in a narrative form. These narrative descriptions are valuable and communicate well within and across diverse technical communities. This paper describes how these narratives can be turned into models, demonstrates a Bayesian network model of one of the research outcomes, and then shows how the network can be calibrated with observations.

We model the social-behavioral theory with a formal mathematical framework called Bayesian Networks (BN) [1]. BN's have gained widespread use in various disciplines as artificial intelligence and decision making tools. BN's are attractive regimes for reasoning because they reduce the joint probability distribution of a set of variables to a more digestible and understandable structure using a set of conditional independence assumptions.

We have constructed BN models based on the theory, descriptions and patterns contained in the open social-science literature. Section 2 presents an example - a small model of formalized mechanisms for increasing political radicalization. This

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 338–345, 2010.

model represents mechanisms described in [2]. Section 3 contains a set of relevant extant data we use from the Minorities at Risk: Organizational Behavior (MAROB) data base [3] to calibrate the model. While not the outcome of a designed experiment, these data were constructed independently of our modeling efforts and so are a good candidate for model calibration and validation. The quantitative comparison of the model to the data would be straightforward via maximum likelihood or Bayesian approaches if the mapping between the MAROB data and the nodes in the BN representation were complete [1]. This is not the case for this model – some of the nodes do not have clear counterparts within MAROB. Thus Section 4 presents (briefly) a Monte Carlo method, Data Augmentation, which is readily adapted to BN's. This computational approach allows for fitting the BN models against incomplete data, while incorporating prior inputs for model parameters not informed by the available data. The outcome of these model calibration calculations are shown in Section 5.

2 Bayesian Network Models for Social Science Patterns

This section demonstrates the use of Bayesian Networks to represent mechanisms and features of organization behavior. BN's are understandable, tractable and sufficiently flexible to represent a wide variety of behaviors and relationships. Software for constructing and using BN's are widely available (e.g. GeNIe [4]). While BN's are more often used in decision support settings, they are also well suited for some modeling tasks. This section describes BN methods to represent behavior science mechanisms. Section 4 will show how BN's provide a quantification of the social science mechanisms.

The approach we have taken for constructing Bayesian network models for behavioral mechanisms follows. For a given set of patterns:

1. Identify key variables and concepts in the reference(s) – for purposes of modeling, this includes both the concept and *states* that concept may take. For instance [2] identifies levels of political radicalization that range from sympathy, identification and active participation political activities.
2. Identify relationships among groups of these variables. These are often in the spirit of "correlation" statements. For instance, the relationship between 'group radicalization' and 'Extremity in like-minded groups' is taken from the fifth mechanism in [2]. The network expresses a positive relationship between radicalization characteristics and group characteristics – notably the existence of group values (achievable by, or consistent with, political radicalization) and competition or pressure to adhere to those values.
3. Construct a BN that is consistent with the relationships.

The approach results in a BN model that is consistent with, and represents at least part of, the technical literature. From a mathematical perspective, the list of relationships from step 2 characterizes the marginal distributions of key variables. The overall model is a joint distribution consistent with the collection of marginal distributions. Some critical characteristics of this process and the resulting models include:

The models can be verified to behave in a manner consistent with the literature. This model verification is accomplished by revisiting the list of relationships, and checking that the model is consistent with each.

The methodology can result in different models from the same sources. While unfortunate, this characteristic reflects, in part, that the narratives often lack complete quantitative information regarding the relationships. There are many different parameter values and network structures that are consistent with a given set of qualitative relationships.

The model is a visual representation of the mechanisms described in the literature. The models provide an overview of some of the key concepts, and allow some exploration of the concepts and their relationships.

Finally, while the models are consistent with the literature or expertise on which they are based - they are not guaranteed to be 'true'. The models depend on the completeness of the literature on which they are based, and then the subsequent interpretation from the technical literature. This is further complicated by the observation that authors of technical papers were not writing with modelers as their intended audience. We discuss this issue further in the example below.

We construct a group radicalization model based on the discussion in [2]. This reference identifies twelve distinct mechanisms that can contribute to increasing political radicalization at three distinct scales: individual, group and society. We chose to model at the group level. The radicalization mechanisms are discussed separately in the reference. Accordingly, the model we developed treats them as essentially separate factors. Figure 1 shows the visual representation of a BN that is consistent with the narrative in [2].

3 Relevant Data and Issues

The MAROB [3] codebook states “The purpose of this project is to answer fundamental questions focusing on the identification of those factors that motivate some members of ethnic minorities to become radicalized, to form activist organizations, and to move from conventional means of politics and protest into violence and terrorism”. The MAROB data contains data on 118 organizations which represent 22 ethno-political groups in 16 countries of the Middle East and North Africa. We note that the data set contains records on approximately 176 primary and ancillary variables relating to various factors of radicalization.

To create data that can be used for model calibration, we had to decide which data variables mapped to the model variables. Further, many of the data variables had multiple states, so we also reasoned a way to map the data variable states to the ternary and binary states of the model variables. Some of this process was straightforward: for example, for a model variable “Group Fissioning” the corresponding data variable was “ORGSPLIT” which is an indicator of whether the organization split in the last year. Another straightforward example is the model variable “Isolated and Threatened” which had corresponding model variables “ORGOPEN” and “ORGLEGAL” which are indicators if the organization is operating clandestinely and if the state has declared the

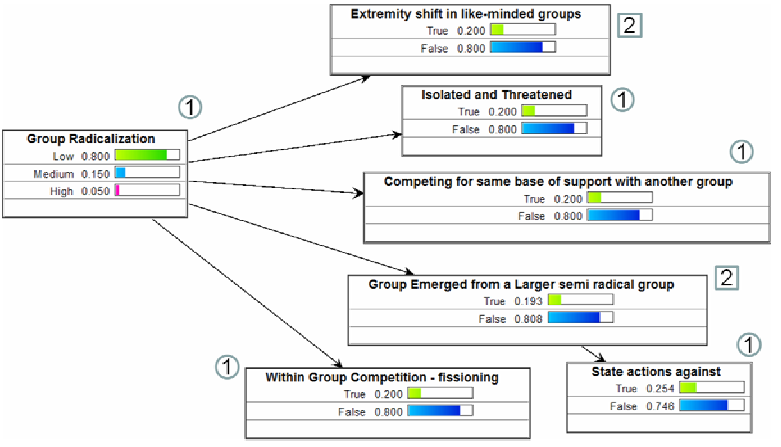


Fig. 1. Group Radicalization Model. Each main branch in the model corresponds with a group radicalization mechanism presented in [2]. The superscripts on the nodes indicate whether there was, in MAROB, some data (1) or no data(2).

organization illegal respectively. Mapping the variable states was done in a similar fashion. Last we note that for the model variable "Group Radicalization" we chose the data variable "ORGMILITANT" which was an indicator if the organization has committed a violent act or possesses violent means such as explosives – this variable, while not a complete indicator of political radicalization as defined in [2], overlaps with the highest level of political radicalization described in that reference. Also, note that this data variable is binary whereas the model variable is ternary, thus we map data values to only the "*Low*" and "*High*" states of the "*Group Radicalization*" variable.

It was difficult to define mappings to all model variables. For example, "Extremity Shift in Like-Minded Groups" and "Group Emerged from a Larger Semi-Radical Group" had no clear MAROB data to model mappings. Thus these variables are completely lacking data. Further, for the variables that did have a reasonable mapping, some of them were only partially complete, that is, they contained a small number of missing values. *Thus a parameter estimation algorithm that can mitigate these issues is desirable*. We will briefly describe it in the next section. Figure 1 shows which variables data. We used a subset of the MAROB by year, 2004, which contained 94 records.

4 Calibration Methods

In this section we briefly describe the formulation of Data Augmentation as a method of estimating the parameters of a BN with incomplete data. Data Augmentation [5] is based on a standard Bayesian estimation model [6], with the characteristic that some of the information is not observed. The model parameters are denoted by θ, the data by Y, and the unobserved data are denoted Z. The relationship among these is summarized by the probability distribution $\Pr(Y, Z \mid \theta)$. Note that this shorthand can hide a wealth of complexity. The underlying model can be a partial differential

equation, a systems dynamics model, a Bayesian network etc. An estimation goal is to obtain the posterior distribution of the parameters given the data: $\Pr(\theta \mid Y)$. A key mathematical result is that this posterior distribution can be obtained as the fixed point of an integral equation. Denoting the posterior distribution as $g(\theta) = \Pr(\theta \mid Y)$, we have the following equation:

$$g(\theta) = \int K(\theta, \varphi) g(\varphi) d\varphi \text{, where}$$

$$K(\theta, \varphi) = \int \Pr(\theta \mid Z, Y) \Pr(Z \mid \varphi, Y) dZ$$

$$\Pr(\theta \mid Y) = \int_Z \Pr(\theta \mid Y, Z) \Pr(Z \mid Y) dZ$$

$$\Pr(Z \mid Y) = \int_\Theta \Pr(Z \mid \varphi) \Pr(\varphi \mid Y) d\varphi$$

This fixed point integral equation is solved via Monte Carlo methods – resulting in a calibration of the model with the available data. Results from this calculation are given in the next section.

We implemented the Data Augmentation algorithm in *R* [7], a statistical programming language suitable for algorithm development. We took heavy advantage of the *gRain* library, a package developed for construction of BN's as well as simple inference and prediction [8]. An alternate approach to Bayesian estimation of the model parameters is presented in [9]. Early work in calibrating BN models is given in [10].

5 Calibration Results

In this section we report select results from the analysis. We present our initial estimates of the model parameters, based on the social science literature, before observing data. Then we present the results of parameter estimation from the data augmentation algorithm.

The variable "Isolated and Threatened" has parameters representing probabilities for the "*True"* state given the level of radicalization; the probability for the "*False"* state can be gleaned through taking the complement. Our initial value for the probability that "Isolated and Threatened" is "*True*" given that "Group Radicalization" is "*Low*" is 0.1 while our initial value for the probability that "Isolated and Threatened" is "*True*" given that "Group Radicalization" is "*High*" is 0.9. These probability statements are consistent with the position in the paper that isolated groups have an increased propensity towards radicalization.

To illustrate that the results from the Data Augmentation algorithm we present plots of the posterior distributions of

$$\Pr(\text{"Isolated and Threatened"} = \textit{True} \mid \text{"Group Radicalization"})$$

for "Group Radicalization" in the "*Low"* and "*High*" states in Figures 2 (a) and (b) respectively. The figure shows the posterior median, which may be considered a point estimate of the model probability, as well as the 95% Highest Posterior Density (HPD) regions. Table 1 summarizes the calculations.

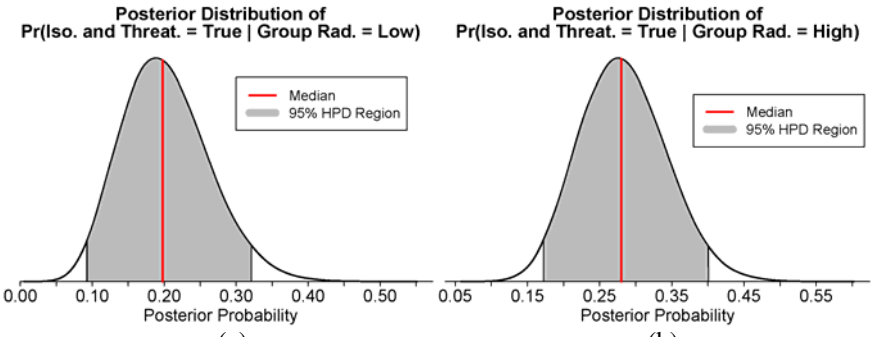


Fig. 2. (a) Posterior Distribution Pr(Isolated and Threatened = *True* | Group Radicalization = *Low*). (b) Posterior Distribution Pr(Isolated and Threatened = *True* | Group Radicalization = *High*).

Table 1. Summary statistics from the Isolated and Threatened Posterior

	Isolated and Threatened = *True*	
Group Radicalization	Initial Value	Estimates
Low	0.1	0.19
95% HPD Bounds		(0.9, 0.32)
High	0.9	0.27
95% HPD Bounds		(0.17, 0.40)

Table 1 indicates that the estimates for "Isolated and Threatened" = "*True*" are 0.19 and 0.27 given that "Group Radicalization" is in the "*Low*" and "*High*" states respectively. The probability estimates increase conditional on "Group Radicalization" in the "*Low*" and "*High*" states as do the initial guesses, however, we see that the estimates to not increase as strongly as the guesses. Further Table 1 shows that the 95% HPD bounds for each probability overlap. We interpret the overlap as very weak evidence that the probability estimates are significantly different from each other.

We next present results for the "Competing" variable. Since this variable is binary, the calibration results only speak to estimating this variable's probability for the "*True"* state given the level of radicalization; the probability for the "*False"* state can be gleaned through taking the complement. Our initial values for the probability that "Competing" is "*True"* given that "Group Radicalization" is "*Low"* is 0.1. Our initial values for the probability that "Competing" is "*True"* given that "Group Radicalization" is "*High"* is 0.9. These statements reflect the position in the reference paper that competition between groups corresponds with an increase in political radicalization level for (at least one) of the groups.

Table 2 contains the summaries of the calibration calculation for the "Competing" variable. The estimates for "Competing" = "*True*" are 0.87 and 0.77 given that "Group Radicalization" is in the "*Low*" and "*High*" states respectively. These results are in the opposite direction as compared with the initial probability assessments. The data estimates indicate that there is a large probability that the groups are competing

Table 2. Summary statistics from the Competing for Same Base of Support Posterior

	Competing = *True*	
Group Radicalization	Initial Value	Estimates
Low	0.1	0.87
95% HPD Bounds		(0.76, 0.95)
High	0.9	0.77
95% HPD Bounds		(0.65, 0.87)

regardless of the level of radicalization. Since the HPD regions overlap there is little evidence in these data that the true probabilities are significantly different from one another.

6 Conclusions

We presented and demonstrated an approach for representing narrative summaries of mechanisms and patterns of human behavior. We represented these using Bayesian networks. An example was shown of a BN representation for political radicalization mechanisms. The BN model was calibrated using MAROB data, resulting in a significant change in the model's parameters.

The approach taken is general, and has wide applicability. The modeling approach can be accomplished independently of the veracity of the hypothesized behavioral pattern, theory and/or mechanism. The model can be verified to behave in a manner consistent with the narrative literature. Given available data, the model can then be critiqued and evaluated in the context of that data. The model can then be validated (or refuted) based on the available data. While not discussed in this paper, the mathematical framework supports comparing the competing narratives by comparing the respective models fit with available data.

Acknowledgements

Amanda White and Landon Sego helped address roadblocks in developing the implementation of data augmentation used to calibrate the model parameters. The work described in this paper was developed within the context of the Technosocial Predictive Analytics Initiative (http://predictiveanalytics.pnl.gov) at the Pacific Northwest National Laboratory.

References

1. Jensen, F.V., Nielsen, T.D.: Bayesian Networks and Decision Graphs, 2nd edn. Springer, New York (2007)
2. McCauley, C., Moskalenko, S.: Mechanisms of Political Radicalization: Pathways Toward Terrorism. Terrorism and Political Violence 20, 415–433 (2008)
3. Minorities at risk organizational behavior dataset (2008), `http://www.cidcm.umd.edu/mar` (retrieved from June 2009)

4. GeNIe Tutorials, Decision Systems Laboratory of the University of Pittsburgh, http://genie.sis.pitt.edu/wiki/GeNIe_Documentation (accessed December 2008)
5. Tanner, M.: Tools for Statistical Inference. Springer, New York (1996)
6. Gelman, A., Carlin, J., Stern, H., Rubin, D.: Bayesian Data Analysis. Chapman and Hall/CRC, Boca Raton (2004)
7. R: A Language and Environment for Statistical Computing, http://www.R-project.org
8. Højsgaard, S.: gRain; A Graphical Independence Networks package in R (2009), http://genetics.agrsci.dk/~sorenh/public/R/gRainweb (retrieved from May 2009)
9. Riggelson, C.: Learning parameters of Bayesian networks from incomplete data via importance sampling. International Journal of Approximate Reasoning 42, 69–83 (2005)
10. Heckerman, D., Geiger, D.: Chickering: Learning Bayesian networks: The combination of knowledge and statistical data. Machine Learning 20, 197–243 (1995)

Social Network Data and Practices: The Case of Friendfeed

Fabio Celli[1], F. Marta L. Di Lascio[2], Matteo Magnani[3], Barbara Pacelli[4], and Luca Rossi[5]

[1] Language Inter. and Comp. Lab, Univ. of Trento
fabio.celli@email.unitn.it
[2] Dept. of Statistical Science, University of Bologna
francesca.dilascio@unibo.it
[3] Dept. of Computer Science, University of Bologna
magnanim@cs.unibo.it
[4] Independent researcher
bpacelli79@gmail.com
[5] Dept. of Communication Studies, University of Urbino
luca.rossi@uniurb.it

Abstract. Due to their large worldwide adoption, Social Network Sites (SNSs) have been widely used in many global events as an important source to spread news and information. While the *searchability* and *persistence* of this information make it ideal for sociological research, a quantitative approach is still challenging because of the size and complexity of the data. In this paper we provide a first analysis of Friendfeed, a well-known and feature-rich SNS.

1 Research Framework

Social Network Sites (SNSs) are undoubtedly one of the most interesting phenomena that bring together new technologies and social practices. They are going through an incredibly fast growth all over the world despite the fact many obstacles like the digital divide still exist. Despite this global success it would be hard to define a single global leader of the SNSs. Facebook, which counts more than 300 million single users mostly clustered in Europe and in the US, is surely a big player but QQ, with a high concentration of users in China, has an even larger user base. It seems that cultural diversity and local identity lead toward the choice of a specific SNS, while the shift toward the adoption of a SNS-model for online interpersonal communications seems to be global [1].

Due to this large worldwide adoption, SNSs have been widely used in many global events as an important source to spread news and information. From the terroristic attack in Mumbai in 2008 to the so-called Twitter revolution in Iran in 2009 SNSs proved several times to be a reliable way to communicate and to spread information in a quick and relatively efficient way. Within this scenario the sociological analysis of SNS based communication is still largely based on a qualitative ethnographic approach aimed at investigating living practices and

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 346–353, 2010.

uses of the SNS [2,3,4]. This approach gave us the opportunity to gain an effective insight in SNS users' lives, motivations and communicative strategies but failed in giving us a general description of how SNSs work and deal, as complex entities, with the diffusion of information.

Aim of this paper is to move a first step into a new direction of sociological SNS research coping with the topic from a multidisciplinary perspective. Within this proposed approach SNSs could be defined, at the same time, as the best and the worst place for sociological research. They can be defined as an optimal place because of the new and emerging properties that communication shows in these contexts. Information in SNSs, as boyd highlighted [5] can be defined also by *searchability* and *persistence* which are two positive characteristics for any researcher. Data can be searched and retrieved easily. At the same time the large amount of data that is published online every second can easily discourage any attempt to investigate online phenomena from a quantitative point of view. Data are out there, they can be searched and retrieved but that is still a great challenge. This paper will present some preliminary results of a larger research project that accepted this challenge and is dealing with a large quantity of SNS data in order to obtain a wider understanding of many unsolved issues in SNS research.

The #SIGSNA project, which stands for Special Interest Group on Social Network Analysis, started his research by analyzing a well known microblogging and social network service called Frienfeed (http://friendfeed.com). Friendfeed has been chosen because of several technical and sociological aspects. From a technical point of view, that will be described in the next section, Friendfeed offers a great level of access to the contents that are produced by the users. Everything that has not been marked as private is available online in RSS format. From a sociological point of view Friendfeed offers a very complex social dynamic constructed on a microblogging service (like the well-known Twitter) with the opportunity to comment the entries of other users. This very simple characteristic makes Friendfeed a microblogging platform able to host huge and complex conversations (made through comments). This is something very similar to what happens in Facebook, where conversation can arise as a sequence of comments to a specific status update.

2 The Social Data Set: Extraction and Structure

Data has been extracted from the Friendfeed application by monitoring the public URL http://friendfeed.com/public, where the system publishes a sample of recent posts. In the following, we will indicate with *post* any text entry or comment posted by a user, with *entry* a new conversation started by a user, and with *comment* a comment to an entry.

The URL was monitored for two weeks, from September 6, 2009, 00:00 AM to September 19, 2009, 24:00 PM, at a rate of about 1 to 2 updates every second (depending on network traffic). During the monitoring phase, all the identifiers of entries appeared on the public page have been saved on our local servers, for

later retrieval. After the end of this phase, we collected all the distinct entry identifiers (9.317.499) and retrieved the corresponding XML representations using the Friendfeed API. At this point, the data has been exported to CSV files, to remove unnecessary XML formatting and for database import. Then, at the end of the monitoring period, we have computed the network of users and followers. Starting from one random user, and retrieving all the connected graph of followers, we have extracted a related data set of more than 400.000 users, with about 15 million subscription relationships. The structure of the corresponding dataset is the following:

```
Entry(PostID, PostedBy, Timestamp, Text, Language)
Comment(PostID, EntryRef, PostedBy, Timestamp, Text, Language)
Like(User, EntryRef, Timestamp)
User(ID, Type, Name, Description)
Network(Follower, Followed)
```

All field names should be self-understandable. The only generated field is `Language`, which currently contains the most probable language identifier, e.g., `it` for Italian, etc. The final total number of posts is 10.454.195, considering both entries and comments and without private entries which could not be retrieved, for an amount of more than 2GB textual data, and 512.339 likes.

3 Statistical Analysis

Statistical analysis of the collected data aimed at offering a comprehensive description of the whole network and some deeper investigation of how a culturally defined part of the network (the Italian speaking sub-network) works. A general overview of the Friendfeed social network is indicated in Table 1.

Table 1. Min, max, mean and standard deviation values of variables observed in the whole network

	UsrFing	UsrFed	Post	Entry	Com	Like	ComR	LikeR
min	0	0	1	0	0	0	0	0
max	3,072	412	438	316	225	2,583	422	1,235
mean	2.73	3.09	28.01	27.17	0.85	1.93	0.73	1.39
sd	21.08	17.79	46.48	44.96	4.98	22.30	5.64	15.97

A preliminary introduction to the labels is required: UsrFing is the number of users that follow the user, UsrFed is the number of users followed by the user, Com is the number of comments made by the user, ComR is the number of comments received by the user, LikeR is the number of likes[1] received by the

[1] A *like* is a simple way to communicate some kind of appreciation toward an entry of another user. Instead of commenting by writing something a user can simply express his level of agreement with the published sentence by hitting the *like* button below the entry. This system is not unique to Friendfeed and it can be found also in Facebook.

Table 2. Correlation matrix of variables observed in the whole social network

	UsrFing	UsrFed	Post	Com	ComR	LikeR	Like	Entry
UsrFing	1.00	0.67	0.34	0.49	0.84	0.66	0.27	0.30
UsrFed	0.67	1.00	0.46	0.86	0.59	0.47	0.52	0.38
Post	0.34	0.46	1.00	0.35	0.28	0.24	0.22	0.99
Com	0.49	0.86	0.35	1.00	0.56	0.42	0.55	0.25
ComR	0.84	0.59	0.28	0.56	1.00	0.78	0.31	0.23
LikeR	0.66	0.47	0.24	0.42	0.78	1.00	0.41	0.20
Like	0.27	0.52	0.22	0.55	0.31	0.41	1.00	0.16
Entry	0.30	0.38	0.99	0.25	0.23	0.20	0.16	1.00

user, Like is the number of likes made by the user, Entry is the number of entries wrote by the user. Post is a derived element and it is the number of Comments and Entries made by the user.

Table 1 shows a lively network with an average of posts equal to 28.01, which means, in the two weeks of our sampling, more than two posts (entry or comment) every day. Despite this high rate of posting, Entries are much more than comments (mean 27.17 vs 0.85) showing a large amount of users that speak alone.

The correlation matrix represented in Table 2 can be used to have a more detailed picture of the whole network population and its habits. From the analysis of these values, it is possible to point out some not so obvious results: the correlation between entries done (Entry) and comments done (Com) is quite low as well as the correlation between comments done and comments received. These two data could suggest on one hand a distinction between the posting activity and the commenting activity and, on the other hand, the lack of reciprocity between the comments done and received.

On a more general level, we can observe in Figure 1(a) that the production of contents in the Friendfeed network has an heavy right tailed distribution that is well known in many web based services. This confirms the well-known fact that a small part of the users is very active and responsible for the largest part of the produced content. At the same time the largest part of user base contributes with very few entries. The last descriptive picture (Figure 1(b)) shows a bubble-plot of the entries related to the comments done. The size of the bubble is related to how spread is that specific cross through the network.

Figure 1(b) allows us to point out a few interesting aspects:

- The largest part of the entries does not have any comment and the largest part of the entries with no comments is made by low-activity users (with less than 29 entries in the sample)[2].
- There is a small but still significant part of users that produces only comments and no entries within the time frame of the sample.

[2] Notice that the class intervals for the two variables compared have been chosen on the basis of the sample quantiles of their distribution function.

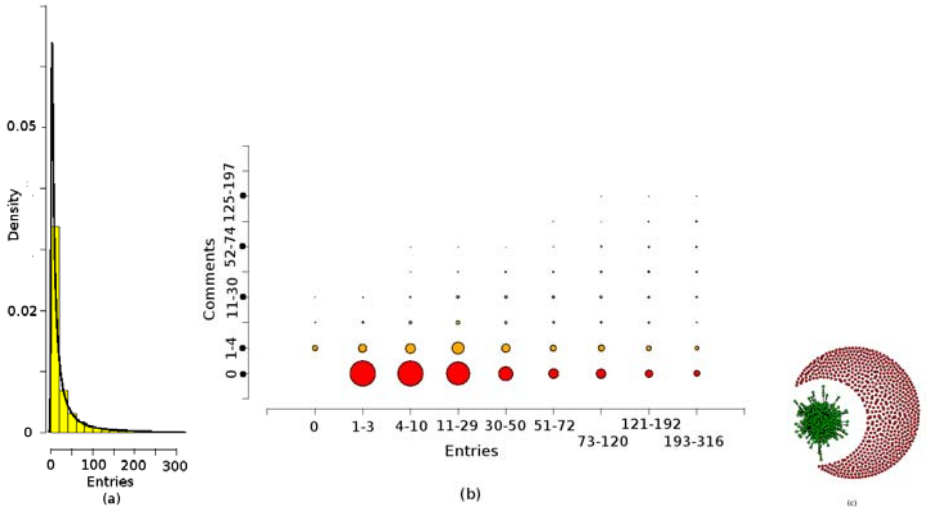


Fig. 1. (a) Density histogram with kernel density estimates of entries in the whole network, (b) relationship between number of posted entries and number of comments done, and (c) graphical representation of comments between Italian users

Finally, Figure 1(c) shows the cluster of the network of *comment* relationships for a 10.000 user subset of the Italian dataset.

As stated before, a specific goal of the #SIGSNA project is to start investigating SNSs by comparing how they are used in different cultural contexts. A preliminary analysis has been done using the language as a way to identify specific cultural contexts. The language identification has been obtained using specifically developed software called SLIde (Simple Language Identifier in Perl) [6], which enabled the creation of language annotations without requiring inefficient connections to Web-based services. Language identification appears to be a suboptimal strategy to identify the cultural context of the users. Many users can use several languages to address to different audiences. This will be surely true especially for languages, like English, widely spoken worldwide.

In order to be able to describe the limitations of the chosen method we have calculated the **language fidelity level** of every language in the sample. The language fidelity level shows the average level of posting in different languages for every user that posted, at least once, in a specific language. This level, obtained by calculating the mean values of the number of posts in each language divided by the total number of posts of each user with an entry in that language, allows us to establish the level of average fidelity toward a language. Several languages show a high level of fidelity (close to mean value 1.0), such as Italian and Portuguese. This means that users that write in that language usually keep writing in the same language all the time.

The language fidelity index has a double value for the #SIGSNA project: on one hand it allows us to identify the languages that can be used as good

Table 3. Mean and standard deviation values of the variables observed into clusters 1 and 2

	UsrFing	UsrFed	Com	ComR	LikeR	Like	Entry
mean 1	2.30	2.90	1.28	0.96	1.06	1.73	22.41
sd 1	9.50	9.22	4.28	4.05	5.73	8.51	18.96
mean 2	29.41	29.69	6.88	7.00	12.41	12.36	149.59
sd 2	93.35	63.84	17.97	22.90	52.09	45.37	65.57

indicators of specific cultural contexts and, on the other hand, it suggests the existence of specific nation-wide sub-networks that are loosely connected with the larger Friendfeed network.

The Italian sub-network has been further investigated with a cluster analysis of the users. Cluster analysis has been performed by using the *clara* algorithm: a partitioning method for finding clusters [7] into very large datasets. The number of clusters k has been chosen on the basis of the overall silhouette width [7] by varying k from 2 to 100. The average silhouette width is useful for evaluating the goodness of both the obtained clustering and the selected number of clusters, and it turned out that for our dataset the best number of clusters was two.

The cluster average widths are 0.85 and 0.13 for the two clusters, respectively whereas the average width of the clustering is 0.59 indicating that the clustering is quite appropriate. By Table 3 we note that the first cluster looks like the sample of users weakly active in the network whereas the second one looks like the sample of most active users. Notice that the biggest cluster is the first one.

Cluster analysis of Italian users gave us the opportunity to point out the existence of two different groups of users within the Friendfeed social network. A larger loosely connected and weakly active group coexists with a smaller heavy engaged group. This suggests, as we are going to discuss further in the sociological analysis, a wide range of uses of the Friendfeed social network.

4 Sociological Analysis

The statistical analysis of the Friendfeed social network depicts an interesting scenario. The descriptive analysis suggests a lively social network with a high level of production of content. Even if the overall level is quite high there are large differences between users' level of participation in the process. This suggests a highly personal use of Friendfeed according to many individual needs. Heavy users can post new updates continuously while light users could post a message once in a while.

An interesting aspect is the use, shown by the descriptive analysis, of Friendfeed only as a conversational space. Several users, in fact, didn't use Friendfeed to actually post something about themselves but just to comment someone else's posts. This suggests that Friendfeed can be used in a different way from how Twitter is used — Twitter is probably the most famous microblogging site. While in Twitter every conversation is a connected sequence of micro-posts, in Friendfeed conversations may take place in a dedicated comment space. This makes

Friendfeed conversations much more similar to what happens on Facebook than to what takes place on Twitter.

Another aspect that makes the conversational practices of Friendfeed similar to Facebook conversations is the identity of the audience. In Friendfeed, as well as in Facebook, due to specific architectural choices the potential audience of a user's comments is larger than the set of user's followers. By commenting someone else's update in Friendfeed (as well as in Facebook) a user moves herself into a semi-unknown place populated by semi-unknown users composed by all her friends and her friends' friends. Within this perspective Friendfeed can be considered both a microblogging service that allows you to share short thoughts and information with a network of friends and, at the same time, a purely conversational space where you can chat or discuss in a semi-protected environment mainly composed of your friends and their contacts.

In addition to a descriptive analysis of communicative practices taking place on Friendfeed we moved a first step toward a series of comparative analysis of SNS use in different cultural contexts. The analysis of the used language and the cluster analysis of the Italian posts gave us the opportunity to suggest some preliminary considerations. The Italian network of Friendfeed users seems to address his communication mainly to a local/national audience. The high level of language fidelity that has been observed suggests this. The average Italian Friendfeed user writes mainly in Italian and this could suggest the nationality of his *perceived* audience. The network here seems to be very closed on a Geocultural basis and connections with different international networks seem to be very rare.

Cluster analysis (Table 3) shows the existence of two groups of users within the Italian Friendfeed network. A small highly dedicated users group coexists with a larger but less active group. This suggests the high level of flexibility that Friendfeed allows. A double use of the medium is confirmed and the difference seems to be mainly based on the amount of communication that users produce. **Weak users** (group 1 in Table 3) are characterized by a couple of entries every day with few comments. This pattern of use seems to be the most classical use of microblogging saw as a way to update your online status and to share short thoughts with your friends. The **heavy users** group shows a completely different scenario. The level of average daily entries rises up to more than ten entries per day with a large level of comments done and received and even a greater level of likes done or received. These data suggest what we could define as a continuous stream of sharing that resembles the idea of life-stream. This comes together with a high level of conversational use of the service itself suggested by the high level of comments and likes and by the reciprocity of these. For this group of users Friendfeed seems to be a perfect platform which is able to host fruitful conversations starting from the sharing of a large quantity of information. Obviously, as we are going to discuss further in the conclusions, a qualitative analysis of the entries produces by the two groups is required in order to better understand the differences pointed out by the cluster analysis.

5 Conclusions and Research Perspectives

This paper presented the first results of the #SIGSNA research project. Aim of the #SIGSNA project is to develop a comprehensive analysis of social interactions that take place in the Friendfeed SNS. In this paper we showed a first descriptive analysis of the whole network that pointed out the existence of a large variety of uses inside the SNS. In addition to that, we used a language identification system to make comparative analysis of SNS uses in different cultural contexts. As a preliminary investigation we have also performed a cluster analysis of the Italian Friendfeed network, with which we have identified the existence of two different clusters of users: weak users and highly dedicated users.

The #SIGSNA project is characterized by the large database of entries that has been collected during the sampling period. Due to the large dimension of the database and to the high quality of the collected data the presented results have to be considered just as a first bite of the whole research that is still in progress. At the same time, our analysis is highlighting some computational limitations of traditional social data analysis tools, which cannot deal with the large amount of information produced by SNSs — in particular, traditional and text clustering algorithms implemented into widely used statistical tools could not be applied to the whole network, which presents hundreds of thousand users, millions of arcs, and millions of text posts. These limitations will drive the development of scalable techniques for the analysis of large and complex networks, which are necessary to deal with the size of current real social datasets.

References

1. Cosenza, V.: Osservatorio facebook (2009), http://www.vincos.it/osservatorio-facebook (retrieved on August 31, 2009)
2. boyd, d., Ellison, N.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication 13(1) (2007)
3. Siiback, A.: Online peer culture and interpretative reproduction on children's social networking profiles. In: COST conference proceedings (2009)
4. Hardey, M.: ICT and generations constantly connected social lives, the good the bad the challenging. In: COST conference proceedings (2009)
5. boyd, d.: Taken Out of Context: American Teen Sociality in Networked Publics. PhD thesis, University of California-Berkeley, School of Information (2008)
6. Celli, F.: Slide: Simple language identifier in perl. Technical report, University of Trento (2009)
7. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)

Predictability in an 'Unpredictable' Artificial Cultural Market

Paul Ormerod[1,*] and Kristin Glass[2]

[1] Volterra Consulting, London, UK
pormerod@volterra.co.uk
[2] New Mexico Institute of Mining and Technology, New Mexico, USA
kglass@icasa.nmt.edu

Abstract. In social, economic and cultural situations in which the decisions of individuals are influenced directly by the decisions of others, there is an inherently high level of ex ante unpredictability.

We examine the extent to which the existence of social influence may, paradoxically, increase the extent to which the choice which eventually emerges as the most popular (the 'winner') can be identified at a very early stage in the process. Once the process of choice has begun, only a very small number of decisions may be necessary to give a reasonable prospect of being able to identify the eventual 'winner'.

We illustrate this by an analysis of the music download experiments of Salganik et.al. (2006). We derive a practical rule for early identification of the eventual 'winner'. We validate the rule by applying it to similar data not used in the process of constructing the rule.

Keywords: social influence, unpredictability.

1 Introduction

Enormous resources are devoted to the task of predicting the outcome of social processes in domains such as economics, public policy, and popular culture. But these predictions are often woefully inaccurate. The two most striking characteristics of cultural markets, for example, are *inequality*, in that hit songs, books, and movies are many times more popular than average, and *unpredictability* (for example, Arthur 1989, Bentley et.al. 2007).

Consumer choice in such industries is governed not just by the set of incentives described by conventional consumer demand theory, but by the choices of others (Potts et.al. 2008), so that the payoff of an individual is an explicit function of the actions of others.

In Salganik et al. (2006), researchers constructed an online music market and examined the role social influence played in the songs which participants chose to download. The experiment revealed that increasing the extent to which participants were able to observe the selections of others led to an increase (decrease) in the popularity of

* Corresponding author.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 354–359, 2010.

the most (least) popular songs and a decrease in the predictability of song popularity based on quality. Experimental studies, such as those conducted in social psychology (Asch 1953) reach similar conclusions regarding the effects of social influence.

This paper examines the extent to which the existence of social influence may *increase* the extent to which winners can be identified at a very early stage in the process of consumer choices in a market. Section 2 describes the data, section 3 sets out some initial analysis, and section 4 derives a prediction rule.

2 The Data

The Salganik et al. experiment created an artificial 'music market' in which participants downloaded previously unknown songs either with or without knowledge of previous participants' choices. Increasing the strength of social influence increased both inequality and unpredictability of success.

We examined data for 18 experimental worlds, in each of which the same 48 songs were available for downloading. The detailed description of the available data for each of these worlds is available at http://opr.princeton.edu/archive/. In 16 of the worlds a social signal is present. In 8 of these worlds, the person making the choice of whether or not to download was given information on the previous number of downloads carried out by other people, with the songs sorted into popularity at that time. We denote these experiments as being 'strong positive externality process' or strong PEP for short.

Table 1. Various information on the distributions of the final outcomes of the experiments

mean/median	max	N	max/N	experiment
1.3	57	659	8.65	11
1.77	154	1021	15.08	12
1.24	81	834	9.71	21
2.02	158	968	16.32	22
1.39	65	733	8.87	31
1.96	114	892	12.78	32
1.13	66	871	7.58	41
1.7	165	1103	14.96	42
1.21	68	755	9.01	51
1.85	161	1109	14.52	52
1.16	61	944	6.46	61
1.96	135	941	14.35	62
1.17	69	1013	6.81	71
1.77	154	1149	13.4	72
1.14	44	819	5.37	81
2.27	179	926	19.33	82
1.09	77	1571	4.9	91
1	79	2193	3.6	92

In a further 8 worlds, the same information was provided, but it was not sorted into rank order. We denote these experiments as being 'weak positive externality process' or weak PEP for short. Finally, in two of the worlds there is no social signal at all, designated 'no PEP'.

Table 1 sets out information on the final outcomes in each of the experiments.

***Notes*:** Mean/median is the mean number of downloads across the 48 songs at the end of the experiment divided by the median. 'Max' is the number of downloads of the 'winner', the most frequently down loaded song, and N is the total number of downloads. Max/N is 'max' as a percentage of N. The final column is simply the identification tags we assigned to each experiment

3 A Heuristic Prediction Rule

Our aim in this paper is heuristic. Specifically, we examine whether a rule can be discovered which will enable *ex ante* the top ranked song at the end of each experiment to be identified. In other words, we are not trying to predict the exact number of downloads (or market share) at the end of each experiment, but to see if the 'winner' of each experiment (i.e. the top ranked song at the end) can be identified *ex ante*.

A key characteristic of processes of agent choice or selection in which the decisions of others are taken directly into account is that the final outcome of any such process will typically exhibit considerable right-skew (for example, Simon 1955, Bentley et.al. 2009).

We examine the data on a step-by-step basis and to see at what point the outcome could be regarded as exhibiting a right-skew distribution. For a Gaussian distribution, the theoretical mean of the data is equal to the theoretical median. Denote by MM the ratio of the mean to the median. In an empirical setting, MM may deviate from 1 even if the data are Gaussian especially in a small sample. But the deviation is extremely unlikely to be more than 1.1, even with a sample size as small as 20 (calculations available from the authors).

In contrast, in right-skew distributions, the theoretical MM is distinctly larger than 1. For an exponential distribution, with rate parameter λ, the mean is $1/\lambda$ and the median is $\log(2)/\lambda$. So the MM theoretically is $1/\log(2)$, or around 1.44. For a lognormal, where μ is the mean of the natural log of the variable and σ is the standard deviation, the theoretical median is $\exp(\mu)$ and the theoretical mean $\exp(\mu + \sigma^2/2)$, so again MM > 1. And for the power law, empirical estimates of MM will in general give a value > 1 even if the population mean does not exist.

4 Results

We therefore calculated the mean/median value at each step of each experiment (though in the very early stages this ratio does not exist given than the median number of download is zero). We averaged this across the 8 'strong' and 8 'weak' positive externality experiments and across 2 experiments with no such externality.

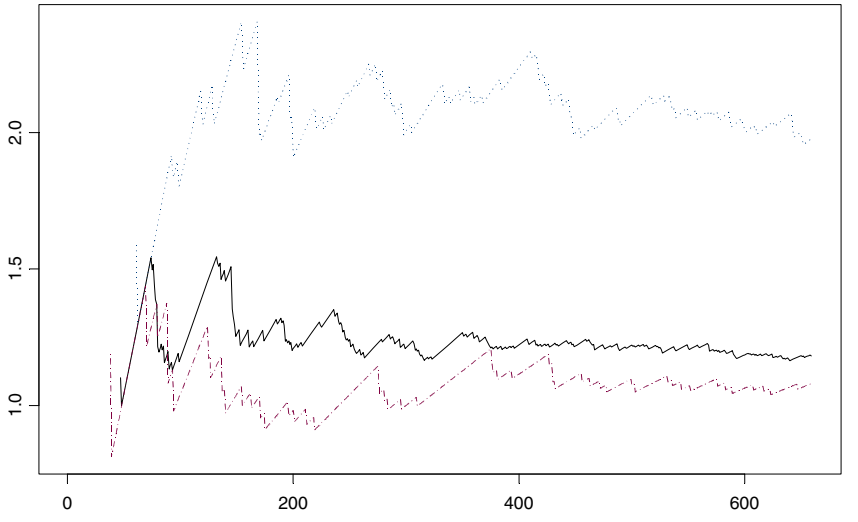


Fig. 1. Dotted line at top is average at step k of the mean/median of the 8 strong PEP experiments; solid line is average at step k of the mean/median of the 8 weak PEP experiments; broken line at bottom is average at step n of the mean/median for the 2 non-PEP experiments. For n close to zero, the median is zero. The data is plotted up to step 659, the length of the shortest experiment.

It is evident that at a fairly early stage in the process, the different types of experiment become differentiated using the mean/median criterion. The next question is therefore whether the empirical mean/median is a useful tool with which to make early identification of eventual 'winners' in the experiments.

As an initial exercise, we selected the first time in each experiment that the mean/median > 1.10, with the next step also being above 1.10. We compared the rankings at this step, τ say, with the rankings in the final step, N.

Specifically, we examined whether the eventual overall winner, the one with the most downloads at time N, can be identified in any way at time τ. Table 2 sets out information on this, along with the percentage of total steps in the experiment which corresponds to step τ.

Column 1 is identifies the experiment in the database we used, and the numbers have no significance as such. Column 2 shows the percentage of total steps in the experiment at which the mean/median > 1.10 for the first time. Column 3 shows the number of downloads of the market leaders at that time. Note that in general it is very small. Column 4 indicates by yes/no whether the winner at time N at the end of the experiment was also the unequivocal leader at time τ. Column 5 indicates by yes/no whether the winner at time N was one of a group of joint leaders at time τ, and column 6 shows the number of joint leaders at time τ.

In nine of the experiments, the eventual winner was either the unequivocal of the joint winner at time τ. Step τ as a percentage of the total number of steps (individual downloads) in the experiment varied between 3.06 and 8.05.

Table 2. Outcome of the use of the decision rule in identifying eventual winners

experiment	τ/N	maximum single download at step τ	winner at time N and winner at time τ	Joint winner at time N, winner at time τ	number of joint winners
11	8.04	2	no	yes	6
12	4.11	7	yes	n/a	n/a
21	5.4	9	no	no	n/a
22	4.75	3	no	yes	4
31	4.5	4	no	no	n/a
32	5.05	4	no	yes	3
41	3.1	2	no	no	n/a
42	3.63	4	no	no	n/a
51	7.02	4	no	yes	2
52	4.26	8	yes	n/a	n/a
61	3.5	4	no	no	n/a
62	5.74	4	no	yes	3
71	3.06	3	no	yes	2
72	5.31	13	yes	n/a	n/a
81	5.74	8	no	no	n/a
82	5.11	8	yes	n/a	n/a
91	2.42	3	no	no	n/a
92	1.41	3	no	no	n/a

In experiment 21, at step τ, where τ is 6.4 per cent of N, the eventual winner was placed joint second. In experiment 32, the eventual winner was third at step τ. The rule was less successful in the other experiments, but not completely without value.

Of the 8 experiments which exhibit strong positive externality processes, the winner at time N can always be identified very early, either unequivocally or as part of a small group, using the mean/median > 1.10 criterion. In addition, as mean/median evolves over time, it rapidly becomes apparent which experiments are strong positive externality processes.

So the simple statistic, the mean/median, appears to be a useful way of a) identifying at an early stage whether a process is governed in part by positive externalities in agent choice and b) identifying at an early stage in processes which do show evidence of positive externalities the choice which will eventually ‘win’ the process.

We checked the validity of the MM rule with 2 further data sets from Salganik which were not used in the process of generating the rule. These had older, more male, and more international participants that were recruited differently from those in the experiments used to develop the rule. So the two provide a useful test of the rule.

In one of the data sets, the eventual winner was also the winner at time τ, when $\tau/N = 6.17$. In the other, the eventual winner was ranked second at time τ, and the eventual second was the winner at time τ. In this case, $\tau/N = 5.64$.

5 Conclusion

In markets where social influence is important in determining whether or not an agent decides to adopt a particular mode of behavior or buy a particular product or brand, a large literature shows that successful *ex ante* prediction of the eventual winner is either very difficult or impossible.

However, the existence of social influence means that it is often possible to identify the eventual winner at a very early stage of the process of choice by participants in the market. We illustrate this with an analysis of the artificial cultural market created by Salganik et.al (op.cit.). We derive a rule for early identification of the eventual winner, which we verify by using it successfully on two further experiments which were not part of the data sets used to create the rule.

Acknowledgement

We are grateful for comments from three anonymous referees for the 2010 International Conference on Social Computing, Behavioral Modeling, & Prediction.

References

1. Arthur, W.B.: Competing technologies, increasing returns, and lock-in by historical events. Economic Journal 99, 116–131 (1989)
2. Asch, S.E.: Effects of group pressure upon the modification and distortion of Judgements. In: Guetzkow, H. (ed.) Groups, leadership and men: Research in human relations. Russell and Russell, New York (1953)
3. Bentley, R.A., Lipo, C.P., Hahn, M.W., Herzog, H.A.: Regular rates of popular culture change reflect random copying. Evolution and Human Behavior 28, 151–158 (2007)
4. Bentley, R.A., Ormerod, P., Batty, M.: An evolutionary model of long tailed distributions in the social sciences (2009), http://arxiv.org/abs/0903.2533
5. Potts, J., Cunningham, S., Hartley, J., Ormerod, P.: Social network markets: a new definition of the creative industries. Journal of Cultural Economics 32, 167–185 (2008)
6. 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)
7. Simon, H.A.: On a class of skew distribution functions. Biometrika 42, 425–440 (1955)

Improving an Agent-Based Model by Using Interdisciplinary Approaches for Analyzing Structural Change in Agriculture

Franziska Appel, Arlette Ostermeyer, Alfons Balmann, and Karin Larsen

Theodor-Lieser-Straße 2, 06120 Halle (Saale), Germany
appel@iamo.de

Abstract. Structural change in the German dairy sector seems to be lagged behind. Heterogeneous farm structures, a low efficiency and profitability are persistent although farms operate under similar market and policy conditions. This raises the questions whether these structures are path dependent and how they can eventually be overcome. To answer these questions we use the agent-based model AgriPoliS. The aim of our project is to improve assumptions in AgriPoliS by using it as an experimental laboratory. In a second part AgriPoliS will be used in stakeholder workshops to define scenarios for the dairy sector and communicate and discuss results to practitioners and decision makers.

1 Introduction and Motivation

Farms in the different parts of Germany operate under relatively similar market conditions and policy environments. In spite of that, a huge regional heterogeneity in terms of farm sizes and specialization is noticeable. This project focuses on the German dairy sector. One reason is that structural change in this sector seems to be particularly lagged behind. The very most dairy farms either operate with inferior techniques or apply them in a less economical way. Also the regional heterogeneity of farm structures is particularly large with, e.g., many small farms in the southern parts of Germany (e.g. Bavaria) and a relatively low number of large dairy farms in the north-eastern parts of Germany (e.g. Saxony-Anhalt). A second reason is that the dairy sector is particularly affected by the ongoing liberalization of the European Union's Common Agricultural Policy (EU CAP). Accordingly, dairy farmers, their representatives and politicians are highly concerned about the future of this sector. Lately the extension of the milk quota which should stepwise lead to a complete abolishment in 2015, causes a fall of the EU milk prices. Because of that, many dairy farmers are threatened in their existences and counteract the CAP reform via strong protests (Deutsche Welle 2009).

An important aim of this research is to analyze structural change in agriculture. Relevant questions are; what are the determinants of structural heterogeneity, is structural change path dependent (cf. David 1985, Arthur 1989, Balmann 1995) and can such path dependences be overcome? In this regard the question arises how fast the structure of the dairy sector is able to adjust to the new EU CAP and how farms

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 360–366, 2010.

can cope with that situation. To address these issues, it is aimed firstly to evaluate the behavioral foundation of agents (farmers) in the agent-based model AgriPoliS (Agricultural Policy Simulator, cf. section 2.1). This will be done through incentive-based participatory experiments with a modified version of AgriPoliS in which persons can take over the role of managing one of the farms while others remain computerized.

A second objective is to analyze structural change of two selected regions in Germany by using an adopted AgriPoliS version. The regions are heterogeneous in their agricultural structures and dairy farming plays a major role. Stakeholder platforms will be established in the study regions in order to discuss assumptions of AgriPoliS, to identify potential scenarios regarding policy instruments and general technological and economic trends, as well as to analyze and discuss simulation results.

2 Methods

2.1 Participatory Laboratory Experiments with AgriPoliS

AgriPoliS is an agent-based model in which structural change is modeled as an endogenous process (Happe 2004; Happe et al. 2005). It can be viewed as an "experimental laboratory" for analyzing structural change in agriculture and builds on Balmann (1995; 1997).

The main idea of the model is to consider the interactions between different agents (farmers) which, in turn, will affect their actions. In AgriPoliS farms have an endogenously evolving factor endowment and interactions are captured in the markets for products, labor, capital, land and quotas. As in Balmann (1997) and Berger (2001), the behavioral foundation in AgriPoliS is that farms are assumed to maximize profits or farm household income. This is implemented in a normative way by using mixed-integer programming where decision making is rational but myopic and non-strategic and expectation formation is generally represented as adaptive (considering trends).

The determinants of path dependence in AgriPoliS include sunk costs, frictions on land market and policy incentives to stay in business (Balmann 1995). Also the behavioral objective to maximize farm income may be a reason for path dependence in AgriPoliS as agents may ignore long-term trends. However, even a small individual divergence from rational behavior or selfishness can affect the behavior of all other agents and therefore influence market developments (Ockenfels 2009).

In the current version of AgriPoliS, agents are assumed to maximize profits or incomes (homo economicus) but act myopically. As a result, path dependence can be affected by the agents' behavioral model. Until now, several alternatives exist to the assumption of myopic and non-strategic income maximization of AgriPoliS: e.g. a global optimization over all farms within AgriPoliS (Kellermann and Balmann 2009) and genetic algorithms (Kellermann and Balmann 2009). The question is, however, whether it is realistic to assume such high rationality. Whether the computer agents in AgriPoliS are "smart enough" with regard to strategic decisions can be discovered by using a fundamentally different approach: to directly include persons into AgriPoliS

and let them replace the computerized agents. Such approaches can be understood as a kind of role playing game (e.g. Barreteau et al. 2001 and Bousquet et al. 2002). Therefore, human players have to be able to participate in AgriPoliS and compete against the computer agents. The players will have to decide on production, investment, renting land, buying quota, continuing or giving-up farming. For this purpose AgriPoliS has to be converted to a hybrid between role-playing game and management game. An approach of coupling agent-based simulations with role-playing games can be found in Barreteau and Bousquet (1999).

In the proposed agent-based participatory simulations, both agents and players will be exposed to different economic, technological and policy scenarios. Observed differences between the behavior of players and that of computer agents provides a starting point for further analyses of behavior-based path dependence. It furthermore facilitates improvements of AgriPoliS, so that the agents display more realistic behavior. Pahl-Wostl and Ebenhöh (2004) elaborate on how to represent human behavior in agent based models and suggest specific attributes and heuristics.

Guyot and Shinichi (2006) discuss several steps of building an agent-based participatory simulation. The first step is to build a "domain model" which is in our case AgriPoliS. In a second step this "domain model" has to be converted such that prospective players can understand it. Guyot and Shinichi (2006) refer to this converted model as "design model". Kellermann (2002) developed PlayAgriPoliS which can be understood as such a "design model" of AgriPoliS. PlayAgriPoliS was developed to provide a tool which allows accessing AgriPoliS in an easy and intuitive way with regards to education and teaching purposes. In PlayAgriPoliS the player takes over the role of an agriculture minister who has to fulfill pre-election promises with a bounded budget. The player in PlayAgriPoliS does not make decisions on farm-level. In our case, however, we want real persons to manage a farm in AgriPoliS (i.d. to replace an agent). Therefore, a new design model has to be developed. In a last step, the design model is modified to the actual agent-based participatory simulation, which hereafter will be referred to as game.

There are several important aspects that have to be considered when designing this game. First of all, the model must be simple enough to allow easy and quick assess to the game (Barreteau et al. 2001). Despite the necessary simplifications, it must at the same time be realistic for the participants in order to adopt realistic behavior (Guyot and Shinichi, 2006). Another challenge is to provide the players with all necessary information without assailing them with to much data. For this purpose a useful interface has to be designed. An interface similar to that of PlayAgriPoliS will be constructed in which the player is given a "sector report" as basis for his/her decisions. This report shows all relevant key data on the basis of a representative sample of farms. The interface furthermore allows the player to have a closer look at representative farms in detail. This provides the basis for the player to compare his/her farm with others. In addition, the player is provided with balance sheets of his/her own farm.

The decision tools are another important part of the interface. The agents can rent land, invest in human capital, stables or machinery, produce crops or livestock, work off-farm or quit farming. As we are especially interested in strategic planning and decisions, such as enlarging the farm or investing in a new branch of production or

exiting, the final production program itself will be defined by linear programming which will also deliver information for renting and investment decisions.

When designing these kinds of games it has to be ensured that the players do not lose their motivation to participate. Bean (2001) emphasizes characteristics for this purpose. First of all, it must be fun for the player to participate. One thing that makes participation entertaining is competition. In our case the player is incited to perform better than the computer agents. More important is that the simulated situation is realistic enough, so that the players act as they would have acted in reality. Moreover, the players receive a payment according to their economic success in the game. The second important characteristic is accessibility. The game should not last longer than planned (Bean 2001). Guyot and Shinichi (2006) also mention that the presence of the organizer often is necessary to support the players in handling the interface. Finally, clarity is necessary. That means the player should know in the beginning of the game what the content and aim of the game are. It is also important to choose a terminology which is easily understandable. As Bean (2001) advises, several tests for usability are planned during the game development.

After developing the game, laboratory experiments will be carried out in which a single player take over the role of a manager of a specific farm while all other agents remain computerized agents of AgriPoliS. This allows the comparison of each experiment with the outcome of a standard simulation in which the replaced agent remains computerized. As mentioned above, in order to create incentives for each player to maximize income, the players get an honorary according to their economic success in the game. Two types of players are considered: on the one hand master and PhD students and on the other hand farmers and agricultural experts. After each experiment, we will present the players the results of the simulations and ask for their strategy and reasons for some important decisions. The experiments will be analyzed in order to identify characteristic strategies and how they deviate from standard simulations as well as their impact on the farm's evolution and the whole sector. Moreover it is aimed to analyze whether different types of players operate in different manners. Finally, the outcomes will be analyzed regarding the question whether and how the standard agents should and could be adapted to get more realistic simulation results. The players' observed behavior that differs from the original behavioral assumptions of the model is used to improve the decision routines in AgriPoliS (cf. Guyot and Shinichi 2006).

2.2 Participatory Analyses of the German Dairy Sector

The next step is to analyze the opportunities to overcome path dependence of structural change and problems of CAP liberalization on a sector level. Therefore, AgriPoliS will be adapted to two regions in Germany: Allgäu (Southern Bavaria) and Stendal (Northern Saxony-Anhalt). While dairy farming in the Allgäu region is small-scaled, farms in the Stendal area are predominantly large-scaled. The different natural, historical and economic conditions in the two study regions result in differentiated structural problems. Whilst dairy farms in the Allgäu are inhibited in their growth due to high competition, farms in Stendal are confronted with other challenges such as the high share of external factors , e.g. credit capital (while equity

capital is low), rented land and hired labor. Hence, the farms in the Stendal area are more susceptible to risk.

To what extent dairy farmers are able to overcome structural deficits and to what extent subsidies and quotas solidify the structural problems of the dairy sector is analyzed with the improved model AgriPoliS. The different scenarios to be analyzed will be developed using participatory methods. In doing so, the knowledge, experiences, and different views of local actors will be utilized. A number of stakeholder workshops in the regions are therefore planned to identify conflicts relating to structural change in the dairy sector, to develop the scenarios of interest, and to discuss results and policies.

Stakeholders can be found on three levels. There are directly affected persons – in our case the farmers. Also the views of a more objective level, e.g. representatives of the food chain and agricultural associations, consultants, and regional agricultural authorities are of relevance. On a third level an interested public, politicians, tax payers and other interest representatives will be involved in the scenario building. Thus, different views of local actors can be covered and a comprehensive picture of impacts on agricultural structural change in the regions can be gained.

In the workshops stakeholders can build scenarios by deciding e.g. how much subsidies farmers get or if there are restrictions regarding the farm or herd size etc. Thinkable scenarios will be related to the dairy sector. Therefore, the milk quota regulation is particularly relevant. Furthermore, policy instruments such as the agricultural investment promotion program, compensation payments for less-favored areas, or the option of payments for land release are possible instruments which can be analyzed to determine their impact on structural change. The policy scenarios are accompanied by the assumption that direct payments of the CAP will be reduced after 2013 as funds from the first pillar will further increasingly be shifted to the second pillar. After developing scenarios, AgriPoliS will be used to run simulations. The results will be presented to the stakeholders in a second meeting in order to discuss them and to adjust the scenarios for a second set of simulations. The strategic aim is an ongoing networking with relevant stakeholders for the project phase. The establishment of stakeholder platforms allows a continuous exchange between research, practice and politics. The agent-based model AgriPoliS can thereby be improved and adopted to real conditions in the regions by the planned queries to the stakeholders. In return participants can gain insights in determinants of structural change.

3 Conclusion

To clarify the objectives of our research they are summarized in the following:

In a first step we want to improve AgriPoliS in a way to make it more realistic in terms of behavioral foundations. By using agent-based participatory simulation, we want to get hints about how real farmers act compared to computer agents. With the results of these simulations we want to answer the following questions:

- What is the impact of the agents' behavioral rules and to which extend is the outcome of simulations with AgriPoliS dependent on them? We want to make sure for further use of the model that results are not ascribed to wrong behavioral assumptions.

- Are agents modeled as too naive (i.e. long-term trends are ignored) and unable to overcome structural deficits? Have real persons, in contrast, a better ability to deal with a changing economic environment (cf. via strategic thoughts and actions, acting anticipatory, awaiting, risk awareness)?
- How can AgriPoliS be used to analyze structural change in the German dairy sector? How "smart" should the agents be modeled when analyzing structural change? Ockenfels (2009) argue that even a small individual divergence from rational behavior or selfishness can affect the behavior of all other agents and therefore influence market developments. Since we want to analyze structural change and its constraints, we have to allow the agents to depart from the as hitherto uniform myopic and non-strategic behavior.

In the next step we want to answer especially following questions by discussing with stakeholders about structural problems in the dairy sector:

- How are dairy and other farms affected by policy instruments and measures? – On a sector level we want to analyze which policy instruments and measures concern (dairy) farmers to which degree. Regarding the ongoing liberalization of agricultural markets we furthermore want to analyze if dairy farmers and other stakeholders of the dairy sector follow specific mental models which create specific concerns against this liberalization?
- To what extent will a changing policy environment (for example the abolishment of milk quotas) affect structural change within the dairy sector? – In a second step we will discuss possible effects of changing policy measures and implementing new ones. This policy analysis will be used to discover instruments which encourage path dependence of farm structures.
- How can assumptions, scenarios and results be communicated between researchers and stakeholders? – Finally, concerning a long-term aim, we plan to communicate our results to stakeholders, especially to policy makers. The establishment of a network with regional stakeholders will give us the opportunity to orient our research to practical topics and needs. Through our participatory analyses we might find reasonable and feasible policy measures to solve structural deficits which can be used in policy advices.

As mentioned in the introduction, dairy farmers are confronted with a decreasing milk price and simultaneously increasing factor prices especially for fodder. The political changes regarding an ongoing liberalization of agricultural markets and the related abolishment of the milk quota in 2015 provide challenges in view of the competitive position of the individual farm. The particular aim is to analyze how to overcome structural deficits in agriculture. By showing the results to farmers and discuss with them we intend to broaden their insights into structural change, impacts of policy instruments, and the role of mental models.

References

Arthur, W.B.: Competing Technologies, Increasing Returns and Lock-In by Historical Events. The Economic Journal 99, 116–131 (1989)

Balmann, A.: Pfadabhängigkeiten in Agrarstrukturentwicklungen - Begriff, Ursachen und Konsequenzen. Duncker & Humblot, Berlin (1995)

Balmann, A.: Farm-Based Modelling of Regional Structural Change. European Review of Agricultural Economics 25(1), 85–108 (1997)

Barreteau, O., Bousquet, F.: Jeux de rôles et validation de systèmes multi-agents. In: Gleizes, M.-P., Marcenac, P. (eds.) Ingénierie des systèmes multi-agents, actes des 7èmes JFIADSMA. Hermès (1999)

Barreteau, O., Bousquet, F., Attonaty, J.-M.: Role-playing games for opening the black box of multi-agent systems: Method and teachings of its application to Senegal River valley irrigated systems. Journal of Artificial Societies and Social Simulations 4(2) (2001), http://jasss.soc.surrey.ac.uk/4/2/5.html

Bousquet, F., Barreteau, O., d'Aquino, P., Etienne, M., Boissau, S., Aubert, S., Le Page, C., Babin, D., Castella, J.-C.: Multi-agent systems and role games: An approach for ecosystem co-management. In: Janssen, M. (ed.) Complexity and ecosystem management: The theory and practice of multi-agent approaches, London, Edward Elgar, pp. 248–285 (2002)

Bean, M.: The Four Key Attributes of Successful Training Simulations (2001), http://forio.com/resources/face/ (30.07.09)

Berger, T.: Agent-based spatial models applied to agriculture. A simulation tool for technology diffusion, resource-use changes and policy analysis. Agricultural Economics 24(1), 85–108 (2001)

David, P.A.: Clio and the economics of QWERTY. American Economic Review 75(2), 332–337 (1985)

Deutsche Welle: German dairy farmers protest against low milk prices (2009), http://www.dw-world.de/dw/article/0,4277853,00.html (25.05.2009)

Guyot, P., Shinichi, H.: Agent-Based Participatory Simulations: Merging Multi-Agent Systems and Role-Playing Games. Journal of Artificial Societies and Social Simulation 9(4) (2006)

Happe, K.: Agricultural policies and farm structures - Agent-based modelling and application to EU-policy reform. Studies on the Agricultural and Food Sector in Central and Eastern Europe 30, IAMO (2004)

Happe, K., Balmann, A., Kellermann, K., Sahrbacher, C.: The use of agent-based modelling to establish a link between agricultural policy reform and structural change. In: Arfini, F. (ed.) Modelling Agricultural Policies: State of the Art and New Challenges, pp. 138–165 (2005)

Kellermann, K.: PlayAgriPoliS – Ein agentenbasiertes Politikplanspiel. Diploma thesis, Humboldt-Universität, Berlin (2002), http://www.iamo.de/PlayAgriPoliS/diplomarbeit.pdf

Kellermann, K., Balmann, A.: How smart should farms be modeled? The behavioral foundation of bidding strategies in agent-based land market models. Journal of Operations and Quantitative Management, Special issue: Decision Making in Complex Systems (2009) (conditionally accepted)

Ockenfels, A.: Marktdesign und Experimentelle Wirtschaftsforschung. Perspektiven der Wirtschaftspolitik 10(special issue), 31–53 (2009)

Pahl-Wostl, C., Ebenhöh, E.: Heuristics to characterise human behaviour in agent based models. In: iEMSs 2004 (2004), http://www.iemss.org/iemss2004/pdf/abm/pahlheur.pdf

Exploring the Human Fabric through an Analyst's Eyes

Nadya Belov[1], Jeff Patti[1], Saki Wilcox[1], Rafael Almanzar[1], Janet Kim[1], Jennifer Kellogg[2], and Steven Dang[2]

[1] Lockheed Martin Advanced Technology Laboratories
3 Executive Campus, Suite 600, Cherry Hill, New Jersey, United States of America
{nbelov, jpatti, swilcox, ralmanza, jkim}@atl.lmco.com
[2] Lockheed Martin Information Systems and Global Services,
700 North Frederick Avenue, Gaithersburg, Maryland, United States of America
{jennifer.l.kellogg, steven.c.dang}@lmco.com

Abstract. The nature and type of conflicts drastically changed in the last half of the twentieth century. Wars are no longer limited to the field; they are supplemented with guerrilla warfare and other asymmetric warfare tactics including domestic terrorism. Domestic terrorism has demonstrated a need for improved homeland security capabilities. Establishing and maintaining the understanding of the key players and the underlying social networks is essential to combating asynchronous warfare tactics. Herein, we identify the key challenges addressed by our Collection/Exploitation Decision System (CEDS) that assist analysts in maintaining an up-to-date understanding of dynamic human networks.

Keywords: dynamic social network modeling.

1 Introduction

The nature of military and homeland conflicts has changed drastically in recent years. Traditional adversarial tactics have given way to guerrilla warfare and terrorism. Success in counterterrorism, whether as part of military conflict or thwarting an individual, depends on the development and maintenance of accurate situational awareness. The success of maintaining accurate situational awareness in theaters faced with asymmetric attacks depends on the ability to reason about dissimilar information that provide a coherent human fabric [7]. Given the dynamic and unpredictable nature of asymmetric warfare, situational awareness relies heavily on the understanding of the key players and the underlying social networks. Effective collection and exploitation of intelligence information from a variety of sources poses a significant challenge enabling accurate knowledge of the underlying human networks. Lockheed Martin is developing the Collection/Exploitation Decision System (CEDS) to address the above collection and exploitation challenges.

With the advent of small, fast computing devices and mobile applications (e.g., twitter), crowd truth is quickly replacing ground truth that may not be available due to the nature of the sensor—biological, biased human sensors [6]. Lockheed

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 367–374, 2010.

Martin has developed CEDS under an internal Human Network research effort to explore the impact [a socio-cultural understanding of the local population] where the military operates during counterinsurgency and stability operations. The knowledge of the human terrain is as important in these operations as knowledge of the physical terrain. Understanding of the local people, their customs and value systems yields a clearer picture of their needs, motivations and intentions. This knowledge is a key component to effective counterinsurgency operations and enables the establishment of effective local partnerships.

We present an operational scenario that highlights the motivation for a system such as CEDS. We discuss the key technologies leveraged in the development of the system. We conclude with a discussion regarding future work.

2 Motivation

Imagine military operations where soldiers or law enforcement operations with police on patrol report on their surroundings, people and events in a natural manner. The reports are augmented with biometric identifiers ensuring positive identification and increasing overall accuracy. Patrol reports are spatiotemporally tagged using context and the soldier's exact location. Developing an intelligence report would entail a combination of a few spoken utterances tied to a quick camera snap.

Imagine a soldier using a video camera pans across a congregation of people to send a video of the crowd to his operations command for identification and review. The operations command would use software to automatically extract and identify individuals from this video. Upon the combination of these intelligence reports and other open source data from a variety of sources, analysts can apply models to predict potential future key instability events in the same region. These capabilities will support operations both in conflict areas and at home.

3 CEDS

One of the main challenges in attaining an accurate understanding of the situation is the timely and accurate acquisition of intelligence reports from the field. Belov and Gerken discuss a mixed-initiative fusion associate system that provides intelligence report authoring, querying, fusion and persistence [1]. However, their system relies on human intervention to help ascertain confidence of certain reports and fuse them into the existing human network. The CEDS builds on [1] and further reduces human manual intervention by leveraging biometric data to support positive identification. CEDS provides operators with robust intelligence collection and human network exploration capabilities.

CEDS (see Fig. 1) integrates three previously developed systems: the Biometric Analysis and Identification System (BAIS) (see Section 3.1), the Core Human Network System (CHANS) (see Section 3.2), and the Spoken Language Integrated Environment (SLICE) (see Section 3.3) to deliver a comprehensive collection and exploitation system.

Recognizing that the size, fit, and weight of a computing device is a concern of the warfighter, we re-architected both BAIS and SLICE. We created client-server applications and took steps toward deploying the client on a small mobile computing

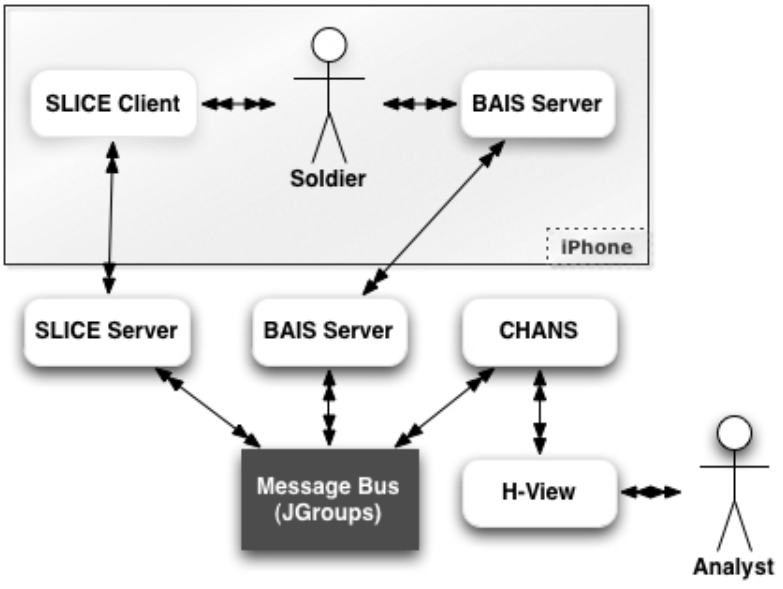


Fig. 1. The architecture of the Collection/Exploitation Decision System (CEDS)

device such as the iPhone™. Both SLICE and BAIS clients communicate with their respective servers via custom messaging protocol. We leverage JGroups for communication among the SLICE, BAIS and CHANS systems. Reports made using SLICE as well as BAIS's positive identifications and new person enrollments are communicated to CHANS and in turn fused and displayed to the analyst via H-View.

3.1 Biometric Analysis and Identification System (BAIS)

Biometric data has been used in many public venues, such as airports, for programs like CLEAR. Additionally, the Transportation Security Administration (TSA) lists biometrics as a key innovation and technology thrust. The TSA already mandates pictures and fingerprints of all foreign visitors upon arrival to the United States of America.

The challenge of using biometric data has shifted from collection to exploitation. There exist many unique biometric identifiers (e.g., fingerprints, retina scans) and numerous sensors to acquire the identifiers. Success in using biometric data is defined by the ability to find the right information when needed. Lockheed Martin Information Systems and Global Services (LM IS&GS) business area has developed several biometric intelligence solutions with internal funding that combine grounded analytic techniques with biometric and knowledge management data to create comprehensive human profiles.

CEDS leverages LM IS&GS's BAIS, which links disparate databases to crosscheck and verify a person's identity using multiple biometric modalities. BAIS uses a client-server architecture allowing the installation of the light client onto a mobile phone. A user can use the phone's camera to take a picture and send it to a server for identification or enter further personal information (e.g., name, height, weight, eye color).

CEDS also uses BAIS with the Faces in a Crowd video analysis system. The Faces in a Crowd system, also developed by LM IS&GS, provides the ability to track and extract faces from a video using advanced real-time facial detection techniques. The resulting face images are supplied to BAIS for identification. The BAIS technology, especially when paired with Faces in a Crowd, provides CEDS with a powerful, real-time identification capability.

3.2 Core Human Network System (CHANS)

Successful discovery and tracking of individuals and groups is dependent upon the ability to aggregate and persist dynamic updates from a variety of sources from the field. Lockheed Martin Advanced Technology Laboratories (LM ATL) has developed the Core Human Network System that provides the capability to persist the dynamics of human networks [5]. CHANS enables continuous updates and development of effective human and entity relationships through the integration of disparate, human-generated information sources and network analysis technologies. The dynamic human network data stored in CHANS is represented in CoreNetML, an edge-rich, spatiotemporally attributed representation. CoreNetML allows the specification of multiple relations between any two nodes or entities. In doing so, CoreNetML recognizes that entities, especially people, may have many simultaneous relationships, each unique in nature and with varying strengths.

As a decision aid to an intelligence analyst, CHANS visualizes the networks and their ever-changing dynamics through a component called H-View. H-View (see Fig. 2) enhances the exploration and analysis processes by facilitating manipulation of the nodes and edges and intelligent filtering based on context. Context specification allows an analyst to filter networks based on the nature of relationships, locations of entities or activities and even the nature of the reports that contributed to the evolution of the human network.

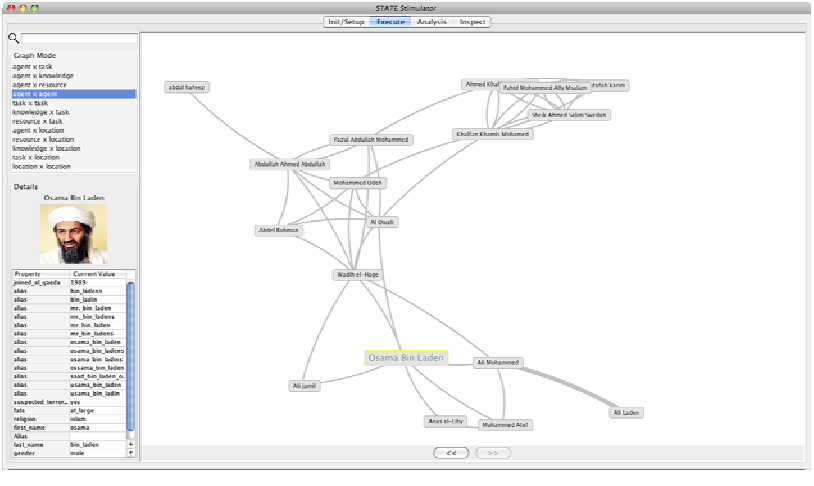


Fig. 2. H-View visualization component of the Core Human Network System (CHANS)

The human network information persisted in CHANS is detailed and can be augmented with a variety of attributes, including activities, previous locations, information certainty, and biometric values. In CEDS, CHANS uses the biometric information made available by BAIS (see Section 3.1) as well as spatiotemporal and activity information supplied through SLICE reports (see Section 3.3)

3.3 Spoken Language Integrated Environment (SLICE)

LM ATL has developed a system that allows a warfighter on the ground to share information by generating on-the-go observation reports. The system, Spoken Language Integrated Environment (SLICE), was developed using internal research funds. SLICE allows a warfighter on the ground to share intelligence information by generating and querying reports. Reports include information about activities, persons and other entities, such as vehicles and IEDs. The design and development of the system leverages the Interaction Design and Engineering for Advanced Systems (IDEAS) process [2] by combining the best practices from the User-Centered Design (UCD) theory with software development throughout the entire engineering process. The IDEAS process leverages Subject Matter Experts (SMEs) to guide the entire design and development processes.

The success of any multi-modal system where speech is one modality depends on the robustness of the speech recognition and synthesis components. SLICE leverages a grammar-based, speech recognition system developed as part of LM ATL's Wearable Intelligent Reporting Environment (WIRE) [3, 4] system with a domain-specific grammar and dictionary. SLICE further improves usability through its use of a confirmation strategy. The confirmation strategy allows the system to interpret the user's utterance and then re-synthesize the interpretation.

The CEDS effort leveraged SLICE for speech-based intelligence reporting. As part of the effort, SLICE was ported to a mobile phone platform, specifically an iPhone™. The port to a mobile platform forced a shift from the standalone architecture to a client-server one. The client, responsible for recording and sending the utterances to the back-end server, resides on the iPhone™. The client is implemented as a typical iPhone™ application using the Cocoa development paradigm. The shift in the architecture paradigm from stand-alone to a client-server is necessitated by lack of speech-to-text interpretation software for the iPhone™. The previously described confirmation strategy is also employed in iPhone™-based SLICE system using a text-based re-iteration of the report. An iPhone™-based SLICE system demonstrates a powerful solution on an inexpensive computing platform that can be quickly deployed to a large force population.

4 CEDS: Concept of Operations

The following scenario demonstrates exemplars of improving situational awareness using collection and exploitation technologies provided in CEDS described in Section 3. The scenario describes how two police patrols each supplied with a handheld equipped with CEDS can team with an analyst to share information and create an improved operational picture. An analyst stationed in the police

department's command center is able to find hidden links among people and activities based on the intelligence information collected in the field. Newly assessed intelligence information helps the analyst guide the actions of the patrols.

4.1 Morning Patrols at the Orlando Airport

Multiple mobile phones are configured with CEDS and deployed to a police squad for patrolling the Orlando Airport. The police department received an anonymous call informing of a possible attack on one or more Wally's World theme parks. When the mission begins, the officers are asked to perform a patrol and recon the Orlando Airport area. The police offers use CEDS to create reports, via SLICE, detailing any suspicious activity or persons arriving at the airport. Should the officers question anyone, they can use CEDS, via BAIS, to identify well-known suspicious individuals or enroll new ones based on observed activity.

At 9:09 a.m, the patrol notices a suspicious meeting between two men at a cafe just outside the arrivals lounge in the airport. The officer uses his phone to take a video of the meeting from a distance and sends the video feed accompanied by a report detailing activity, location and time to the police headquarters.

4.2 Police Headquarters

An analyst stationed in the operations center at the police headquarters receives reports as they are generated from all patrols equipped with CEDS. The analyst receives the video feed generated by the patrol at the Orlando Airport and uses Faces in the Crowd and BAIS (see Section 3.1) to extract faces and identify the suspect individuals.

The analyst is notified through H-View that one of the individuals in the video feed is a well-known terrorist. The other individual has not previously been enrolled in BAIS and cannot be identified. As a result of this information, the analyst radios the patrol at the airport and the patrols dispatched to all the Wally's World theme parks to be on the lookout for the known terrorist. The analyst further instructs to if possible, take a picture and interview the terrorist's associate.

4.3 Afternoon Patrols at the Wally's Kingdom

Based on the morning's reports and instructions received from the police headquarters, the mission is to be on the lookout for the well-known terrorist. Once located, the terrorist's associate is to be questioned and enrolled into BAIS.

Two officers patrolling the main entrance to the Wally's Kingdom notice suspicious activity near the ticketing booth. They see three men, each holding an identical, black duffel bag, involved in a hushed discussion. The officers approach the group and request to ask a few questions. The three men agree to the interview with hesitation.

One of the officers asks to take a picture of each of the individuals. The pictures are supplied to the officer's clients for identification and return significant results: two of the three individuals are well-known, dangerous terrorists. Furthermore, one of the two well-known terrorists is the same one spotted earlier in the day at the Orlando Airport.

The officer composes a report detailing his findings and sends it to police headquarters. While an analyst is assessing the report, the officer proceeds to interview the unknown associate inquiring about his name and origin and destination Back at the center of operations in the police headquarters, upon the receipt of the report from the patrol in the Wally's Kingdom, the analyst radios instructions to detain the group and bring them in for questioning.

4.4 Back at the Police Headquarters

The Wally's Kingdom police patrol returns to the police headquarters with the detained individuals. They are interviewed, and their belongings are search. The search reveals various bomb-making materials in significant quantities ready to be mixed and detonated on-demand.

5 Concluding Remarks and Future Work

The recent rise of insurgent organizations and significant increase in unconventional attacks they mount has forced the armed forces and homeland security to expand their intelligence processes to include real-time, field-based reporting, as well as tracking and reasoning about human networks. Often, the motivations and operations of insurgent groups are not well understood. The development and maintenance of an accurate operational picture provides a significant advantage to intelligence analysts in preventing such attacks. CEDS brings together three core systems, each with a clear role in collection and exploitation of human networks. Through the use of CEDS, users in the field are able to quickly identify persons of interest and report on suspicious activity. Through H-View, CEDS, equips analysts with an improved exploration capability through rich visualization of human networks and their properties.

LM ATL plans to continue developing technologies that improve the exploration and analysis of human networks. The next step in the development of CEDS is its augmentation with visual analytics that involve the analyst in the development and assessment of the human network. To the analysis process, the analyst adds deeper understanding of the underlying culture and motivations as they may evolve.

References

1. Belov, N., Gerken, P.: Mixed Initiative Soft Fusion Data Associate. In: Twelfth International Conference on Information Fusion, pp. 1897–1902 (2009)
2. Regli, S.H., Tremoulet, P.D.: IDEAS: Interaction Design and Engineering for Advanced Systems. American Psychological Association, Division 21, A New Collaborative Frontier: Innovative Approaches and Applications, Fairfax, VA (2007)
3. Orr, M., Hastie, H., Miksch, D., Flanders, J., Corrad, C.: Capturing Critical Information Using the Wearable Intelligent Reporting Environment (WIRE). U.S. Dept. of Homeland Security, Science and Technology Directorate, Boston, MA (2005)

4. Craven, P., Orr, M., Hastie, H.: WIRE: A Wearable Spoken Language Understanding System for the Military. In: Human Language Technology Conference (HLT) North American Chapter of the Association for Computational Linguistics (NAACL) Workshop, Rochester, NY (2007)
5. Belov, N., Martin, M.K., Patti, J., Reminga, J., Pawlowski, A., Carley, K.M.: Dynamic Networks: Rapid Assessment of Changing Scenarios. In: Second International Workshop on Social Computing, Behavior Modeling, and Prediction, Phoenix, AZ (2009)
6. Pentland, A.: Reality Mining of Mobile Communications: Toward a New Deal on Data. In: Behavior Modeling, and Prediction, Phoenix, AZ (2009)
7. Waltz, T.: Understanding the Human Terrain by Fusion of Social and Geospatial Data 3(4), 9–11 (2007)

Mitigating Issues Related to the Modeling of Insurgent Recruitment

Erica Briscoe, Ethan Trewhitt, Lora Weiss, and Elizabeth Whitaker

Georgia Tech Research Institute, Atlanta, Ga, USA
{Erica.Briscoe,Ethan.Trewhitt,Lora.Weiss,
Elizabeth.Whitaker}@gtri.gatech.edu

Abstract. Modeling the specific motivations and influences related to an individual's decision to become involved in insurgent warfare presents its own collection of unique challenges. The difficulty of the problem often necessitates simplifications that, while making the task more manageable, may inadvertently 'smooth away' critical aspects of the problem. Augmenting the challenge is that research into the motivations of terrorism has found there is *not* a definitive set of variables that serve as reliable indicators of an individual's involvement. This paper addresses techniques aimed toward mitigating issues that manifest in the modeling of insurgent recruitment so that these complications do not lessen the viability of models that are used in the prediction and evaluation of terrorist activity.

Keywords: Human Social Cultural Behavioral Modeling, Cognitive Modeling, Terrorism, Psychology, Insurgent Behavior, Recruitment.

1 Introduction

One of the principal contemporary challenges of the military is developing effective methods of dealing with the rise and persistence of insurgent movements aiming to undo the efforts of Western nations in Iraq and Afghanistan by attempting to unify the support of local populations. Past and recent experience in these Muslim states has indicated that these insurgencies, which are active at both a global and local level, are especially "complex, unstable, and harder to comprehend than purely national insurgencies" [5]. For many, the war on terror appears endless and irrational, unwinnable through force alone, which is as "unrealistic as attempting to prevail in a struggle against hate or racism by military means" [1]. This incredible complexity has led many to consider alternatives to traditional warfare tactics, particularly through the use of models of human behavior to aid in the development of methods for more effective interaction with adversaries and for use in analysis of military situations. The hope is that if adverse behavior is represented in a proper modeling framework, critical causal relationships and patterns of behavior will emerge that can then be used to inform counterterrorism strategies. This research into modeling of human behavior, especially that for military purposes, has emphasized the need for a varied set of modeling tools, each approach exhibiting its own strengths and weaknesses (see Zacharias, MacMillan, & Van Hemel [16] for an excellent overview).

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 375–381, 2010.

An attractive avenue for reducing the number and severity of terrorist activities is the prevention of people becoming involved in terrorism from the onset, intervening far to the "left of boom". The decision of an individual to become involved in terrorism depends on the environmental context and various precipitating factors. This paper identifies methods to address problems specific to modeling the motivations and influences related to an individual's decision to become involved in terrorist or insurgent behavior.

2 Issues with Modeling Behavior of Insurgents

It seems intuitive to adapt long-standing and well-researched empirical models of human behavior to the application of reducing insurgent behavior; however, there is a dangerous assumption of generalization in this practice that may lead to an amalgam of issues, the most critical of which are described below.

(i) *Theories based on controlled settings*. The use of psychological theories to explain the decision to engage in terrorism within a construct of individual psychology substantially relies on psychological models that are based on work that has been conducted in controlled settings or for targeted categories of individuals. An often-heard criticism of computational models based on experimental psychology is that they are limited to behaviors within this specific controlled context. While it may be feasible to claim that these results extend into 'real-life' situations, the application of behavioral models created with predominantly western approaches to insurgent behavior in the middle-east may be an unrealistic stretch. This extension is a particular problem in the application of decision-making models that attempt to understand how individuals decide to become involved in insurgent activity based on unknown preferences and influences.

(ii) *Obtaining data from largely inaccessible populations*. The severity of a lack of generalization may be lessened through the assertion that while a particular model may not exactly 'fit' a novel population, the computational framework itself is generalizable (under the broad assumption that human behavior models generalize to all humans). Even assuming that the theories are somewhat applicable, a much more significant question is whether it is even possible to have the data necessary to fit models to largely inaccessible populations. Though their use seems intuitive, traditional models were created with the assumption that an adaption will be possible, when given the right type and the right amount of data. This is more than just an issue of a mere lack of data, such as that which could conceivably be acquired given the right tools and access. Knowing that data is not available, *a priori*, certainly brings into question whether it is worthwhile to force fit the data that we *can* obtain into models that were created with the explicit assumption that other data could be acquired. This problem begs the question as to whether new models, created specifically with the type of data that *is* available, are a necessary realization of insurgent behavior modeling.

(iii) *Variability in qualitative variables*. Information concerning the motivations and influences on terrorist behavior to be used in modeling usually must be derived from a variety of sources, as there is a noticeable lack of empirical datasets for modeling [8].

Newspaper articles, government reports, and expert interviews are some of the most fruitful sources of material. Unfortunately, most of this information is descriptive and qualitative, rather than in a form that could easily be translated and applied within a quantitative framework. For example, some believe that, were the economic conditions of communities in Iraq improved, the extent of radicalization of the population would be decreased [10]. How much difference an improvement in economic conditions would make on radicalization is a very ill-defined and therefore makes the predictive modeling of the effect of economic policy difficult. This problem is compounded by the variability between psychological (or psychographic) and demographic variables.

(iv) *Conflicting information*. When collecting information from multiple sources, relevant knowledge often contains contradictions that can range from mere differences in terminology to significant gulfs in theoretical underpinnings. For example, frequently there is a necessary reliance on expert opinion to augment the knowledge required for the construction and use of models. This dependence creates an environment conducive to the discovery of conflicting information, as demonstrated recently in the reported dispute between Bruce Hoffman and Marc Sageman about jihadist terrorism [14]. Sageman's book, Leaderless Jihad [13], characterizes jihadist terrorism as a primarily 'bottom up' phenomenon, where small local cells are formed by radicalized amateurs on their own initiative, without direction or support from a central or affiliated jihadist terrorist organization. Hoffman [6] disagrees with this position and argues that central organization exerts ideological influence and plays a significant leadership and operational role in respect to jihadist cells. This dispute emphasizes how a difference of opinion can have great repercussions for those who attempt to model particular behaviors. Sageman's ideas about how terrorism begins and is spread require a much different modeling approach than does Hoffman's. While it may be relatively straightforward to manipulate input variables in a model, differences in opinion about theoretical constructs, such as this one, have a much more substantial impact on modeling.

Information may not only be conflicting because of differing opinions, sometimes it is the case that apparent contradictions indicate an actual potential incongruity in the world. For example, an often-cited contributing factor to terrorist behavior is the lack of education among contested populations. Seemingly, an obvious intervention would be to provide means of education. Increasing education, however, may have a contradictory affect. According to Atran [2], increasing literacy rates may allow for a greater exposure to terrorist propaganda. Likewise, decreasing the poverty of an area, which seemingly would mitigate the financial incentive to become involved in insurgent behavior, might in reality mean a "redistribution of wealth that left those initially better off with fewer opportunities than before" [2]. Even more strongly, Horgan [8] points out that the attempt to weaken terrorism through financial manipulation is "practically worthless" from a counterterrorism point of view. This problem is compounded by the inclusion of cultural variables, as, when modeling other cultures, it is often difficult to avoid bias in applying cultural information. In counterinsurgency efforts, for example, there is a "faulty assumption: that insurgents are motivated by a desire for Western or "modern" democratic and capitalistic values" [15]. These differences in opinion and counterintuitive effects require that the

modeling of a particular strategy not be carried out in isolation, which might allow for the exclusion of potentially detrimental results.

(v) *Modeling groups as collections of individuals*. Terrorists interact with their environment and environmental variables influence terrorist behavior. Therefore, to understand the behavior of terrorists it is important to investigate it not as a collection of isolated events, but as manifestations of a comprehensive system that includes players at both an individual and group level. Not only do individual terrorists learn, but so do terrorist groups, where if they do not adapt as a group, they are less likely to be able to continue their existence amid a superior force [3]. Representing group knowledge in models, however, is more than just the aggregation of individual knowledge because a group "structures, stores, and influences what and how its members learn" [9]. This may be of specific importance when, for example, policy makers would like to influence those aiming to become terrorists, but cannot get at them directly. Instead, they may create programs that affect the radical institutions within sympathetic groups that may have an indirect effect on the beliefs of the terrorists. Modeling constructs must therefore consider theoretical concepts that incorporate group as well as individual behavior, which in turn adds to the amount of data and relevant research required for accurate modeling.

3 Approaches to Mitigating These Issues

In spite of the difficulties associated with modeling human behavior, there is still value in creating and using behavioral models, especially when the purpose of the model has been explicitly defined before development. Models have a variety of applications; therefore, it is important to make the intended use of model well-defined. Because human behavior is such an extraordinarily difficult subject to model, predicting exactly which individuals will become involved in terrorist activities is, at best, overly optimistic, and at worst, completely unrealistic. As recommended by Davis and Arquilla [4], it is more reasonable to identify influencing factors on an outcome rather than pinpointing what that outcome will be, therefore, a more realistic goal of modeling is to provide insight into potential outcomes. For example, a system dynamics model may investigate whether the interdiction of cross-border cash flow limits insurgents' ability to offer financial inducements to the population [10], though the exact effect is unpredictable. Models are also useful as a means of communication, where an explicit representation of the process through which an individual becomes a member of a terrorist organization can be used as a guide to furthering discussion about intervention.

Several of the previously mentioned issues can be mitigated through the use of a flexible modeling framework, one wherein perspectives from different subject matter experts or different psychological theories can be readily interchanged and explored. This also allows for iteration of the model development so that new models are not continuously created each time some limited data is obtained. Instead, it allows for feedback into evolving representations of the domain given the type of data that is available, while also realizing it may be only one perspective or snapshot of the domain being assessed. For example, a modeling framework may be constructed so as to allow the incorporation of exchangeable submodels, which represent a subset of the

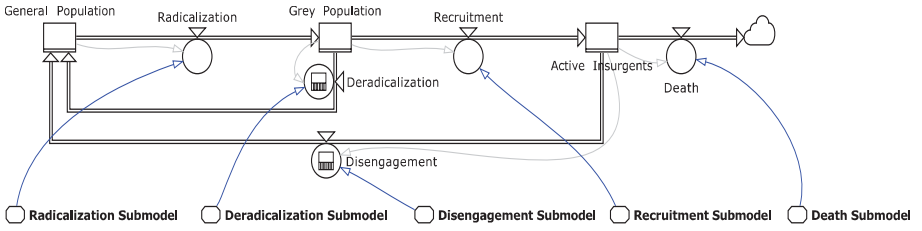


Fig. 1. A high-level system dynamics model of the insurgent recruitment process that accommodates various submodels

larger system. Figure 1 presents a high-level system dynamics model of the recruitment process that accommodates different submodels into the process, as shown along the bottom of the figure. These insertion points can contain a library of submodels that reflect conflicting viewpoints, differing theories, or models based on particular datasets. For example, the main model may be executed with a disengagement submodel that incorporates the practice of paying people to disengage from the activity. Likewise, it could be executed with a different disengagement submodel that reflects people who disengage because of education opportunities, or with one that incorporates a separate, highly detailed behavioral theory. This interchangeability exists for all of the submodels encompassed in the higher-level model and also allows for model revision as more information becomes available. For example, as data is gleaned from the success of an intervention method, that information can be codified by updating existing submodels or by developing another submodel, so that tradeoffs can be explored.

Model flexibility also facilitates the problems associated with understanding the relationship between individual and group behavior, which is critical to elucidating the process by which people are recruited into terrorist groups. Because understanding the ideology of a group allows only limited insight into its internal decision making processes [12], the manner in which individual preferences of the group-members combine into a group-decision and the manner in which individual preferences change because an individual is a member of a particular group are critical types of interactions. This complexity of modeling the influences may be mitigated through the use of multiple submodels, for example, those that capture influence processes at different levels, such as the strategic, organizational, and psychological levels outlined by McCormick [11]. Another option is to utilize various types of modeling frameworks within the submodels, as one particular type of model may not be adequate for the intended analysis. For example, it may be beneficial to use both a social network model and an agent-based model to capture the spectrum of influence on an individual subject to recruitment. Only by using a variety of models (or models that allow for a variety of theoretical constructs) may all possible broader trends be discovered [3].

Despite the inevitable problems that arise from applying psychological theories for use within an analytical tool, there are several options for improving implementation. One potential area of mitigation for models that incorporate theories would be the development of a well-defined formal approach for evaluating the degree to which individuals exhibit the traits or tendencies present in a theory that are to be included in

the model. For example, the well-known Hofstede dimensions [7] may be an adequate representation of culture for a particular population; however, it may be necessary to understand to what degree that cultural characterization is valid under this circumstance. An approach that formally evaluates the relevance of a theory with, for example, a degree of applicability, would be helpful in preventing a misinterpretation of its application. This type of evaluation may also highlight the need for additional studies that could be used to fill in the gaps between theory and available data.

Many of the previously mentioned issues, such as the limit to which past data can be generalized, may be remedied by focusing subject matter experts on identifying information that is likely to be critically different within the specific domain or culture on which the modeling is focused. This identification may be made easier with the construction of a set of metadata to structure both the (potentially limited) theoretical information and the (sparse) data used within the various psychological and cultural models. A consistent set of metadata helps limit the variability that exists when modelers are not explicit with their methods of coding quantitative variables from qualitative data and may help to identify dependencies relevant for sensitivity analysis to evaluate uncertainty and changes in data.

4 Conclusion

Given the challenge of modeling human behavior in general, it is no surprise that attempts at modeling specific behavior, such as terrorism, are not more straightforward. This difficulty, however, should not diminish the usefulness of such endeavors, as models may help analysts to understand core factors that contribute and lead to terrorist acts. Even a general understanding can facilitate planning in such a way as to minimize the causal forces that result in terrorist activities. The grand challenge to model builders is not only to construct realistic and robust models, but, in the near term, to embrace what is learned from iterative and ongoing modeling efforts, to incorporate submodels or components of models that can be swapped, to enable inclusion of psychological and political theories that may not be fully developed, and to inform modeling efforts by facilitating the implementation of processes that address the significant amount of uncertainty associated with behavioral modeling.

Acknowledgments. This work was supported in part by ONR contract N00014081-0481.

References

1. Alexander, K.: Terrorism and Global Insecurity: A Multidisciplinary Perspective. Linton Atlantic Press, Louisville (2009)
2. Atran, S.: Genesis of Suicide Terrorism. Science, 534–539 (2003)
3. Bale, J.: Jihadist Cells and "I.E.D." Capabilities in Europe: Assessing the Present and Future Threat to the West. Unpublished report (2009)
4. Davis, P., Arquilla, J.: Thinking About Opponent Behavior in Crisis and Conflict: A Generic Model for Group Analysis and Discussion. RAND, Santa Monica (1991)

5. Gompert, D.: Heads We Win-The cognitive side of counterinsurgency (COIN). RAND, Santa Monica (2007)
6. Hoffman, B.: The 'Cult of the Insurgent': Its Tactical and Strategic Implications. Australian Journal of International Affairs 61, 312–329 (2007)
7. Hofstede, G.: Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations, 2nd edn. Sage Publications, UK (2001)
8. Horgan, J.: The psychology of terrorism. Routledege Taylor & Francis Group, London (2007)
9. Jackson, B., Baker, J., Chalk, P., Cragin, K., Parachini, J., Trujillo, H.: Aptitude for Destruction: Organizational Learning in Terrorist Groups and Its Implications for Combating Terrorism v. 1. RAND, Santa Monica (2005)
10. Kilcullen, D.: Countering Global Insurgency. Journal of Strategic Studies 28(4), 597–617 (2005)
11. McCormick, G.: Terrorist Decision Making. Annual Review of Political Science 6(6), 473–507 (2003)
12. Ross, J.: A Model of the Psychological Causes of Oppositional Political Terrorism Peace and Conflict. Journal of Peace Psychology 2(2), 129–141 (1996)
13. Sageman, M.: Leaderless Jihad: Terror Networks in the Twenty-First Century. Pennsylvania Press, Philadelphia (2007)
14. Sciolino, E., Schmitt, E.: A Not Very Private Feud Over Terrorism. New York Times, June 8 (2008)
15. Snodgrass, T.: Is counterinsurgency a viable strategy for America? American Thinker, October 4 (2009)
16. Zacharias, G., Macmillan, J., Van Hemel, S.: National Academies Press, Washington (2008)

An Application of Epidemiological Modeling to Information Diffusion

Robert McCormack and William Salter

Aptima, Inc.
12 Gill Street Suite 1400
Woburn, MA 01801
{rmccormack,wsalter}@aptima.com
http://www.aptima.com

Abstract. Messages often spread within a population through unofficial - particularly web-based - media. Such ideas have been termed "memes." To impede the flow of terrorist messages and to promote counter messages within a population, intelligence analysts must understand how messages spread. We used statistical language processing technologies to operationalize "memes" as latent topics in electronic text and applied epidemiological techniques to describe and analyze patterns of message propagation. We developed our methods and applied them to English-language newspapers and blogs in the Arab world. We found that a relatively simple epidemiological model can reproduce some dynamics of observed empirical relationships.

Keywords: Information Diffusion, Epidemiological Modeling.

1 Introduction

As the ability to disseminate information to individuals and populations has become nearly instantaneous and ubiquitous, the impact of that information can have important consequences. More and more, messages spread via Internet and other electronic mechanisms, such as mobile phones, blogs, and social media websites-e.g., Facebook and Twitter. Ideas that spread - often rapidly - have been termed "memes" [4]. While most memes are benign, such as a joke, some are malignant and threatening. These could be instructions for creating improvised explosive devices or calls for terrorist action, such as with the attack on the Danish embassies in the wake of the printing of cartoons depicting Mohammed.

Some memes spread quickly, some slowly, and many not all. Some may experience periodic episodes of prevalence, such as fashion trends. The ability of a message to spread may be dependent on a number of factors including the individuals involved, the medium of communication, the social context, and the message itself. Intelligence and counter-intelligence agencies must be able both to understand the dynamics of message propagation and to predict the consequences of various actions on the information propagation mechanisms used.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 382–389, 2010.

In this paper, we describe a successful approach to understanding the dynamics of meme propagation within the context of heterogeneous populations and Internet media. Our underlying hypothesis is that the spread of memes is in many ways analogous to the spread of disease. Mathematical epidemiology provides a wide range of quantitative models proven to be both analytically rich and operationally useful. Many epidemiological models proceed from assumptions about the human behavioral and social dynamics which lead to the spread of disease, and we believe that similar phenomena drive the spread of memes. The degree to which diseases and memes are similar is open to debate, but at least this approach may provide better insights into the dynamics of information propagation, and, at best, may yield a set of computational techniques useful for modeling meme spread.

1.1 An Overview of Mathematical Epidemiology

The application of quantitative mathematical methods to biological systems has proven to be both theoretically and practically rewarding. From Malthus's observations that geometric growth of human populations exceeds the arithmetic growth of food supplies [13], through the Lotka-Volterra models of interacting predators and prey which helped explain the decrease in aquatic predators due to overfishing [14], and more recently the analysis of the human genome, computational models help explain and predict a wide variety of biological phenomena. In the first half of the 20th century, Kermack and McKendrick introduced quantitative models aimed at capturing epidemiological phenomena [12] to explain the rapid rise and subsequent decline in the number of infected patients often seen in epidemics. Their models formulated the notions of infection and recovery as non-linear processes. Even with simple, low-parameter models, they reproduced the large-scale dynamics of epidemics such as the Black Plague and cholera. In the latter half of the 20^{th} century, Anderson and May [2] refined epidemiological modeling in terms of closed mathematical systems and produced useful real-world results. Indeed, mathematical epidemiology has become one of the key drivers in setting public health policy.

Some basic parallels between the spread of disease and ideas are obvious: ideas "spread" by being "transmitted" from "carriers" to "uninfected" people; such "transmission" requires "exposure" – the person being "infected" must encounter the idea – some kinds of people may have "resistance" to some kinds of ideas, while others may be more "susceptible." We wanted to determine if such parallels were more than metaphorical - if mathematical methods derived from epidemiology could actually capture observed patterns in data.

2 Modeling and Analysis of Information Diffusion Across Websites

As a first step in modeling information spread, we obtained a suitable data set for study. After investigating various data sources, we selected two websites

to analyze initially: a news website and a blog. We obtained approximately 4500 articles from The Pakistan Times (www.pak-times.com) between May 2007 and May 2009 and 4400 articles from Baithak Blog (baithak.blogspot.com) between August 2005 and May 2009. News websites and blogs, in general, operate in different ways. News websites are (often original) sources of information, while blogs tend to repeat or react to information found on news websites. We chose these two data sources, in part, to test if our models could find this type of relationship: the publication of information and the subsequent repetition or reaction to that information across the websites.

2.1 Extracting Memes

To study memes, we had to define them pragmatically. Richard Dawkins introduced the concept in his 1976 book *The Selfish Gene* [4] as a principle to explain the spread of culture and ideas. Most work in memetics remains philosophical in nature, with little agreement about the definition, importance, or even existence of memes. This project, however, was not concerned with the ontological status of memes or their role in cultural continuity, but rather sought to define a unit of information in such a way that we can track its prevalence and spread over time and enable useful application and analysis of epidemiological models. We used "topics" derived from natural language processing techniques to operationalize memes. We discuss the methods and algorithms employed to extract these topics, but we note that the epidemiological analyses are conceptually independent of how memes were extracted. Indeed, in future research we plan to expand the implementation of memes to other definitions and extraction techniques. Indeed, different methods for defining and extracting memes may prove useful in different contexts.

A class of powerful natural language processing algorithms, referred to as latent topic methods, uses statistical and/or algebraic methods to extract the conceptual content of free text documents. A topic is typically represented mathematically as a weighted set or distribution of words and can be thought of as a significant conceptual unit that is found within the text. Imagine someone gives you an article about the death of Archduke Franz Ferdinand and asks you what the article is about. You may say it is about Austria-Hungary, Serbia, an assassination, World War I, and royalty. These are the types of topics that these latent methods attempt to extract. In the initial phase of this research, we focused on such topics. Such methods have been used, by Aptima and others, in a variety of domains for extracting useful concepts from text.

In order to extract topics from our data set we applied Aptima's Latent Variable Analysis (LaVA^{TM}) toolkit, which includes a number of mathematical and statistical language processing tools. In particular, we used an algorithm in LaVA called probabilistic latent semantic analysis (PLSA). PLSA [8] is a statistical latent variable method for finding words that tend to be associated across documents, and extracting such groupings without requiring that all the grouped words always co-occur. For example, if "bombing" often occurs with "terrorist," "insurgent," and "IED" in some documents, and "explosion" (but not

"bombing") occurs with those same three words in others, a latent variable method would correctly group "bombing" with "explosion," even though they might never co-occur.

We trained a 100-topic model on the data using PLSA. Throughout this paper, we use Topic 66 as an example, but note that the described analyses were performed for all 100 topics. One can often get a sense of the topic from the words with highest probability for the topic. For example, the top words for Topic 66 are, in order of descending probability, "Iran, Iranian, Nuclear, Missile, Israel, Iraq". They are concerned with Iran, Nuclear Weapons, and the Middle East so one might interpret this topic as having to do with Iran's relationship with other Middle Eastern countries and their involvement with nuclear weapons. Indeed, an examination of the full set of 100 topics (not presented here) provides a plausible conceptual overview of the types of information contained within the data.

2.2 Correlating Memes across Websites

To investigate how memes change over time, we determined the prevalence of each topic on each day for both sources by extracting the relevant topics from each article in the data source using the PLSA model. The set of topics for each day is then an aggregation of the topics found for that day's articles. These analyses yielded 100 "waveforms" or "time-series" for each source, where each wave form represents the prevalence of one topic over time in the data source.

To investigate meme"spread" between data sources, we performed a standard cross-correlation analysis on the waveforms to find examples in which changes in topic prevalence in one data source were correlated with similar changes in the other. Cross-correlation is a measure of similarity that can be calculated between two continuous or discrete signals, where one is fixed in time, and the other is allowed to vary as a function of time-lag. The two waveforms are compared at all different time-lag values, producing a distribution of similarity measurements as a function of time-lag. One can picture this as fixing one signal and sliding the other backwards and forwards in time to find the best match. These cross-correlations yield two potentially important pieces of epidemiological information: the maximum correlation of the signals and is the time-lag between sources that yields that maximum. These two parameters give us an initial understanding of the spread of information within the data set for that meme. As an example, Figure 1 shows the time-series signal of Topic 66 for the Pakistan Times on top and the corresponding topic time-series for the Baithak Blog below it. Prima facie, the two signals appear to show interesting dynamics. There is a large spike in activity around July-August 2007 in the Pakistan Times followed by a similar increase in the Baithak Blog. Near the end of 2008 and the beginning of 2009, we see what seem to be periodic dynamics in both signals. Thus, upon inspection, the data from the two sources appear to be related.

The maximum correlation is between about a 9- and 12-day lag, with a positive lag interpreted as the Baithak Blog lagging behind the Pakistan Times. Thus, when prevalence of Topic 66 increases in the Pakistan Times, there is a better

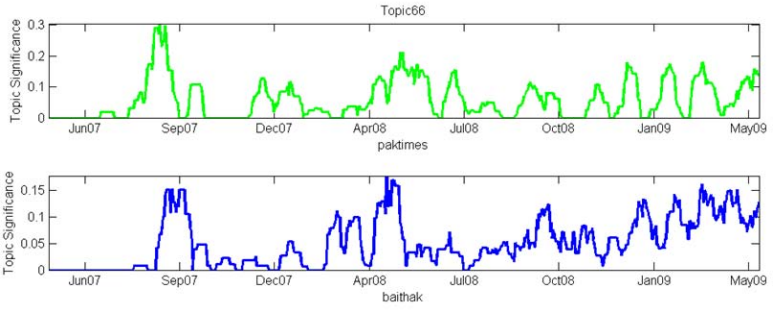


Fig. 1. Time-series graphs for Topic 66: Iran and Nuclear Weapons. The top graph represents the strength of Topic 66 over time for The Pakistan Times, the bottom for Baithak Blog.

than random expectation of a similar increase in the Baithak Blog about 10 days later. While there is a strong correlation between the two data sources on this topic, it is by no means perfect, nor does it speak to mechanisms of causation. Rather, it characterizes the empirical statistical expectation of behavior between the data sources. The cross-correlation analysis was applied to all 100 topics. Many topics showed a strong correlation, while others did not. Indeed, one would expect that some topics simply do not spread between the data sources.

2.3 Epidemiological Model

Our goal in developing an epidemiological model is to attempt to reproduce the kinds empirical relationships discovered and documented above. We constructed and parameterized epidemiological models for all 100 topics, using the maximum correlation and the time-lag which produces that maximum correlation for each, as well as the period of each topic using a simple sine wave fitting method.

The cross-correlation analysis showed empirical relationships between sources on certain topics. As a preliminary epidemic model, we attempt to capture that relationship using the three parameters. We are not making any claims about the nature (or, indeed, existence) of causal mechanisms. Whatever the mechanisms may be, our objective was to discover empirical relationships and to produce an epidemiological model to describe them. Given a topic time-series from one data source, the historical correlation and time-lag between two data sources on that topic, and the period of the topic in the second data source, we formulate an epidemic model to approximate the dynamics of the second topic time-series.

The specific model we used is based on an unrestricted random walk, a standard stochastic epidemiological modeling technique. This is a Markov process in which a random variable moves between a set of states (here, positive integers) based on certain probabilities. In a one-dimensional random walk, we are only concerned with the probabilities of the random variable increasing or decreasing. (This is the type of simple epidemiological model used to provide insight into how the number of infections changes over time.) We specify a particular

instance of a random walk model based on three parameters: the correlation, the time-lag, and the period of the topic, as well as the dynamics of the topic from the influencing source. The model is given as:

$$P\{I_{n+1} = i + 1 | I_n = i\} = \beta \widehat{I}_{n-d} \quad (1)$$

$$P\{I_{n+1} = i - 1 | I_n = i\} = \gamma i \quad (2)$$

$$P\{I_{n+1} = i | I_n = i\} = 1 - \beta \widehat{I}_{n-d} - \gamma i \quad (3)$$

where β is the maximum correlation based on the historical cross-correlation analysis, d is the time-lag of that maximum correlation, γ is the period estimated from the historical time-series signal, and $\widehat{I}$ is the time-series signal from the influencing source. This model defines the probabilities of an increase, a decrease, or no change in the state space at each time step. We assume that the time step of the model is small enough that only one change occurs in that interval. (1) gives the probability of an increase in the number of infections based on the current number of infections. In our case this is the probability of the topic occurring on the website in the next time step (tomorrow) based on the probability of the topic occurring in the current time step (today). It is defined as the maximum correlation times the probability that the topic occurred in the first website d days ago. This simple equation is intended to capture the observed time-lagged relationship between the data sources. The probability that the level of infection will decrease in the next time step is given by (2); we expect to see the probability of the topic appearing on the website decrease based on the period of that topic. Finally, the probability of no change is one minus the other probabilities (3).

Thus, this model estimates the probability of a topic increasing or decreasing on the website based on the value of that topic for the other website and the specific parameters capturing the temporal dynamics of the relationship between the websites. We simulate this model by choosing a random number from a uniform distribution at each time step. If it falls within the threshold of increase we add one to the current state, if it falls within the threshold of decrease we subtract one, otherwise we do nothing. In order to get a general understanding of the dynamics produced by the model we simulate 100 sample paths of the random walk and find the mean of those paths.

Now, we revisit Topic 66 in Figure 2 which displays the model approximation of the topic time-series on the bottom and the actual topic-time series on top. The correlation between the approximated and actual signals is 0.40. While not a perfect correlation, the approximation captures some of the interesting dynamics of the original signal. Indeed, we do not expect a perfect approximation, due to the numerous influences on the website that are not captured in the model. Furthermore, the actual observed empirical relationship is approximately 0.40 as well. Nevertheless, including only one related source and three parameters, the simple epidemiological model approximates the actual time-series surprisingly well. Similar epidemiological model analyses were performed for the remaining topics but are not presented here.

We believe that the work reported here shows the potential utility of applying epidemiological models to information spread. A number of more precise

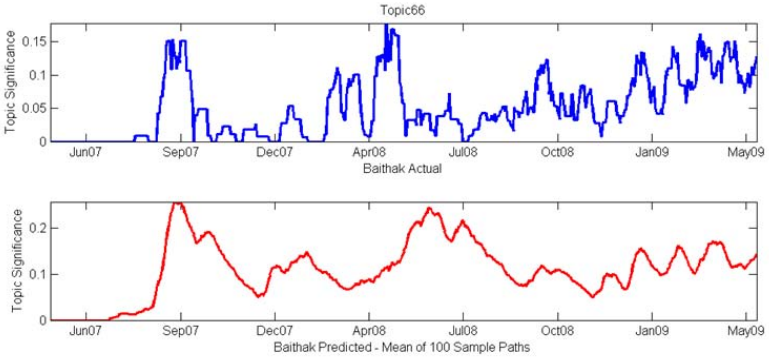


Fig. 2. Mean of 100 sample paths from the epidemiological model (bottom) compared with the actual time-series (top)

methods can be used to obtain data and extract topics, and more advanced epidemiological models can be applied. In addition, a number of theoretically interesting, and potentially practically useful, extensions can be made: organizing topics hierarchically, investigating ways in which memes may "mutate" into other memes, investigating memetic "markers" (such as particular phrases [10]).

3 Applications to Social Media

Ideas increasingly spread by "non-traditional" Internet media, particularly Twitter. When this work was originally planned (December 2007), Twitter barely existed. Now, it is widely used and can serve as a valuable source of information about real-time and near-real-time events of military and political relevance - tracking movements of terrorists and authorities during the Mumbai attacks, organizing rallies in Spain in response to the Madrid bombings, organizing rallies to protest the June 2009 election in Iran, monitoring the locations of troops sent out to suppress those protests, etc.

Twitter (and similar sites) differs in important respects from "old media" Internet sites like newspapers, blogs, chat rooms, and the like. The stringent limit on the number of characters of a tweet means that tweets tend to contain only one or two memes, a simplifying virtue in their analysis. And, critically, Twitter tends to be a real-time medium. Twitter also has a strong geographical component, supports tagging of tweets, and has an induced network structure among users, of followers and followings.

Twitter use is evolving rapidly - tweets can contain weblinks; the induced network structure can perhaps be exploited for social networking, marketing, other commercial and nefarious purposes; and new uses are constantly being developed. The larger world of social networking applications is evolving rapidly as well. However Twitter and social networking evolve, the "tweet stream" will constitute a valuable source of data. Its differences from more traditional web

media make it important to address and present technical opportunities to incorporate its real-time inputs into a rich analytical framework. For example: a temporally extended database of the tweet stream can be analyzed; memes extracted from the real-time flow of tweets can be fed into rich epidemiological models; the tweet stream can be monitored for indicators and warnings derived from epidemiological models, in addition to being monitored for emerging events; and the geographical aspects of tweets can be used to add a key dimension to the tracking of memes spread.

Acknowledgements

This paper reflects work performed on a Phase I SBIR project entitled E-MEME (Epidemiological Modeling of the Evolution of Messages), contract N00014-09-M-0189, sponsored by Dr. Rebecca Goolsby of ONR, whom we thank for her valuable insights, resonant questions, and overall support.

References

1. Allen, L.J.: An Introduction to Mathematical Biology. Prentice Hall, Englewood Cliffs (2007)
2. Anderson, R.M., May, R.M.: Infectious Diseases of Humans, Dynamics and Control. University Press, Oxford (1992)
3. BlogTrackers: Combining Domain Knowledge and Novel Search Capabilities, http://www.public.asu.edu/~huanliu/projects/BlogTrackers/
4. Dawkins, R.: The Selfish Gene. Oxford University Press, Oxford (1976)
5. Diekmann, O., Heesterbeek, J.A.P.: Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation. John Wiley & Son, Ltd., Chichester (2000)
6. Edelstein-Keshet, L.: Mathematical Models in Biology. McGraw-Hill, Boston (1988)
7. Griffiths, T.L., Steyvers, M., Tenenbaum, J.B.: Topics in semantic representation. Psychological Review 114(2), 211–244 (2007)
8. Hoffman, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACMSIGIR conference on research and development in information retrieval, Berkeley, California, USA, pp. 50–57 (1999)
9. JSpider, http://j-spider.sourceforge.net/
10. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proc. 15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (2009)
11. Intel Math Kernel Library, http://software.intel.com/en-us/intel-mkl/
12. Kermack, W.O., McKendrick, A.G.: A Contribution to the Mathematical Theory of Epidemics. Proc. Roy. Soc. Lond. 115A, 700–721 (1927)
13. Malthus, T.R.: An Essay on the Principle of Population, as it Affects the Future Improvement of Society, 6th edn., Murray, London, UK (1826)
14. Volterra, V.: Variations and fluctuations of the number of individuals in animal species living together. In: Animal Ecology. McGraw-Hill, New York (1931)

A Social Network Analysis Approach to Detecting Suspicious Online Financial Activities

Lei Tang, Geoffrey Barbier, Huan Liu, and Jianping Zhang*

Data Mining and Machine Learning Laboratory
Arizona State University
Tempe, AZ 85287-8809
{L.Tang,gbarbier,Huan.Liu}@asu.edu
* The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102
jzhang@mitre.org

Abstract. Social network analysis techniques can be applied to help detect financial crimes. We discuss the relationship between detecting financial crimes and the social web, and use select case studies to illustrate the potential for applying social network analysis techniques. With the increasing use of online financing services and online financial activities, it becomes more challenging to find suspicious activities among massive numbers of normal and legal activities.

Keywords: Social Networks, Social Network Analysis, Crime detection.

1 Introduction

Advances in information processing technologies make financial crimes easier to commit, more frequent, and more difficult to detect. Financial institutions are required to submit a Suspicious Activity Report (SAR) to the Financial Crimes Enforcement Network (FinCEN), an agency of the Department of Treasury, when observing transactions that are suspicious. A large number of SARs submitted since 1996 were attributed to Wire Transfer Fraud, which constitutes 45% of the 53,590 cases [5] in 2007 and 2008. Scalable and novel techniques are needed to prevent crimes from happening, or limit the impact of crimes. Many existent data sources, such as SARs and Currency Transaction Reports (CTRs), may be leveraged for detecting financial crimes. One way of using these data sources is to apply information extraction techniques to identify financial entities and relationships between individuals, organizations, and events. However, there are too many links to scrutinize effectively without advanced analysis methods. The application of social network analysis to discover a community of people involved in unusual money transactions may be useful in revealing criminal financial activities.

* The views presented in this article are those of the author, and do not necessarily reflect the views of MITRE.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 390–397, 2010.

SARs document many cases where applying social network analysis could help detect criminal activity and preventc rimes. Nine law enforcement cases for financial crimes were reported in the Oct. 2007 monthly SAR activity review [4] and more crimes are highlighted every month in subsequent editions. In one of the 2007 cases, the bankrupt defendant made over 1,000 deposits at several different bank locations into his accounts totaling more than $500,000. These "transactions took place at several different locations in amounts less than $10,000 and, in one instance, five transactions occurring in one business day. The defendant was making cash deposits totaling up to $30,000 monthly and writing numerous $2,000 checks to himself and family members."

Additional examples highlight the social structure in the network of criminal activities. A perpetrator acts as a sink with different bank accounts receiving high numbers of deposits from different locations (Figure 1) within a short time frame. A perpetrator works as a distributor (Figure 2) who writes checks to self and family members (presumably, a small number of people). Stores providing check cashing services to selected customers with normal activity including daily deposits of large volumes of third party checks and corresponding withdrawals. The networking pattern is the store is both the sink and the potential distributor (as it received large amounts of cash after depositing third-party checks.) The deposits are distributed in cash so it is not known who received the money.

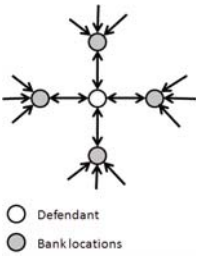


Fig. 1. Sink

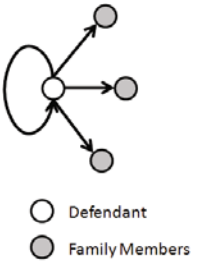


Fig. 2. Distributor

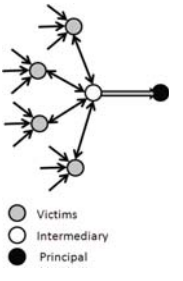


Fig. 3. Online Scam

Other cases include stealing from numerous accounts for relatively low amounts of money via computer intrusion, credit card fraud, debit card fraud, and wire transfer fraud. In one case, PayPal[1] noticed the irregular and distinctive patterns similar to those in illegal drug activity. In this case, small amounts of money from numerous accounts flow to those of the defendant's through the identity theft of hundreds of Americans and many bank frauds. From a network analysis perspective, there was an unusually large number of inlinks to the defendant's node. The defendant in another case was allegedly an intermediary in a criminal scheme (Figure 3). "Victims would send money to him, mostly through wire transfers and personal checks that he deposited in various accounts, and he would withdraw the money in structured transactions. With cash in hand, he would purchase money orders to send to the principals in the scheme."

[1] https://www.paypal.com/

Social network analysis techniques have been emerged to gain insights into criminal activity in both academia and industry [12]. In this paper, we highlight select financial crime cases reported in [4] and focus our discussion on how social network analysis techniques might be applied to help detect financial crimes. We also discuss a relationship between detecting financial crimes and the social web. In the experiments, we introduce some social network analysis techniques including detection of anomaly and overlapping communities, and report some interesting results achieved using Enron email data. We conclude with discussions of challenges and potential future work.

2 Social Web and Online Financial Activities

Online banking allows customers to conduct financial transactions on a secure website operated by their retail or virtual banks, credit unions or building society[2]. PayPal accounts are an example of such applications[3]. PayPal was founded in 1998 as a global online payment company and enables the most popular worldwide payment types in 190 countries and in more than 20 currencies for over 78 million users. The convenience and flexibility of online banking also offers a fertile environment for sophisticated and high-tech criminals to plan and commit financial crimes.

Emerging social web such as friendship and professional networks (MySpace, Facebook, Linked-In, Orkut, etc.) provide a new type of environment for surreptitious financial transactions at even larger scales than seen previously. A string of unabated Internet bank scams shows that hustlers have found the cyberworld to be just as profitable as the physical world. In 2008, criminals stole 30,000 pounds from victim they profiled through facebook [8]. The online banking scams can often operate for several months without triggering suspicion. Suspicious financial activities can be structured and covert to evade timely detection. With increasing use of online financing services and increasing online financial activities, the challenge of how to proactively find suspicious activities among massive numbers of normal and legal activities in the emerging environment of social webs and online banking is still not adequately addressed.

Social network analysis could be effective in combining with traditional investigative approaches. Social network analysis can: (1) exploit tips - confirm or reject the legitimacy of tips, and follow the lead to find potential associates in a tip-based triggering system[4]; (2) help understand various roles of an actor of interest such as a sink, intermediary, distributor, and the relationships between the actor and others for subject-based data mining [6]; and (3) can help find outliers or odd partners: most actors with similar statistics perform some routine tasks. SNA methods and techniques such as structural equivalence, blockmodels, and relational algebras [11] might be useful for investigating financial crimes.

[2] http://en.wikipedia.org/wiki/Online_banking/

[3] http://www.aotalliance.org/summit2007/2007_presents/501_ecommerce_bankingw.pdf

[4] Tips are still one prevalent useful means to pinpoint suspects.

3 Case Studies of Social Network Analysis

How can a social network analysis approach help detect suspicious transaction patterns? Specifically, we study how we can find potential interesting actors for further study, how we can focus on an actor to discover his/her role in the network, and how we can structurally interrogate the data to improve precision. In this section, we show some case studies using a SNA approach.

Access to most financial transaction data is limited. To illustrate the potential of social network analysis to positively impact the ability to detect on-line criminal financial activity, we provide a case study based on data from a criminal case involving significant financial fraud in the United States, the enron email corpus [7]. The Enron email corpus was released by the Federal Energy Regulatory Commission (FERC) in 2002 to enable the public to understand why FERC investigates Enron's financial problem. The data we used was obtained from University of Massachusetts[5], which contains 250,484 unique emails received by 149 people (mainly senior managers in Enron company). Here we investigate only the internal communication among the 149 people, reducing to 27,224 unique emails. Based on their email communications, we construct social networks between the 149 people. The social network is represented as a directed graph with edge weights being the email frequencies between two persons. We construct a network for each month from Nov 1998 to Jun 2002. The Enron corpus, the first publicly-available email communication network data, has stirred a surge of network analysis on large-scale networks [3,2,10,9]. In this work, we emphasize the discovery of anomaly and interesting patterns in the network.

3.1 Suspicious Activity Detection Based on Pattern Difference

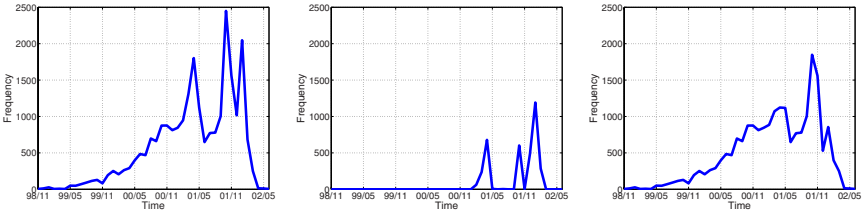
As introduced earlier, some financial crime cases involve high frequency of local interactions. For many social network analysis approaches, the first step is the construction of a network. However, as we show here some suspicious activities from few individuals might affect the pattern presented in the whole network. These suspicious activities can be detected by studying the pattern differences after removing certain individuals.

We propose that those actors whose removal results in a sharp pattern change might be considered as candidate for further investigation. In order to find them, we construct an activity vector $\mathbf{v}$ for each actor. Let $\mathbf{a}$ denote the total activity in a network (or a local network). The pattern difference of the removal of one actor i can be computed as

$$difference = distance(\mathbf{a}, \mathbf{a} - \mathbf{v}_i) \tag{1}$$

The number of emails sent by one particular user each month is the activity vector for the user. To compute the difference, we adopt one minus cosine similarity of two vectors for the distance function. Since actors are seldom active thorough the whole period (1998/11 - 2002/06). we consider only those months one actor is active when we do the computation.

[5] http://ciir.cs.umass.edu/%7Ecorrada/enron/



(a) Emails of all actors **(b)** Emails of Pete Davis **(c)** Emails excluding Pete

Fig. 4. Email traffic patterns presented in the network

The person whose removal leading to the biggest pattern change is *Pete Davis* based on our computation. Figure 4a summarizes the email traffic information between the 149 persons. Clearly, three spikes can be observed in Apr. 2001, Oct. 2001 and Jan. 2002. Figure 4b plots the emails sent by Pete Davis in each month. A 3-spike pattern is also observed. However, after removing Pete from the email communication, the pattern drastically changed as shown in Figure 4c. Only two spikes (Apr 2001 and Oct 2001) remain. Both Apr. 2001 and Jan. 2002 look normal now compared with its neighboring months. By checking the Enron chronology events[6], we noticed that a turning event happened during Oct. 2001 which explained the spike at that moment. In Oct. 2001, Enron announced the huge loss and the Securities and Exchange Commission (SEC) launched inquires into Enron finance. This is one example to show that local pattern change can affect the global phenomenon.

More interesting results can be found by checking the content of emails sent by Pete. we noticed that most of his emails were sent to the same group of people and followed the same pattern, like a real-time transaction reminder generated by machines every hour. Such a strong garbage email communication inside a small clique does shift the whole global trend. It is the pattern difference between networks before and after removing one person that leads us to locate Pete Davis and find out his "abnormal" email communication presented in this data set.

This case study also underscores the need to be cautious of spam and outliers when constructing a social network. Focusing just on the global traffic or statistics could mislead to some "fake" events or patterns as in Figure 4a. To really understand the cause, it is desirable to zoom into those small groups to discover some trend. With Pete Davis's data removed, we move on to investigate communities that may be manifest in the data.

3.2 Communication Pattern Discovery and Key Actors

In the introduction, we highlighted a few financial crimes involving a group of entities with frequent financial transaction, that is, more frequent than normal activities. Hence, finding meaningful communities can contribute to the discovery of financial crimes.

[6] `http://www.usatoday.com/money/industries/energy/2006-01-23-enron-chronology_x.htm`

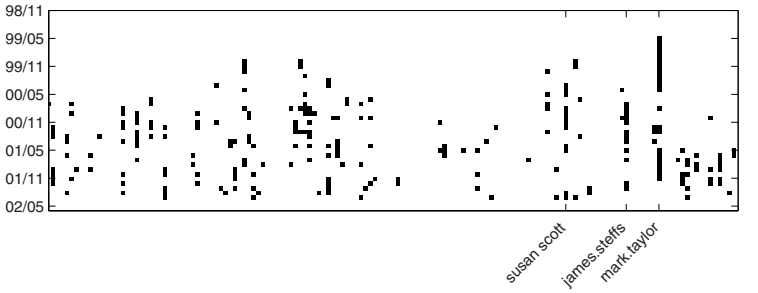


Fig. 5. Key Actors in Enron data - 44 months, 149 people

Discovery of communities in social networks is one of the most studied areas in social network analysis [11,1]. Many community detection methods often require users to specify the number of clusters, and the returned communities are disjoint. Alternatively, we develop a strategy to extract communities without requiring pre-specified number of clusters and allow communities to overlap. The algorithm for discovering overlapping-communities is below:

1) Use a depth-first search to find all maximal cliques.
2) Based on the overlap of maximal cliques, merge them into clusters. One applicable simple strategy is: if two maximal cliques have an overlap $> \epsilon$, merge them into one cluster. Those actors connecting two cliques are *key actors*.
Based on the resultant community and key actor information, we can easily detect network dynamics, and further check each group's sustainability to external influence or attacks.

We applied the algorithm to the Enron data. We only consider those connections with mutual communications between two actors. During the clique merge in step 2, ϵ is set to 1. That is, as long as two cliques overlap, we merge them into one community. Figure 5 shows the key actors in each month. Each row is a month, and each column represents a person (in total 149 people). Each black cell represents a key role. Some interesting patterns can be observed. Most people appear and disappear occasionally as key actors. This is the effect of randomness in a temporal change. More interesting actors are those who are consistently play a key role in some group. Three people who are key actors for at least 10 months are *Susan Scott*, *James Steffes* and *Mark Taylor*. They play a key role in reality as well. For instance, James Steffes is the vice president in Enron for government affairs. To our surprise, none of the key actors is high level executives such as CEO of Enron. For instance, the CEO of Enron, Jeff Skiling, never appeared as a key actor. Actually, most of his emails were sent to some people outside Enron or a broad range of receipts within Enron. Some emails were even sent by his assistant on behalf of him.

As we found key actors for each cluster, it is easy to identify those stable and vulnerable communities. In other words, if the number of key actors in a group is small, then the cluster is more like a star-typed community. So if the key actor is removed from this community, or some links between the key actor and

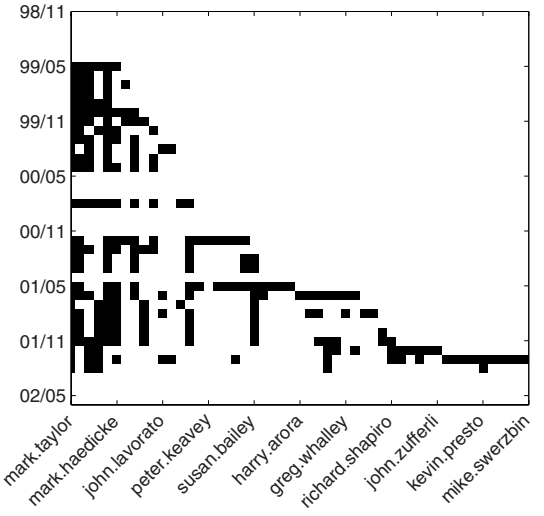


Fig. 6. People Assigned to the same community as Mark Taylor

other members are removed, this cluster is often decomposed into singletons and isolated tiny groups. If the number of key actors is large in comparison with the number of members in a community, then it is a clique-typed community. Hence, removing any key actor has a less observable effect on the community. This kind of community is more stable. Such kind of structure analysis is especially important if we need to break down financial criminal activity.

Once we find key actors in each small group, it is desirable to detect the cluster membership change as well. In other words, does he/she always talk to the same group of people? Figure 6 shows the members that are assigned to the same cluster as Mark Taylor (the first column represents mark.taylor) at different time points certain timestamps and Mark stays active from May 1999 until Feb 2002. Interestingly, a small group of people (showing in the first few columns): tana.jones, louise.kitchen, sara.shackleton stay very closely with Mark Taylor all the time. On the other hand, some people joined Mark Taylor's group at some later time, e.g., greg.whalley. Some only communicated with him occasionally or at some special event point (see kevin.presto in Jan. 2002 at the bottom). This kind of temporal change can help us identify the trend change of each group.

4 Conclusions and Future Work

As financial crimes continue to increase, SNA techniques may be a useful tool to guide investigators to suspects, and aid administrators in detecting unlawful activities. We introduced examples of suspicious financial activity cases and discussed their relationship with social networks. We apply social network analysis to a surrogate data set extracted from Enron email communications and show how some suspicious activity can be uncovered by comparing the pattern difference presented in a network. We also presented an algorithm to find overlapping

communities as well as key actors of each community. Based on the key actor and cluster membership evolution, it is possible to perform further analysis to find out suspicious activities.

This work highlights the potential of SNA to detect online suspicious financial transactions. Some interesting challenges and questions remain. For example, the very thin line between legal and illegal financial transactions makes the detection of criminal activity a non-trivial task. Additional research needs to be done to uncover answers to: What are the right metrics and triggers for distinguishing between normal activity and criminal activity? Besides communities, are there any other frequent financial crime communication patterns? And how to unravel them? It is also desirable to continue experiments with large-scale networks and different data sets to address further challenges. Since real-world transaction networks are of colossal size, future work needs to be done to determine how well SNA approaches will scale up to data sets with enormous financial transactions as well.

Acknowledgments. This work is, in part, supported by AFOSR.

References

1. Aggarwal, C., Wang, H. (eds.): Managing and Mining Graph Data. Springer, Heidelberg; Expected in early 2010
2. Chapanond, A., Krishnamoorthy, M.S., Yener, B.: Graph theoretic and spectral analysis of enron email data. Comput. Math. Organ. Theory 11(3), 265–281 (2005)
3. Diesner, J., Frantz, T.L., Carley, K.M.: Communication networks from the enron email corpus "it's always about the people. enron is no different". Comput. Math. Organ. Theory 11(3), 201–228 (2005)
4. FinCen. The SAR activity review: the trends, tips, and issues (October 2007), `http://www.fincen.gov/sarreviewissue12.pdf`
5. FinCen. The SAR activity review - by the numbers (June 2009)
6. Jonas, J., Harper, J.: Effective counterterrorism and the limited role of predictive data mining. Policy Analysis 11 (2006)
7. Klimat, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) ECML 2004. LNCS (LNAI), vol. 3201, pp. 217–226. Springer, Heidelberg (2004)
8. O'Hare, P.: Facebook boaster robbed of pounds 30,000; conmen use details off net. Daily Record, p. 23 (April 2008)
9. Priebe, C., Conroy, J., Marchette, D., Park, Y.: Scan statistics on enron graphs. Computational & Mathematical Organization Theory 11(3), 229–247 (2005)
10. Shetty, J., Adibi, J.: Discovering important nodes through graph entropy the case of enron email database. In: LinkKDD 2005: Proceedings of the 3rd international workshop on Link discovery, pp. 74–81. ACM, New York (2005)
11. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge (1994)
12. Xu, J., Chen, H.: Criminal network analysis and visualization. Commun. ACM 48(6), 100–107 (2005)

Opponent Classification in Poker

Muhammad Aurangzeb Ahmad and Mohamed Elidrisi

Department of Computer Science and Engineering, University of Minnesota
{mahmad,elidrisi}@cs.umn.edu

Abstract. Modeling games has a long history in the Artificial Intelligence community. Most of the games that have been considered solved in AI are perfect information games. Imperfect information games like Poker and Bridge represent a domain where there is a great deal of uncertainty involved and additional challenges with respect to modeling the behavior of the opponent etc. Techniques developed for playing imperfect games also have many real world applications like repeated online auctions, human computer interaction, opponent modeling for military applications etc. In this paper we explore different techniques for playing poker, the core of these techniques is opponent modeling via classifying the behavior of opponent according to classes provided by domain experts. We utilize windows of full observation in the game to classify the opponent. In Poker, the behavior of an opponent is classified into four standard poker-playing styles based on a subjective function.

Keywords: Opponent Classification, Opponent Modeling, Poker.

1 Introduction

A game is a structured or a semi-structured interaction between two or more entities where there usually some incentives involved in playing the game. Games can be classified into games with perfect information and games with imperfect information. In games with perfect knowledge the complete state of the game is observable. Examples of such games include Tic-Tac-Toe, Backgammon, Go, Chess etc. The most common approach in solving these games is to build a game tree and determine the best strategy to use given one's position in the game tree. Such games are readily amenable to brute force search approaches. A history of the study of this class of games demonstrates that many of these games were solved as the computational resources became available.

Games with imperfect information on the other hand are fundamentally different because the complete state of the game is not observable. In many such games either the search space is intractable or even in cases of tractability many of the techniques employed in games with perfect information do not do well in games with imperfect information. In theory even if one could state the optimal strategy for imperfect information games like Poker, computing the strategy would be prohibitively expensive and given the large uncertainties involved may not always work. Thus alternative techniques to traditional approaches used for perfect information games have to be employed. Poker is a game with imperfect information often used as a test bed for

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 398–405, 2010.

research in AI [6]. There are many variants of the game of Poker, the one that has come to prominence in recent years is known as Texas Hold 'em. In this paper use this variant of Poker for opponent modeling and opponent classification. In Texas Hold 'em there are can be multiple opponents but we limit our study to the cases where there is only one opponent. We explore the game for the limit version of the game where there is a limit on the amount of money for each round of bets.

One way to model opponents is to try to model how humans play such games since humans play these games in manner which is significantly different from how computers play these games e.g., example in chess the best chess playing program computes billions of possible moves in a second, the best chess playing humans only consider a few moves at every turn. In this paper we want to classify agents based on insights about how humans play imperfect games. In addition to modeling how humans play the game of Poker the results from agent modeling can have implications for the wider AI research community as well e.g., insights from opponent modeling in Poker may be transferrable to other domains as well [6].

2 Paper Preparation

Research into perfect information games began not long after the invention of computers. Checkers was one the first games to be analyzed with computers with the earliest program being written in 1952 [19]. Most of the techniques used to tackle per-fect information games involve representing the moves as a search tree and thus the task of determining the optimal policy to play in a game amounts to searching the game tree. Some games like Tic-Tac-Toe have a relatively small game tree, even more complex versions of this game like the 4 x 4 x 4 version have been solved [2]. Exam-ples of games that have been solved or where a computer can play better than any human player include Othello [16], Go-Moku [3], Connect-Four [1] and even Checkers [17]. There are some games like Go which are perfect information games but for which traditional approaches based on search do not seem to fare well [15].

Work in the computational study of games with imperfect information has largely been concentrated in modeling the game of Bridge [21]. While Bridge has not been solved, steady progress has been made in the field [9]. While Scrabble can also be considered to be a game with imperfect information, it has been shown that [20] it is amenable to brute force approaches and an optimal strategy can be computed given the uncertainty involved in not knowing the opponent's letters and the unseen letters. In games like Poker the factor of uncertainty cannot be ignored however. Research into analysis of poker has a long history in mathematics, economics and psychology, but given the complexity of analyzing the complete game of poker simplified variants of poker were studied [12]. Game Theoretic approaches to Poker have also been applied but some researchers acknowledge that the computational complexity of game may be too great to be amenable to a game theoretic solution [14]. The dramatic increase in interest in Poker in the Computer Science community has been facilitated by access to cheap and fast computational resources. Billings et al. [7] note that a poker playing agent should have the following characteristics: Assessing hand strength and potential, betting strategy, bluffing, unpredictability and opponent modeling. Billings [8] describe an agent for playing Poker that computes probabilities for folding,

raising, betting and then uses simulations to compute expected values of betting decisions. Aaron Davidson [11] describes various strategies for opponent modeling in Poker in his master's thesis which in turn is expanded upon by Terence Schauenberg [18] in her master's thesis.

3 The Modeling Approach

Our approach departs away from the traditional approaches to solving games which mainly involve constructing a game tree and searching through the tree to find an optimum solution. Humans do not tend to build a complete game tree for the game nor do they keep the complete game history of the game states. They rather look at relatively short window of the history of the game and the expected states to classify the behavior of the opponent to some general behavioral categorization. In the case of Poker behavioral psychologists have identified [13] four types of behavioral categorizations: loose passive, loose aggressive, tight passive, and tight aggressive. A loose agent is an agent that tends to take actions that are not correlated with the strength of their hand while a tight in an agent who is completely bound by the strength of the hand. An aggressive agent is one that acts strongly by raising the bets while passive involves taking the average action which is call and sometimes folding. Our strategy can be described as follows: Given a hand we agent first compute the odds for the cards, we then define thresholds that can be used by an agent to play Poker, use deterministic rules, try probabilistic variants of these rules, use different levels of past history to guide action and finally use non-correlation.

3.1 Formulas

The first step in opponent behavior classification is computing the odds of a hand. In our approach we pre-compute the strength of all possible hands in a game of poker. We do this by first generating all possible hands (52*51*50*49*48=311,875,200) and then playing each hand against every other possible hand. Now we have to compute the results of playing each of the more than three hundred million hands against all other hands which is an $O(n^2)$ operation and implies than more 97 quadrillion comparisons have to be made which is prohibitively expensive. However one can significantly decrease the number of comparison by noting that many of the hands generated are permutations of one another and thus they do not have to be recomputed. To reduce the number of such comparisons we force an ordering on the cards which takes into account the transitivity of the hands. The number of comparisons that have to be done can be computed by using the following formula.

$$\rho_h = \prod_{i}^{52} |c_i| \prod_{j=i+1}^{52} |c_j| \prod_{k=j+1}^{52} |c_k| \prod_{m=k+1}^{52} |c_m| \prod_{n=m+1}^{52} |c_n|$$

Where ρ_h is the number of times that comparisons have to be made and c is a unique card. Thus only 2.6 million hands have to be evaluated. Once we have generated the hands the same formula can be used for computing the number of comparisons that

we have to make in order to compute the probabilities for the hands. Here the probability of a hand is defined as the number of times a given hand can beat all the other hands divided by all other possible hands. This can be computed based on the following formula.

$$P_{hi} = \frac{\sum_{j \neq T, j \neq i} w_{i,j}}{|T| - 1}$$

Where P_{hi} is the probability of the hand, T is the total number of possible hands and w_{ij} is an indicator function whose value is one if the hand i can defeat hand j, otherwise its value is zero. The end result of this process is a look up table which has all possible 2.6 million ordered hands and their associated probabilities.

3.2 Poker Agent Models

First we describe our most basic agent which uses a deterministic model for game play. When a new game starts the agent computes the probability of winning by using the lookup table and then uses a threshold function to make a decision to either fold, call or raise. The values of the threshold are varied with increments of 0.05 until the optimal results were obtained, e.g., if the probability is less than or equal to 0.05 then always fold, if the probability is greater than 0.05 but less than 0.55 then always call and if the probability is greater than 0.55 then always raise. A shortcoming of the deterministic model is that it always behaves in the same manner if the state of the game is the same regardless of the previous history of the game. This makes it vulnerable to being exploited by another poker agent which can learn its behavior. To address this problem we created a stochastic model where the idea is to use thresholds but within the threshold decisions we introduce controlled randomization. Consider the case when the probability of hands is 0.89, if the deterministic agent is playing the game then it will always raise. If the opponent takes into account the history of game play with the deterministic agent then it will fold or pull out of the game. The stochastic agent avoids this problem by varying its behavior e.g., it generates a random number and if it is less than 0.5 then it calls otherwise it raises given that the probability of the hands is 0.89. Both these models and do not take into account the history of gameplay and cannot exploit the behavior of the adversary if it is known of have a certain style of playing. This problem can be mitigated by considering the history of gameplay. Based on this observation we describe a model which uses a set of rules for decision making. These rules override the expected decisions for whatever model that the agent is using. On the other hand if there are no rules that match the history of gameplay between the agent and its opponent then the agent uses either the deterministic or the probabilistic model. We use two different versions of the action based history models which use different rules and different length of history for decision making.

Taking an action that is non-correlated with the strength of hand and possibility of winning is a widely used in poker to increase winnings. In Poker this called Bluffing and Reverse Bluffing. Both of these methods assure maximum utilization of strength of hand or weakness. Bluffing is defined as taking an action like raise or call when the hand strength implies that one should fold. Taking what is called as strong action

when your hand is weak typically makes the opponents believe that one has a strong hand and then fold accordingly. Bluffing and reverse bluffing are instances of deception, so in this approach we try to capture the deception of opponents. Deceiving actions are in fact part of the optimum best policy. For instance, all Nash equilibrium calculation approaches have deceiving actions. There is however a semantic issue here because bluffing and deception in general implies some degree of intention. However one can never know for certain the intention of the opponent and so use the term "non-correlation of actions" to describe this phenomenon. This information can be extracted by looking at the game history in cases where showdown is reached.

There are 3 main steps in our non correlation approach. We start by assuming that the agents are playing poker based on a distribution similar that we get from our hand evaluator. Next we capture the opponent's hand at every showdown. Afterwards we build a categorical classification of the non-correlation of the opponent's actions to their hand using our own deterministic subjective function. We compare the non-correlation of our simulation results of actions with the actual opponent's observed actions. According to the comparison, we update our behavioral views of the opponent which are loose passive, loose aggressive, tight passive, tight aggressive. We keep track of four set of score records for the opponent, one for each behavioral class. Once the opponent takes an action that implies a certain class, we increment the score for that specific class. For instance, if the opponent has taken raise action even though they had bad hand then we increment the loose aggressive. The same if they have taken a call action when they had good hand then we increment the tight passive. For any opponent we typically have different values for each class. Typically the value with the highest score is the overall behavioral policy.

In this model, All-call player would be easily detected and classified as loose passive, and all-raise would be classified as loose aggressive. The non-correlation approach does not give us what our next action should be but instead it focuses on classifying the opponent and predicting their playing style. From that class label we can adopt the best counter strategy action for this specific behavioral style which is what many of the human players do. They adopt the strategy that they believe best counters the opponent's current style. If the opponent's behavioral style changes during the game, this will cause the score of the new behavior to increase and therefore adopt a proper policy to counter the newly adopted style.

4 Experiments and Results

In the previous section we described four main models that can be used to play Poker. For experimental evaluation of how these models fare against other models and against one another we had each individual agent play 1,000 games against all the other agents and including the agent itself for the same set of cards. From the four basic models we derived six models which were based on either the individual models or a combination of the models. The results for the six agents playing against one another are given in Table 2. Here we provide analysis of results for the individual agents. **Borg1** is a deterministic poker player part of the University of Minnesota poker research group family of bots. It mainly computes the odds of its card being

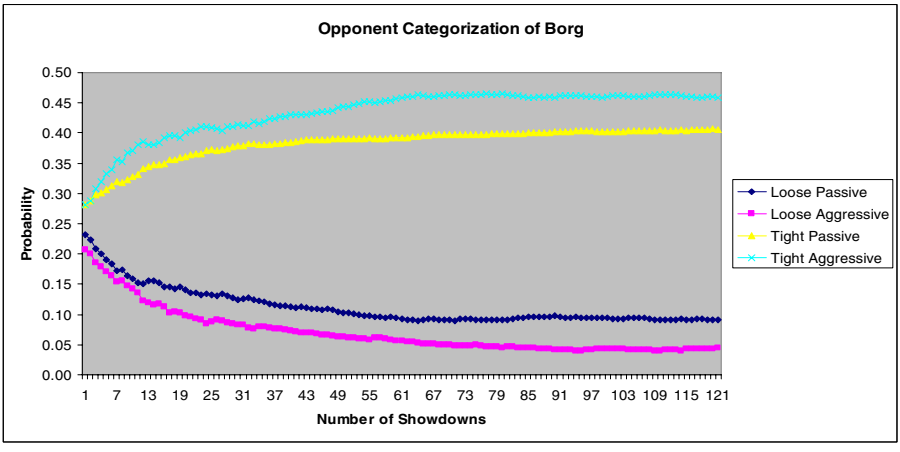


Fig. 1. Categorizing the behavior of Borg1 playing History II

better than the opponent's cards given that both cards are drawn from a uniform distribution. After computing the probability of the opponent's cards being better, it hands the results to the triple decision generator which makes the decision about whether to fold, call, or raise. Overall borg1 could be categorized as a tight aggressive player that sometimes plays loose. Figure 1 shows the behavioral categorization of borg1 which was mainly classified by our model as Tight aggressive. Borg1 typically has been making most of its wins from around 10 % of the games only. Whenever, Borg1 has a high probability hand it tends to raise and continue aggressively to show down.

We have four variations of the bot as detailed in the approach above. The deterministic Model, the stochastic model, and two variations the History based Model. History based one with a window of 3 and History based II with a window of 7. All four variations played against the following 5 bots. Random, All Raise, All Call, All Fold, and Borg1. According to our categorization All Raise is loose aggressive. All Call is loose passive, All Fold is also loose passive and Borg is tight Aggressive. It is clearly evident that deterministic is the best agent against random. It makes sense since it plays the same strategy against random based on its own hand and the actions of random do not follow any particular pattern. Thus it makes sense to play using one's cards. In the game play against All Raise the best strategy is deterministic.

The Stochastic agent actually performs worse than other agents and loses against All Raise and All Call because the learning component is not present. The reason for this result is that there is a small probability that the stochastic agent folds even in the later stages of the game. When it is playing against All Raise then the losses are likely to be high since it always raises regardless of its cards and the game has to go to completion either by playing the whole game or when the other opponent folds. Thus in the cases when the opponent folds its losses are quite high. The same reasoning applies when playing against All Call but the pot of money under consideration will be lesser as the other agent will never be raising. The results for All Fold are quite

Table 1. The result of different agent gameplays

Agent Type	Deterministic	Stochastic	History I	History II
Random	1402	266	693	1080
All Raise	1570	-623	983	715
All Call	394	-185	312	303
All Fold	708	708	744	744
Borg	-1024	-901	-790	-476

similar as expected but both the versions which employ Action Based History seem to have a slight edge over other agents. Lastly we note that although the different versions of the agents do not beat borg1 but each progressive version of the agent fares better than the previous version. The most significant improvement is for the second action based history agent whose performance is dramatically better than the other agents.

5 Conclusion and Future Work

Modeling opponent behavior in imperfect games in an open problem which potentially many applications in other domains as well. In this paper we explored a series of techniques for building Poker playing agents. We take our inspiration from how humans play Poker. In imperfect information games precise modeling of the opponent's behavior is rather complex, evaluating the accuracy of the model is in even more complex. However, classification of the opponent's behavior to some preconceived class structure is more tangible and easier evaluated. It eventually correlates to human behavior in similar imperfect information environments. One way to improve the results is to use genetic algorithms to find the values for the thresholds that maximize the performance. Future work will involve extending this model so that it can give a proper counter strategy for each specific type behavioral style.

References

1. Allen, J.D.: A Note on the Computer Solution of Connect-Four. Heuristic Programming in Artificial Intelligence 1, [7], 134–135 (1989)
2. Allis, L.V., Schoo, P.N.A.: Qubic Solved Again. Heuristic Programming in Artificial Intelligence 3, [9], 192–204 (1992)
3. Allis, L.V., van den Herik, H.J., Huntjens, M.P.H.: Go-Moku Solved by New Search Techniques. In: Proceedings of the 1993 AAAI Fall Symposium on Games: Planning and Learning. AAAI Press Technical Report FS93-02, Menlo Park (1993)
4. Baker, R.J.S., Cowling, P.I.: Bayesian Opponent Modeling in a Simple Poker Environment. In: IEEE Symposium on Computational Intelligence and Games (CIG 2007), Honolulu, USA (2007)
5. Baker, R.J.S., Cowling, P.I., Randall, T., Jiang, P.: Can Opponent Models Aid Poker Player Evolution? In: IEEE Symposium on Computational Intelligence and Games, pp. 23–30 (2008)

6. Billings, D., Papp, D., Schaeffer, J., Szafron, D.: Poker as a testbed for machine intelligence research. In: Mercer, R., Neufeld, E. (eds.) Advances in Artificial Intelligence, pp. 1–15. Springer, Heidelberg (1997)
7. Billings, D., Papp, D., Schaeffer, J., Szafron, D.: Poker as a Testbed for Machine Intelligence Research. In: Mercer, R., Neufeld, E. (eds.) AI 1998 Advances in Artificial Intelligence, pp. 1–15. Springer, Heidelberg (1998)
8. Billings, D., Peña, L., Schaeffer, J., Szafron, D.: Proceedings of AAAI 1999 (Sixteenth National Conference of the American Association for Artificial Intelligence) (1999)
9. Blair, J.R.S., Mutchler, D., Lin, C.: Games with Imperfect Information. In: Proceedings of the AAAI Fall Symposium on Games: Planning and Learning, pp. 59–67. AAAI Press Technical Report FS93-02, Menlo Park (1993)
10. Chaddock, G., Pickett, M., Armstrong, T., Oates, T.: Models of strategic deficiency and poker. In: Working Notes of the AAAI Workshop on Plan, Activity, and Intent Recognition (PAIR), pp. 31–36 (2007)
11. Davidson, A.: Opponent Modeling in Poker: Learning and Acting in a Hostile and Uncertain Environment, University of Alberta M.Sc. thesis (2002)
12. Findler, N.: Studies in machine cognition using the game of poker. Communications of the ACM 20(4), 230–245 (1977)
13. Griffiths, M.D.: The cognitive psychology of gambling. Journal of Gambling Studies (1990)
14. Koller, D., Pfeffer, A.: Representations and solutions for game-theoretic problems. Artificial Intelligence 94(1-2), 167–215 (1997)
15. Popma, R., Allis, L.V.: Life and Death Refined. Heuristic Programming in Artificial Intelligence 3, 157–164 (1992)
16. Rosenbloom, P.S.: A World-Championship-Level Othello Program. Computer Games 2, [6], 365–408 (1979)
17. Schaeffer, J., Burch, N., Bjornsson, Y., Kishimoto, A., Muller, M., Lake, R., Lu, P., Sutphen, S.: Checkers is Solved. Science (2007)
18. Schauenberg, T.: Opponent Modelling and Search in Poker. M.Sc. thesis (2006)
19. Strachey, C.S.: Logical or Nonmathematical Programs. In: Procd. of the ACM Conf., Toronto (1952)
20. Ulgee, I.H.: Letters beyond Numbers. Heuristic Programming in Artificial Intelligence 3, [9], 63–66 (1992)
21. Wheen, R.: Brute Force Programming for Solving Double Dummy Bridge Problems. Heuristic Programming in Artificial Intelligence 1, [7], 88–94 (1989)

Convergence of Influential Bloggers for Topic Discovery in the Blogosphere

Shamanth Kumar, Reza Zafarani, Mohammad Ali Abbasi, Geoffrey Barbier, and Huan Liu

Computer Science and Engineering
Arizona State University
Tempe, AZ 85281
{shamanth.kumar,reza,ali2,geoffrey.barbier,huan.liu}@asu.edu

Abstract. In this paper, we propose a novel approach to automatically detect "hot" or important topics of discussion in the blogosphere. The proposed approach is based on analyzing the activity of influential bloggers to determine specific points in time when there is a convergence amongst the influential bloggers in terms of their topic of discussion. The tool BlogTrackers, is used to identify influential bloggers and the Normalized Google Distance is used to define the similarity amongst the topics of discussion of influential bloggers. The key advantage of the proposed approach is its ability to automatically detect events which are important in the blogger community.

1 Introduction

For thousands of years human beings have been interested in understanding and measuring their surroundings. The Greek mathematician Thales used a method known as triangulation to measure the height of objects. A general concept of triangulation involves using two or more perspectives to accurately locate something else. The approach of using multiple perspectives to gain a more accurate measure or understanding of the details of an object, triangulation, is applied in other domains as well. Surveyors plan roads, geologist study earthquakes, and even biologist use radio triangulation to study wildlife behavior. Similar to how others use triangulation in their fields of work, we believe it is possible to detect significant events and produce a more accurate sentiment of the blogosphere by examining how much independent influential bloggers have in common with each other during a finite time period. We consider a significant event, a news topic, opinion, or other trigger that motivates the influential bloggers to write about some subject during the same period of time. The subject is significant in that it changes the focus to the subject from the previous and usually disparate topics that the independent influential bloggers were focussed on before the significant event.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 406–412, 2010.

2 Motivation

Social Media has played an increasingly significant role in our everyday life. We use Facebook to connect to our friends, YouTube to share videos, etc. The blogosphere is one of the most popular facets of Social Media. It has shown consistent exponential growth over the past few years. State of the Blogosphere[1], a study of bloggers performed annually by Technorati, presents some interesting results on the importance of the blogosphere. In the latest study it was observed that 76% of the bloggers who were surveyed, blogged to express their opinions. Additionally, the blogosphere captures more detailed aspects of public opinion than polling and surveys over a wider geographical area without traditional limits of time, space, and survey administrators. Opinions are associated with topics, which generally correspond to a real world event like the launch of a new mobile phone or the elections in a country. Topic discovery is therefore, an essential first step in the analysis of the blogosphere. An effective way to detect topics significant from the blogging community's perspective is to look at the discussions amongst influential bloggers in a community. The influential bloggers can be considered to be the first to start conversation on a key topic and hence provide a way of detecting what could be termed as the "hot" topics amongst the bloggers.

As the Internet continues to play a major role in communication among people from various communities, organizations, and nations, there is a growing interest in analyzing world wide web content to better understand the culture, sentiment, and social relationships amongst the people that use the web and provide a variety of social information. Computational methods can aid researchers in mining the vast amount of social computing data that is available today and help pave the way for a deeper understanding of how human actions in the "cyber-world" correlate with the "real-world."

3 Related Work

Some of the leading blog indexing services have blog search and topic tracking features. BlogPulse[2] supports conversation tracking, which looks at blog-roll data to track replies to a blog post, which when put together form a conversation. Technorati[3] supports advanced blog search capabilities including searching for other blogs that link to a specific blog post. Technorati also provides a list of key topics, videos, and people being discussed in the current blog posts. BlogScope [3] is an analysis and visualization tool for the blogosphere, which currently indexes 133 million blogs. This web based tool is capable of generating popularity trends for specific topics, identifying information bursts, and performing geographical search. Our tool, BlogTrackers [1], combines unique blog and blogger

[1] http://technorati.com/blogging/article/day-1-who-are-the-bloggers1/
[2] http://www.blogpulse.com
[3] http://technorati.com

analysis capabilities by identifying topics of discussion, influential bloggers in the community, and unique topic tracking features.

Influential bloggers can be identified by analyzing the link structure of the blogs as suggested in PageRank [9] and HITS [7] which can identify important nodes in a community. Authors in [11] propose the InfluenceRank algorithm which combines link-structure information and the novelty of the blog content to detect opinion leaders in a community.

Topic discovery and tracking has been a popular research area in information retrieval. In [2,10], the authors discuss the use of vector spaces in detecting topics. Tools such as [6,5] focus on the temporal analysis of topics to provide context to identified topics.

4 Methodology

An important issue in the analysis of the blogosphere is the detection of topics. In this section, we propose a new approach to detect "hot" topics by using the information from the blog posts of the influential bloggers. An influential blogger can be defined as a blogger whose opinions can influence the opinions of a significant number of other individuals in the community. This makes influential bloggers an important part of the blogosphere who can be targeted for advertising or for gauging the political sentiments of the community. Influential bloggers have their niche topics on which they are able to exert influence on others and we assume that they do not talk about similar topics unless a real world event of high significance such as elections or a new product launch occurs and motivates them to start talking about these topics. By focusing our attention on the topics discussed by the influential bloggers we hope to have a more efficient approach in detecting "hot" topics in the community.

Each influential blogger in the community can be represented as a vector of high frequency terms extracted from his blog posts for a specific period of time. Therefore, every blogger I_i can be defined as a vector of high frequency terms t_j as,

$$I_i = \{t_j | 0 \leq j \leq n\} \tag{1}$$

This vector of high frequency terms can be used to identify the convergence of the influential bloggers by measuring the distance between the average term frequency vector and that of a particular blogger. Due to the inability to use systems like WordNet to identify the semantic relationship between Indonesian terms, we used the Normalized Google Distance(NGD) [4] to measure semantic relationship between the terms of two vectors. The NGD distance measure calculates the normed semantic distance between two terms using the number of Google search results for the terms jointly and independently as a measure of their semantic relationship. The Normalized Google Distance for two terms x and y can be defined as,

$$NGD(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logM - min\{logf(x), logf(y)\}} \tag{2}$$

where f(x) and f(y) are the number of search results returned by Google for the search query x and y respectively, f(x,y) denotes the number of search results returned by Google for the search query containing both x and y, and M denotes the number of pages searched by Google, which is estimated to be around 10^{10}. If two words a and b have a distance smaller than the distance between a and c, then the two words a and b are said to be semantically closer to each other than a and c.

The similarity of influential bloggers, in terms of their topics of discussion can be defined as follows,

$$S_i = P_i \times B_i, \forall i \tag{3}$$

where P_i denotes the number of blogposts published by all the influential bloggers within a given time interval i and B_i denotes the number of influential bloggers during the time interval i. S gives us an estimate of the amount of activity, which if significantly greater than the neighboring time periods indicates a key or a significant event being discussed by the influential bloggers. P is used to scale the value of B to remove irrelevant points from consideration. The number of relevant influential bloggers during any given period B can be defined as

$$B_i = \Sigma b_i, b_i < T \tag{4}$$

where T is a suitable threshold which can be varied to increase or decrease the number of relevant influential bloggers B. If both P and B values are high then this is an indication that there is a significant proportion of blog posts from a significant proportion of influential bloggers, the topics from which together converge towards the average topic of discussion amongst all the influential bloggers.

5 Case Study

In order to explore the convergence of topics of influential bloggers, we identified the top 10 influential bloggers from a dataset comprising of 50 Indonesian

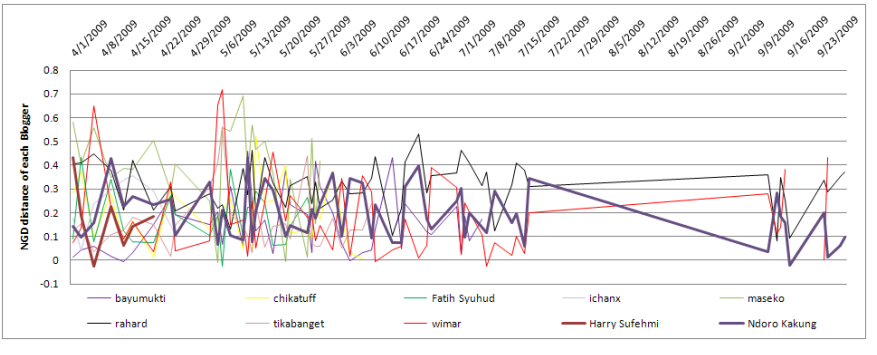


Fig. 1. NGD distance between the top 10 influential bloggers and the average terms during the period April, 2009 to August, 2009

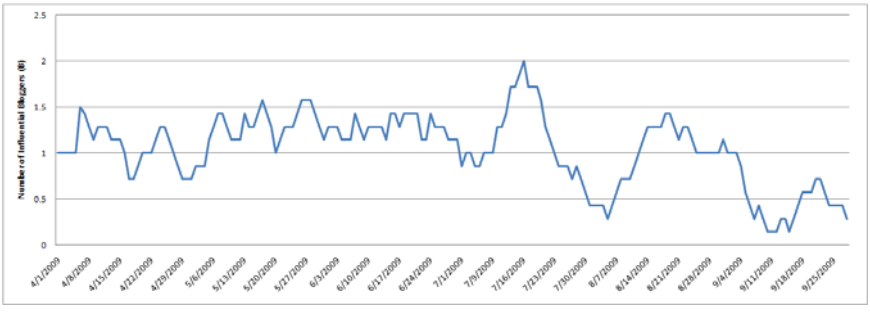


Fig. 2. Number of influential bloggers (B) whose deviation from average below the threshold T

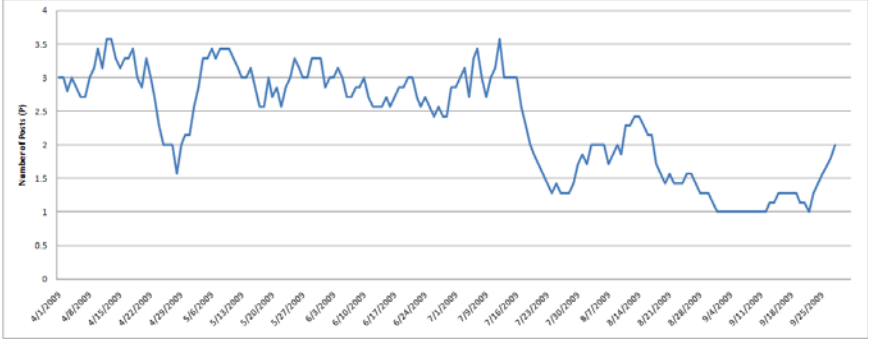


Fig. 3. Total number of blog posts (P) during each time interval

blogs [8] crawled using BlogTrackers for the period April 2009 to September 2009. To represent the topics discussed by influential bloggers, we computed the top 10 keywords and their frequency, which formed the term-frequency vector for each blogger, from the blog posts published by them during all the time periods for the aforementioned period. In this case study we fixed the time interval to consecutive periods of 7 days each. To compute the average term frequency vector for all the influential bloggers presented in Figure 1 we combined the already identified top keywords for individual bloggers for each period and then chose the top keywords from this list as the representative or average keywords for the duration. Using NGD, described in Section 4 we identified the deviation of each blogger from the average term vector by computing the average of the pairwise NGD distance between the terms in the two given vectors. These values for each blogger is presented in the Figure 1.

For the purpose of this case study we fixed the value of the threshold T to 0.2. We computed the number of relevant influential bloggers for each time interval using the definition provided in Eq 4, which is presented in the Figure 2. The value of B as shown the figure should be as large as possible for the convergence

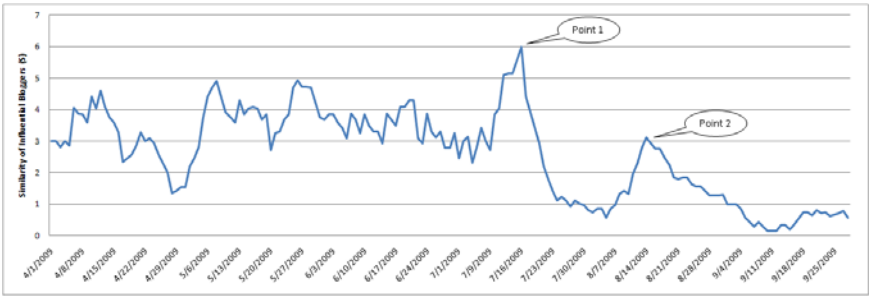


Fig. 4. Similarity (S) of influential bloggers for the period April, 2009 to September, 2009

of the influential bloggers to represent the overall sentiment of the influential blogger community. To compute the weights P for each time interval i we determined the average number of blog posts published during each time interval by all the influential bloggers and these values are shown in Figure 3.

We present the similarity of influential bloggers S computed using the definition from Eq 3 in Figure 4. We hypothesize that the peaks in the figure represent points of convergence of a majority of influential bloggers to the average topic of discussion amongst all the influential bloggers under consideration. The keywords corresponding to these points would most likely represent "hot" topics pertinent from the blogging community's perspective. After analyzing two of the most recent points as highlighted in Figure 4 we found that Point 1 represented keywords such as "president", "sby", "elections", and "political". All of these keywords correspond to the Indonesian presidential elections held on July 8th. It is clear that most of the influential bloggers considered the event to be significant enough to deviate from their normal activity and blog about this topic. On the other hand, when we analyzed Point 2 we found that the high frequency keywords during the period were "father" and "pin" which had no correlation with each other and which could not be associated with any specific event. This is in contrast to our earlier observations and highlights the need for further refinement in our approach to handle such points.

6 Conclusion and Future Work

In this paper we proposed a new approach to detect key topics of discussion in the blogosphere using the information from the activity of influential bloggers. Our case study on the influential bloggers from the Indonesian blogosphere shows that it may be possible to detect "hot" topics in the blogosphere by observing the convergence of topics from the blog posts of influential bloggers. The preliminary results obtained from our study highlight the need for additional work to validate our methodology and approach. The detection of a suitable threshold which can be used to automatically extract the points of convergence of influential bloggers is another relevant area for future research.

Acknowledgements

This work was supported in part by the Office of Naval Research under the grant ONR N00014-09-1-0165.

References

1. Agarwal, N., Kumar, S., Liu, H., Woodward, M.: Blogtrackers: A tool for sociologists to track and analyze blogosphere. In: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, ICWSM (2009)
2. Allan, J.: Topic detection and tracking: event-based information organization. Springer, Heidelberg (2002)
3. Bansal, N., Koudas, N.: Blogscope: a system for online analysis of high volume text streams. In: 33rd International Conference on Very Large Data Bases (2007)
4. Cilibrasi, R.L., Vitanyi, P.M.B.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19, 370–383 (2007)
5. Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: WWW 2004: Proceedings of the 13th international conference on World Wide Web, pp. 482–490. ACM, New York (2004)
6. Havre, S., Hetzler, B., Nowell, L.: Themeriver: visualizing theme changes over time. In: IEEE Symposium on Information Visualization, InfoVis 2000, pp. 115–123. IEEE Computer Society, Los Alamitos (2000)
7. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: 9th ACM-SIAM Symposium on Discrete Algorithms (1998)
8. Kumar, S., Agarwal, N., Lim, M., Liu, H.: Mapping socio-cultural dynamics in Indonesian blogosphere. In: 3rd International Conference on Computational Cultural Dynamics (2009)
9. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project (1998)
10. Schultz, J.M., Liberman, M.Y.: Towards a "universal dictionary" for multi-language ir applications. Topic Detection and Tracking: Event-based Information Organization (2002)
11. Song, X., Chi, Y., Hino, K., Tseng, B.: Identifying opinion leaders in the blogosphere. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 971–974. ACM, New York (2007)

Sentiment Propagation in Social Networks: A Case Study in LiveJournal

Reza Zafarani, William D. Cole, and Huan Liu

Computer Science and Engineering
Arizona State University
Tempe, AZ 85281-8809
{Reza,WCole,Huan.Liu}@asu.edu

Abstract. Social networking websites have facilitated a new style of communication through blogs, instant messaging, and various other techniques. Through collaboration, millions of users participate in millions of discussions every day. However, it is still difficult to determine the extent to which such discussions affect the emotions of the participants. We surmise that emotionally-oriented discussions may affect a given user's general emotional bent and be reflected in other discussions he or she may initiate or participate in. It is in this way that emotion (or sentiment) may propagate through a network. In this paper, we analyze sentiment propagation in social networks, review the importance and challenges of such a study, and provide methodologies for measuring this kind of propagation. A case study has been conducted on a large dataset gathered from the LiveJournal social network. Experimental results are promising in revealing some aspects of the sentiment propagation taking place in social networks.

1 Introduction

Social networks have become popular with the pervasive use of the World Wide Web. With the paradigm shift in the usage of the Web from information consumption to information production and sharing ("Web 2.0"), numerous social media services have emerged. Individuals use different social media services for various purposes and exhibit diverse behaviors. We use Flickr to share pictures with friends, Twitter to update our "status", MySpace to keep in touch with friends, and Blogs to express our interests, opinions, and thoughts. According to recent statistics[1], more than 10 billion photos exist on Facebook, 20 hours of video is uploaded on YouTube every minute, and around 38,400 photos are uploaded every hour on Flickr. With the massive amount of data published every day on these networks, we no longer have a shortage of experimental data; our challenge now is to make sense of the data. That said, the quantity of data gives us the opportunity to analyze the various behaviors of users in social networks and how they differ from their "real-world" social lives. Analogs of some

[1] http://www.labnol.org/internet/data-storage-for-user-generated-content/9656/

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, pp. 413–420, 2010.

real-world behaviors have been studied in the context of online social networks. For example, in [1], the authors introduce a technique to measure the degree of influence users in the Blogosphere have on other users in order to identify the most influential bloggers. In this study we focus on a different problem. Does the amount of [emotional] content users are exposed to on a daily basis in the online social world influence their emotions? And if it does, how can we observe this phenomenon? We aim to develop methodologies and find answers to these questions. In this paper, we present a case study with the following contributions:

- Formally define and study the propagation of sentiment in social networks,
- Quantify and predict the occurrence of a sentiment propagation, and
- Identify salient features that result in a sentiment propagation.

The rest of the paper is organized as follows: Section 2 describes the motivation behind this study. Section 3 presents the problem statement. Section 4 discusses a case study in LiveJournal, the approach used to analyze sentiment propagation in social networks, and experimental results. Section 5 summarizes the related research. Section 6 concludes with future work.

2 Motivations

The following five items describe the basic motivation behind our study.

- **How do individuals influence each other in social networks?** There is growing interest in the community to determine the extent to which participants can influence each other in terms of thoughts and behaviors via social networks.
- **Does sentiment propagate?** There has been extensive research on information diffusion in social networks. However, to the best of our knowledge this is the first study of its kind to consider sentiments as information entities to analyze the propagation thereof.
- **How does sentiment propagate?** Assuming there is a propagation effect, there are many questions that arise in this area. For example, how rapidly does sentiment propagate? What parameters influence the propagation rate? How do propagation speed variations correspond to real world events?
- **What different roles do individuals play in propagation?** It is important to analyze the actors involved in the propagation to understand these roles. For example, the users who initiate the propagation, those who relay the propagation (hubs), and those who block or enhance the propagation, are of interest in our a study.
- **How useful are sentiment analysis tools for sentiment propagation analysis?** It is interesting to find out how effective current sentiment analysis techniques are for analyzing sentiment propagation. For example, we used Normalized Google Distance (NGD) as measure of semantic distance in this study, and it is interesting to analyze how it affected our experimental results.

3 Problem Statement

Before we delve into the details, we will first formally define the problem of sentiment propagation in social networks. Let μ represent an active individual in cyberspace and s a single site. Without loss of generality, we restrict our study to a single website. We denote the set of all active users at site s as Λ_s. Let $m(\mu, t)$ denote the overall sentiment (mood) of user μ at time t. For a set of users $U \subset \Lambda_s$, we define the overall sentiment (mood) at time t as follows,

$$m(U,t) = \frac{\sum_{\mu \in U} m(\mu,t)}{|U|}. \tag{1}$$

Then, the sentiment propagation problem can be formally stated as follows:

Definition. ***Sentiment Propagation in a Social Network:*** *given a user μ, called the target user, a social network site s, and a subset of users $U \subset \Lambda_s$ that initiate the propagation at time t_i, a sentiment propagation has influenced μ at time t_j, iff.,*

$$|m(U,t_i) - m(\mu,t_j)| \leq |m(\Lambda_s,t_i) - m(\mu,t_j)| + b_1, \tag{2}$$

$$|m(U,t_i) - m(\mu,t_j)| \leq |m(U,t_i) - m(\mu,t_i)| + b_2, \tag{3}$$

where $t_i \leq t_j$ and $b_1, b_2 \geq 0$ are the intercepts.

Equation 2 denotes that if a propagation influences a user, then at time t_j, the overall mood of a user should be closer to the overall mood of the group that initiated the propagation than to that of the entire population of users. Equation 3, however, states that as time goes by (time t_j), the user should get even closer to the group that initiated the population in terms of mood. If both conditions are met, then we consider the user to be influenced by the propagation. Note that, b_1 and b_2 are parameters that can be learned or heuristically set.

4 A Case Study in LiveJournal

In order to analyze sentiment propagation, we performed a set of experiments on a dataset gathered in July 2009 from LiveJournal[2]. LiveJournal (LJ) is a web community where users can keep a blog or journal. Besides being quite popular, it includes features of a self-contained community and some social networking features similar to those of other social networking sites. An interesting feature of LiveJournal is that when users post on their blog they have the option of assigning a "mood" to their post. Users can select from a list of 130 moods or type freeform text to specify a mood not in the predefined list. Note that assigning moods to posts is optional and therefore not all posts have moods associated with them. We assume the mood assigned to a post is a sentiment that is generally oriented to the post content. This feature defines the data labels for our experiments.

[2] http://www.livejournal.com

Table 1. LiveJournal dataset statistics

Bloggers	Links	Link Density	Posts	Avg. Posts	Avg. Links	Net Diameter
16,444	131,846	9.8×10^{-4}	475,932	28.94	16	8

4.1 Data Collection

In order to gather a sufficiently large dataset for our experiments, we developed a breadth-first crawler that follows forward links (friendship relationship) on LiveJournal. We used forward links only since following them yields a large portion (or the "Weakly Connected Component") of the network [2]. The crawler starts from a well-known user found manually on LiveJournal, and reaches all users that are within four links of this user. We stored posts for every user found during the crawl. For each post, we stored the post date and the "mood" associated with it, if any. Overall, more than 1 GB of data was gathered. An overview of the dataset crawled is provided in Table 1.

4.2 Data Pre-processing

To ensure the data was sufficiently reliable for our experiments, various preprocessing steps were taken. First, we removed all non-English blog posts. We also removed all posts that did not have a mood associated with them (*non-moody* posts). We also filtered out posts with moods that were relatively infrequent in the corpus[3]. We then removed users who had less than five posts, had no *moody* posts (posts with moods), or had less than five friends. This was done to ensure we had a sample of user behavior sufficient enough to assess the sentimental dynamics which are the focus of this study. After these steps, the remaining posts in our dataset were associated with 285 distinct mood strings. We quantified the mood strings using Normalized Google Distance (NGD) [3]. Normalized Google Distance is a measure used for calculating the semantic distance between words from the hit counts returned by the Google search engine. Terms with similar meanings tend to be "close" in units of NGD, while dissimilar words tend to be farther apart. NGD between words x and y is defined as follows,

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}, \tag{4}$$

where M is the number pages on Google and $f(x)$, $f(y)$, and $f(x,y)$ are the number of hit counts for queries x, y, and $x + y$ on Google, respectively. In order to quantify moods, we first assumed that each mood can be represented using a pair $< positive, negative >$, where the first component is the positiveness and the second term is the negativeness. This approach has been extensively studied in opinion mining literature [4]. In order to obtain these values, a pair of *seed words* (poles) is required. Each seed word represents a pole (positive/negative). The NGD distance between these seed words and each mood

[3] In our experiment, the frequency threshold was empirically found and set to 10.

string provides us with the $< positive, negative >$ pair. Various seed words have been tested in the previous literature [5]. From these, we manually selected 2 pairs: $< excellent, poor >$ and $< happy, sad >$. This manual procedure involved looking at the NGD values produced for different moods using different seed words and ensuring that ostensibly negative (or positive) moods generated a greater negativeness (or positiveness) value. Then, all the mood strings were replaced with two pairs $< distance\ from\ happy, distance\ from\ sad >$, $< distance\ from\ excellent, distance\ from\ poor >$. After this step, the dataset was ready for experimentation.

4.3 Experiment Setup

Without loss of generality, we added a set of assumptions to the formalized version of the problem. First, we assumed that friends have greater influence over a user than non-friends. In other words, for a given user, the set of users that may initiate or transfer a sentiment propagation are his/her set of friends. Note that friends are directly connected to the user via forward links, i.e., are one hop away. The assumption simplifies the model but is still general enough to carry our experiments. Second, we assumed that the intercepts for our propagation constraints are zero, i.e., $b_1 = b_2 = 0$ (see Equations 2 and 3). Furthermore, we assumed that we have a propagation time window during which the sentiment propagation starts and ends. So, instead of arbitrary t_i's and t_j's (see Equations 2 and 3), we assumed that,

$$t_j - t_i \in \{1, 2, 3, 4, 5, 6\} \quad months. \tag{5}$$

For each time period, we analyzed the dataset and checked if the conditions for a propagation are met for each user. We recorded the status of the propagation (i.e., a binary value with zero meaning the sentiment propagation did take place) in that period, along with additional data for each user: the number of posts made by the user in that period, the number of posts made by his/her friends in that period, the user's number of friends, the time window, and the total number of posts in that period. Again, this is a labeled dataset where *propagation status* is the class variable and all the other gathered attributes are its preliminary features that can be employed for classification. Next, we review our experimental results on this dataset, which revealed some aspects of the sentiment propagation taking place in social networks.

4.4 Evaluation Results

We first performed a propagation classification using the aforementioned dataset and 10-fold cross-validation to see how accurately we could predict sentiment propagation. Tables 2 and 3 show the classification results from C4.5, Sequential Minimal Optimization (SMO) [6], Naive Bayes (NB), Logistic Regression (LR), K^* (lazy learning), and Random Forest (RF), using a 10-fold cross validation and for both seed pairs used for the NGD metric. As shown in these tables, both decision tree algorithms (RF, C4.5) outperform other methods and

Table 2. Classification results for $< excellent, poor >$ seed words

	SMO	C4.5	NB	LR	RF	K*
Accuracy	62.65%	**71.10%**	61.22%	62.65%	**72.78%**	69.98%
Mean Absolute Error	0.37	**0.36**	0.45	0.46	**0.34**	0.38

Table 3. Classification results for $< happy, sad >$ seed words

	SMO	C4.5	NB	LR	RF	K*
Accuracy	69.51%	**76.16%**	67.95%	69.44%	**76.91%**	72.34%
Mean Absolute Error	0.30	**0.31**	0.40	0.42	**0.30**	0.35

result in reasonable accuracy in predicting the propagation based on available features. Moreover, it is evident that seed words play a role in the classification performance and should therefore be selected carefully.

Next, we analyzed the attributes of our dataset in order to find discriminatory ones. Table 4 shows the list of attributes and their average values for different class values, i.e., the binary variable that indicates whether or not a sentiment propagation exists. As shown in this table, users that exhibit sentiment propagation, on average, have more friends, make fewer posts, and have have less prolific friends than those of users who are not clearly subject to sentiment propagation. Moreover, propagation influences users in time periods shorter than 4 months. Note that the 4 month propagation window might be due to the nature of our dataset and it can be different in other datasets.

5 Related Work

Sentiment Analysis, also known as *Opinion Mining* or *Sentiment Extraction* is an emerging area of research as a subset of Text Mining. Sentiment Analysis is a newer research area at the crossroads of Text Mining and Computational Linguistics concerned not with the more commonly studied topical analysis of a document, but rather analyzing the overall polarity of opinions or sentiments expressed therein. It involves techniques to automatically analyze the sentiment, attitude, or opinion of textual documents on the World Wide Web, usually in terms of being positive, negative, or neutral. Generally speaking, it aims to determine the attitude of a speaker with respect to some topic. This attitude may indicate evaluational or affectational state (the emotional state of the author) or the intended emotional communication (the sentiment intended to be conveyed to the reader). Sentiment classification, as one of the active topics in this

Table 4. Average feature values for different classes

	# Friends	Time Window	Friend Posts	User Posts
Has Propagation	9.92	3.78	16.06	2.09
No Propagation	9.35	3.85	64.89	8.32

area, deals with categorizing sentiments. In most cases, if not all, the categories are limited to two (bipolar classification) [7]. These two categories represent the *positive* or *negative* sentiments. In other cases, these categories represent the objectivity/subjectivity of the textual excerpts [8,9]. A comprehensive list of computational measures regarding semantic relatedness for approximating the relative meaning of words/documents has been proposed in the literature. Some of these measures use lexical dictionaries such as WordNet, or SentiWordNet [10]. SentiWordNet provides 3 values for each sentiment (positive, negative, neutral). Other methods such as the Point-wise Mutual Information (PMI) [5] or the NGD [3] use search engines such as Google and Altavista to extract semantic similarity. Some early work in the study of information diffusion is [11], which introduced a model of collective behavior based on the concept of an aggregate threshold that must be overcome for individual behavior to spread to other actors. Another prominent model in this area, Independent Cascade Model (ICM), is alluded to in [12]. ICM models diffusion on a stochastic process whereby behavior spreads from one actor to another with a given probability. Some aspects of sentiment propagation has been analyzed before. For example, the intuition behind sentiment propagation in the 'real world' was validated in [13] through an analysis of responses to surveys given to a network of participants over a 20-year period. Moreover, Wu et al. devised a theory on the formation and spreading of opinions in a social network based explicitly on network structure, and make predictions about the pervasive influence of a minority of central actors [14]. Huberman et al. make the important observation that in computer social networks, significant influence occurs mostly between actors with a sufficiently close relationship [15]. This set of relationships is a subset of hard-linked relationships, i.e., 'friends', and is only identifiable by analysis of actual interactions between actors.

6 Conclusions and Future Work

In this paper, we studied the propagation of sentiments in social networks. We presented the challenges we encountered and developed approaches to tackle those challenges. We conducted carefully designed experiments on our dataset, gathered from LiveJournal, in order to analyze the sentiment propagation and to identify salient features that play a role in the propagation. Our preliminary experiments showed that users who have more friends, are less prolific, and who have friends who are less prolific exhibit greater sentiment propagation than users with dissimilar attributes. Also, propagation occurs within time periods of less than four months. In the future, we aim to expand our work by considering a user's overall sentiment as a time-dependent stochastic variable. This will address the shortcomings regarding the discretization of time in our current experiments. As mentioned earlier in our experiments, seed words play a role in the performance of detecting propagation. However, we did not attempt to find the optimum seed words in order to improve the accuracy of the classification methods. We leave this as another line of our future work, i.e., searching for an optimal seed set for improving accuracy.

Acknowledgements

This work is, in part, sponsored by AFOSR and ONR.

References

1. Agarwal, N., Liu, H., Tang, L., Yu, P.S.: Identifying the influential bloggers in a community. In: WSDM 2008: Proceedings of the international conference on Web search and web data mining, pp. 207–218. ACM, New York (2008)
2. Mislove, A., Marcon, M., Gummadi, K., Druschel, P., Bhattacharjee, B.: Measurement and analysis of online social networks. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, p. 42. ACM, New York (2007)
3. Cilibrasi, R., Vitanyi, P., Cwi, A.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
4. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, p. 354, Association for Computational Linguistics (2005)
5. Turney, P., et al.: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 417–424 (2002)
6. Platt, J.: Sequential minimal optimization: A fast algorithm for training support vector machines. In: Advances in Kernel Methods-Support Vector Learning, vol. 208 (1999)
7. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL 2002 conference on Empirical methods in natural language processing, vol. 10, pp. 79–86. Association for Computational Linguistics, Morristown (2002)
8. Wiebe, J., Riloff, E.: Creating subjective and objective sentence classifiers from unannotated texts. In: Gelbukh, A. (ed.) CICLing 2005. LNCS, vol. 3406, pp. 486–497. Springer, Heidelberg (2005)
9. Riloff, E., Wiebe, J.: Learning extraction patterns for subjective expressions. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2003), pp. 105–112 (2003)
10. Esuli, A., Sebastiani, F.: SentiWordNet: A publicly available lexical resource for opinion mining. In: Proceedings of LREC, Citeseer, vol. 6 (2006)
11. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology 83(6), 1420–1443 (1978)
12. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 137–146. ACM, New York (2003)
13. Fowler, J., Christakis, N.: Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. British Medical Journal 337(dec04 2), a2338 (2008)
14. Wu, F., Huberman, B., Adamic, L., Tyler, J.: Information flow in social groups. Physica A: Statistical Mechanics and its Applications 337(1-2), 327–335 (2004)
15. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2008)

Iranians and Their Pride: Modalities of Political Sovereignty

Mansoor Moaddel

Department of Sociology,
Eastern Michigan University,
2361 Sun Valley Drive,
Ann Arbor, Michigan 48108 USA

Abstract. In 2000, we asked a nationally representative sample of 2,532 Iranian adults "which of the following best describes you: I am an Iranian, above all; I am a Muslim, above all; I am an Arab, a Kurd, a Turk, a Baluch, etc., above all?" We also asked them how proud they are to be Iranian; (1) very proud, (2) proud, (3) not proud, and (4) not proud at all. In the 2005 survey of a nationally representative sample of 2,667 Iranian adults, we asked these questions again. The first question was intended to measure national identity and the second national pride. The results showed that between the two surveys the percent of Iranians who defined themselves as "Iranians, above all" went up significantly-from 35% in 2000 to 42% in 2005. Those who said that they were very proud to be Iranian, on the other hand, went down considerably-from 89% in 2000 to 64% in 2005. What is more, national identity and national pride displayed opposing relationships with the norms and values that were rigorously promoted by Iran's religious regime and these relationships grew stronger between 2000 and 2005. The feeling of national pride was positively linked to attitudes toward gender inequality, religiosity, and religious intolerance, but negatively to attitudes toward the West, while national identity had just the opposite relationships with these variables.

These findings have implications not only for theories of political sovereignty but for understanding and predicting the regime's behavior as well. The pertinent theories are predominantly concerned with national sovereignty, which is a special case of political sovereignty, where the sovereignty is territorial. Given that a shift in the identity of Iranians from Islam to the nation entails weakening support for the regime's culture, the sovereignty the regime claims is not national in a strict territorial sense. In fact, since its emergence, the behavior of the Islamic regime has not been dictated by concerns for the country's national interests but by the desire to defend and propagate the faith, which the regime has consistently interpreted in terms of an anti-Western clergy-centered Shi'i revolutionary discourse. It is thus plausible to argue that the nation and the religion form the bases for competing modalities of political sovereignty among Iranians. The regime, however, mobilizes support for its cultural values through the provocation of national pride.

We develop structural modeling in order to test the relation of national identity and national pride to various indicators of the regime culture.

S.-K. Chai, J.J. Salerno, and P.L. Mabry (Eds.): SBP 2010, LNCS 6007, p. 421, 2010.

Author Index