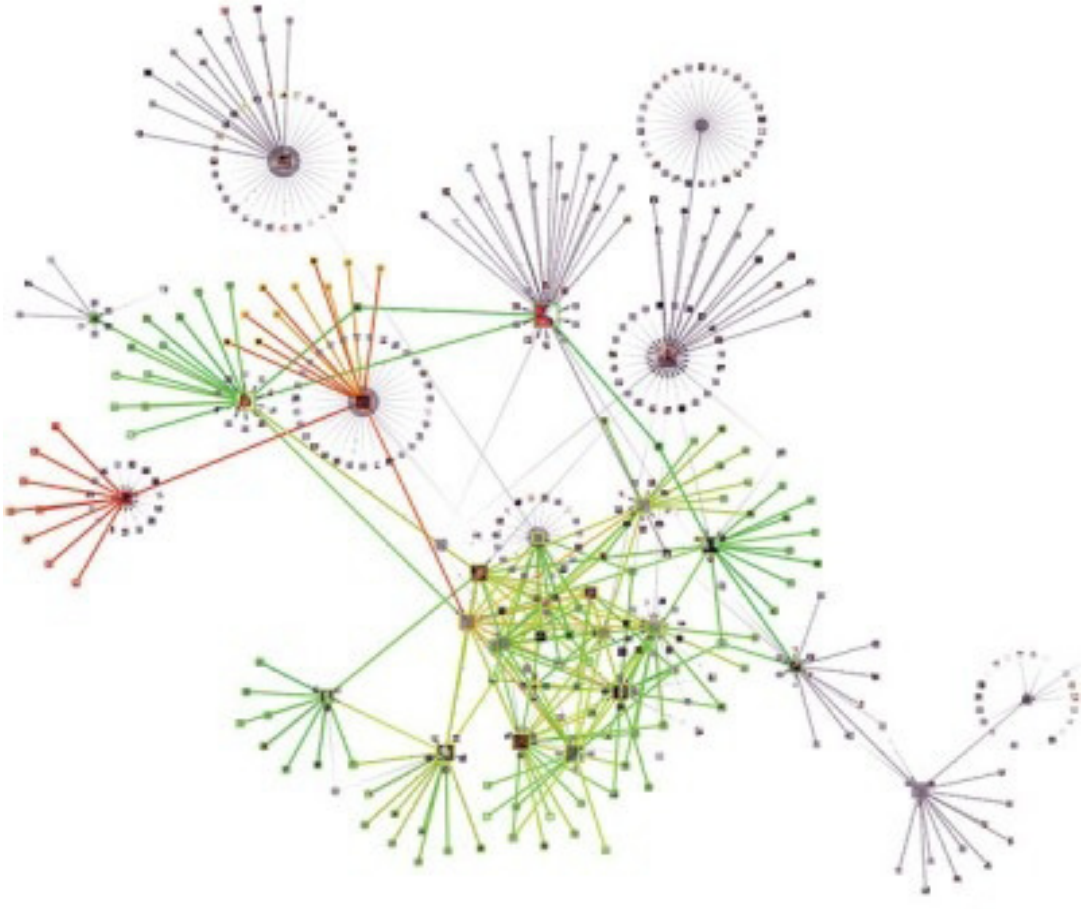
Laboratory for Information and Decision Systems

Interdisciplinary Workshop on Information and Decision in Social Networks

May 31 & June 1, 2011



Editors: Vincent Blondel, Munther Dahleh, Asu Ozdaglar, John Tsitsiklis

Sponsored by

NOKIA

orange™

This workshop is part of MIT's 150th anniversary celebration

MIT +150 INVENTIONAL WISDOM

Interdisciplinary Workshop on Information and Decision in Social Networks

May 31 & June 1, 2011

This workshop is being organized by the **Laboratory for Information and Decision Systems.**

Organizers:

Vincent Blondel, UCLouvain (Belgium) and LIDS, MIT
Munther Dahleh, LIDS, MIT
Asu Ozdaglar, LIDS, MIT
John Tsitsiklis, LIDS, MIT

Support Staff:

For administrative requests and general questions about the conference, please contact Lisa or Jennifer.

Lisa Gaumond – Administrative Coordinator – lisaga@mit.edu; 617-324-1543
Jennifer Donovan – Administrative Support – jdonovan@mit.edu; 617-253-2142

For problems with the website, or technical questions regarding the conference, please contact Brian.

Brian Jones – Technical Support – jonesb@mit.edu; 617-253-4070

Location:

The workshop will take place on the 6th floor of building E14 (the MIT Media Lab) on the MIT campus in Cambridge.

Travel and accomodation:

Workshop participants travelling to Cambridge should arrange their travel and accommodation; no support will be provided by the workshop organizers. There are several hotels close to MIT campus, including the historic Kendall Hotel and the Marriott Cambridge (for both hotels, please ask for the MIT rate). Prospective participants are advised to book early; the workshop week is commencement week at MIT and hotels in the area fill fast.

Plenary Speakers:

Prof. Daron Acemoglu, Department of Economics, MIT
Prof. Nicholas Christakis, Harvard Medical School, Harvard University
Prof. Martin Nowak, Program for Evolutionary Dynamics, Harvard University
Prof. Sandy Pentland, Human Dynamics Laboratory, Medialab, MIT


MAIN ST
BROADWAY
MASSACHUSETTS AVE
VASSAR ST
MEMORIAL DR
AMHERST ALLEY
AMHERST ST
MAIN ST
Kendall Square
Marriott/Residence Inn
Marriott Hotel
Whitehead Institute
University Park Hotel @ M.I.T.
77 Massachusetts Ave.
84 Massachusetts Ave.
Harvard Bridge
Hyatt
Hyatt Regency
E14
E14

Interdisciplinary Workshop on Information and Decision in Social Networks

May 31 & June 1, 2011

Scientific Committee

Chair: Vincent Blondel, UCLouvain (Belgium) and LIDS, MIT

Daron Acemoglu, Economics, MIT
Sinan Aral, New York University
Albert-László Barabási, Northeastern University
Alain Barrat, CNRS (France)
Michael Braun, Sloan School, MIT
Damon Centola, Sloan School, MIT
Nicholas Christakis, Harvard University
Vittoria Colizza, ISI (Italy)
Konstantinos Daskalakis, CSAIL, MIT
Eric Fleury, Ecole Normale Superieure (France)
Julien Hendrickx, UCLouvain (Belgium)
Marta C. González, Civil and Environmental Engineering, MIT
Sanjeev Goyal, University of Cambridge (UK)
Dirk Helbing, ETH Zürich (Switzerland)
Cesar Hidalgo, Media Lab, MIT and Harvard University
Bernardo Huberman, Social Computing Lab, HP Labs and Stanford University
Patrick Jaillet, LIDS, MIT
Tony Jebara, Columbia University
Michael Kearns, University of Pennsylvania
János Kertész, Budapest University of Technology (Hungary)
Renaud Lambiotte, Imperial College (UK)
Juha Laurila, Nokia Research (Switzerland)
David Lazer, Northeastern University and Harvard University
Naomi Leonard, Princeton University
Jure Leskovec, Stanford University
Jukka-Pekka Onnela, Medical School, Harvard University
Alex (Sandy) Pentland, Media Lab, MIT
Carlo Ratti, Senseable City Lab, MIT
Whitman Richards, CSAIL, MIT
Jari Saramäki, Aalto University (Finland)
Devavrat Shah, LIDS, MIT
Zbigniew Smoreda, Orange Labs (France)
Joshua Tenenbaum, Brain and Cognitive Sciences, MIT
Alessandro Vespignani, Indiana University
Eiko Yoneki, University of Cambridge (UK)

Program Outline

Tuesday May 31

8:30 Registration

9:00 Opening and general overview

9:30 Brief talks:
"Communities & Structure"

11:00 Break

11:30 Brief talks:
"Information Propagation"

12:30 Lunch

14:00 Invited plenary presentation
Prof. Nicholas Christakis, Harvard University

14:30 Poster Session

16:00 Break

16:30 Brief talks:
"Influence and Control"

17:30 Invited plenary presentation
Prof. Martin Nowak, Harvard University

18:00 End

Wednesday June 1

8:30 Registration

9:00 Invited plenary presentation
Prof. Daron Acemoglu, MIT

9:30 Brief talks:
"Opinion, Learning and Games"

11:15 Break

11:45 Brief talks:
"Online Networks"

12:45 Lunch

14:00 Brief talks:
"Data Collection &
Computational Social Science"

16:00 Break

16:30 Poster Session

17:00 Invited plenary presentation
Prof. Sandy Pentland, MIT

17:30 End

Tuesday
May 31

8:30 **Registration**

9:00 **Opening and general overview**

9:30 **Brief talks**

"Communities & Structure"

Methods for clustering multi-layer graphs in mobile networks
X. Dong, *Ecole Polytechnique Federale de Lausanne*
P. Frossard, *Eole Polytechnique Federale de Lausanne,*
P. Vandergheynst, *Eole Polytechnique Federale de Lausanne*
N. Nefedov, *Nokia Research Center*

Robustness and limited modular information in networks
James P. Bargrow, *Northeastern U.*
Sune Lehmann, *Harvard U.*
Yong-Yeol Ahn, *Northeastern U.*

Impact of context on information spreading
Dashun Wang, *Northeastern U.*
Zhen Wen, *IBM*
Hanghang Tong, *IBM*
Ching-Yung Lin, *IBM*
Chaoming Song, *Northeastern U.*
Albert-Laszlo Barabasi, *Northeastern U.*

Computing global social balance in large-scale signed social networks
G. Facchetti, *International School for Advanced Studies*
G. Iacono, *International School for Advanced Studies*
C. Altafini, *International School for Advanced Studies*

Stochastic blockmodels with growing number of classes
D.S. Choi, *Harvard U.*
E.M Airoldi, *Harvard U.*
P.J. Wolfe, *Harvard U.*

Fellows: Crowd-sourcing the evaluation of an overlapping community model based on the cohesion measure
A. Friggeri, *LIP / ENS de Lyon – DNET/INRIA*
G. Chelius, *LIP / ENS de Lyon – DNET/INRIA*
E. Fleury, *LIP / ENS de Lyon – DNET/INRIA*

11:00 **Break**

11:30 **Brief talks**

"Information Propagation"

Monopoly pricing in the presence of social learning
B. Ifrach, *Columbia Business School*
C. Maglaras, *Columbia Business School*
M. Scarsini, *LUISS*

Comparing and visualizing the social spreading of products on a large social network
P.R. Sunsdoy, *Corp. Develop. Markets Telenor ASA*
J. Bjelland, *Corp. Develop. Markets Telenor ASA*
G. Canright, *Corp. Develop. Markets Telenor ASA*
K. Engo-Monsen, *Corp. Develop. Markets Telenor ASA*
R. Ling, *IT U.*

Diffusion and cascading behaviors in random networks
M. Lelarge, *INRIA-ENS*

Analysis of tipping points in social networks for diffusion of innovations
S. Lee, H. Kim & K. Jung, *KAIST*

12:30 **Lunch**

Tuesday
May 31

14:00 Invited plenary presentation

Prof. Nicholas Christakis, Harvard University

Introduced by Prof. Asuman Ozdaglar, MIT

14:30 Poster Session

(please note that all posters will be up during both poster sessions)

Topology discovery of sparse random graphs with few participants
A. Anandkumar, A. Hassidim, J. Kelner

Social network-based interventions
W. An

Parsimonious algorithms for decentralized ranking in social networks
K. Jung, B. Kim, M. Vojnovic

Dual approaches to network science
P.O. Perry, P.J. Wolfe

Degree distributional metric learning
B. Huang, B. Shaw, T. Jebara

On joint diagonalization in network analysis
D. Fay, J. Kunegis, E. Yoneki

Visualizing social networks with structure preserving embedding
B. Shaw, T. Jebara

Ranking: Compare, don't score
A. Ammar, D. Shah

Community detection with fuzzy community structure
Q. Wang, E. Fleury

A new leaders-followers algorithm for detecting overlapping communities in social networks
R. Kanawati

Estimation of dynamic social network structure via online convex programming
M. Raginsky, C. Horn, R. Willett

Dynamic network centralities and saltatoric information transmission
P. Csermely, E. Hazai, H.J.M. Kiss, I.A. Kovacs, A. Mihalik, R. Palotai, G.I Simko, K.Z. Szalay, M. Szalay-Beko, S. Wang

Converging an overlay network to a gradient topology
H. Terelius, G. Shi, J. Dowling, A. Payberah, A. Gattami, K.H. Johansson

14:30 Poster Session (cont'd.)

A model of strategic behavior in networks of influence
M.T. Irfan, L.E. Ortiz

Optimal marketing and pricing over social networks
N. Haghpanah, V.S. Mirrokni

Robust collaborative filtering via convex optimization
Y. Chen, C. Caramanis, S. Sanghavi

Mood, sleep and face-to-face interactions in a co-located family community
S. Moturu, I. Khayal, N. Aharony, A. Pentland

16:00 Break

16:30 Brief talks

'Influence and Control"

Controllability of complex networks
Y-Y. Liu, *Northeastern U.*
J-J. Slotine, *MIT*
A-L. Barabasi, *Northeastern U.*

Information flow and active social influence in social networks
G.C. Chasparis, *Lund U.*
J.S. Shamna, *GTech*

Structural analysis of information dissemination in large-scale networks
V.M. Preciado, *UPenn*
A. Jadbabaie, *UPenn*

'Friendship-based' games
P.J. Lamberson, *MIT*

17:30 Invited plenary presentation

Prof. Martin Nowak, Harvard University

Introduced by Prof. Patrick Jaillet, MIT

18:00 End

Wednesday
June 1

8:30 Registration

9:00 **Invited plenary presentation**

Prof. Daron Acemoglu, MIT

Introduced by Prof. Munther Dahleh, MIT

9:30 **Brief talks**

"Opinion, Learning & Games"

On global games of regime change in networks with non-binary payoffs
M. Dahleh, *LIDS, MIT*
A. Tahbaz-Salehi, *LIDS MIT*
J. Tsitsiklis, *LIDS, MIT*
S. Zoumpoulis, *LIDS, MIT*

A differential games framework for consensus in social networks: From Nash equilibrium to mean-field equilibrium
Quanyan Zhu, *U of IL at Urbana-Champaign*
Tamer Basar, *U of IL at Urbana-Champaign*

Iterative learning from a crowd
David R. Karger, *EECS, MIT*
Sewoong Oh, *EECS, MIT*
Devavrat Shah, *EECS, MIT*

Asymptotic learning on social networks
Elchanan Mossel, *UC Berkeley & Weizmann Inst of Sci.*
Allan Sly, *Microsoft Research*
Omer Tamuz, *Weizmann Inst of Sci.*

Opinion fluctuations and persistent disagreement in social networks
Daron Acemoglu, *Dept of Economics, MIT*
Giacomo Como, *LIDS, MIT*
Fabio Fagnani, *Politecnico di Torino*
Asuman Ozdaglar, *LIDS, MIT*

Opinion formation under peer pressure
Vivek S.Borkar, *Tata Inst of Fundamental Research*
Aditya Karnik, *Enterprise Analytics Group, GM R&D, India Science Lab*

Discovery and security in social network models: Graph-theoretic characterizations
Sandip Roy, *Washington State Univ.*

11:15 **Break**

11:45 **Brief talks**

"Online Networks"

On the unpredictability of elections using social media data
Daniel Gayo-Avello, *Universidad de Oviedo*
Panagiotis T. Metaxas, *Wellesley College*
Eni Mustafaraj, *Wellesley College*

Capturing unobserved correlated effects in diffusion in large virtual networks: distinguishing individual preferences, social connections and cultural discourse influence on the adoption of Twitter clients
Elenna R. Dugundji, *Universiteit van Amsterdam*
Ate Poorthuis, *Universiteit van Amsterdam*
Michiel van Meeteren, *Universiteit van Amsterdam*

Emergence of superstars in online social networks
Devavrat Shah, *LIDS, MIT*
Tauhid Zaman, *LIDS, MIT*

Bias in social and mainstream media
Yu-Ru Lin, *Harvard & Northeastern*
James P. Bagrow, *Dana-Farber Cancer Institute, Boston*
David Lazer, *Harvard & Northeastern*

12:45 **Lunch**

Wednesday
June 1

14:00 Brief talks
"Data Collection & Computational Social Science"

The effects of just-in-time social networks on people's choices in the real world
Kwan Hong Lee, *MIT Media Lab*
Andrew Lippman, *MIT Media Lab*
Alex S. Pentland, *MIT Media Lab*

Localizing externalities in social networks: Inducing peer pressure to enforce socially efficient outcomes
Ankur Mani, *MIT Media Lab*
Iyad Rahwan, *ACM*
Alex (Sandy) Pentland, *MIT Media Lab*

Spread of influence in cellular social networks
Abhik Das, *UT at Austin*
Suriya Gunasekar, *UT at Austin*
Sujay Sanghavi, *UT at Austin*
Sriram Vishwanath, *UT at Austin*

Social networks and research output
Lorenzo Ductor, *University of Alicante*
Marcel Fafchamps, *University of Oxford*
Sanjeev Goyal, *University of Cambridge*
Marco J. van der Leij, *University of Amsterdam*

Automated extraction of social networks from meeting transcripts
David A. Broniatowski, *MIT & Synexxus Quant Analytics*

The effects of corruption on organizational networks and individual behavior
Brandy Aven, *Carnegie Mellon*

Social & spatial network dynamics in the U.S. house of representatives
Clio Andris, *MIT*
Frank Hardisty, *Penn State*

16:00 Break

16:30 Poster Session
(please note that all posters will be up during both poster sessions)

A game theoretic perspective on network topologies
S. Lichter, C. Griffin

Dynamical control of DeGroot learning to shape belief structures in social networks
S. Mandyam, U. Sridhar

Convergence to consensus in multiagent systems and groups of humans: Mathematical results and experimental findings
J. Lorenz

Randomized optimal consensus for multi-agent systems with time-varying communication graphs
G. Shi, K. H. Johansson

Computing symmetric functions on memory bounded social networks
E. Mossel, A. Prakash, G. Valiant

Social consensus through the influence of committed minorities
J. Xie, B.K. Szymanski, S.Sreenivasan, G. Korniss, W. Zhang, C. Lim

On the equilibrium of binary choice models with positive externalities
J. Tipan Verella, S.D. Patek

Mining social networks for customer churn prediction
W. Verbeke, K. Dejaeger, T. Verbraken, D. Martens, B. Baesens

Making sense of the chat dialogs: the network, the communities and the text
F. Shah

Exploiting social network analysis for career paths recommendations
M.F. González-Gutiérrez, O. Morajko, M.P. Miquel, D. Monreal

The market economy of trips
D. Papanikolaou

Social distance drives the convergence of preferences in an online music sharing network
L. Tran, M. Cebrian, C. Krumme, A. Pentland

Wednesday
June 1

16:30 **Poster Session (cont'd.)**

Modeling the coevolution of network structure and node state in a student dorm
W. Dong, A. Madan, A. Pentland

The "Friends and Family" study: Progress report and initial results
N. Aharony, W. Pan, A. Pentland

Information flow in networks: Trendsetters, bellwethers and shepherd dogs
Y. Altshuler, A. Pentland

Network manipulation (with application to political issues)
P.T. Metaxas

Social relevance
E.A. Baatarjav, R. Dantu

Identification of social network actors: A fuzzy based context-dependent approach
M. Fazeen, R. Dantu, P. Guturu

Social media and the 25 January Revolution: Social firestorm or tempest in a teapot?
K. Glasgow

That's what (best) friends are for
D. Liben-Nowell

Fertility decisions and their sensitivity to social networks and family policies
T. Fent, B. Aparicio Diaz, A. Prskawetz

Adopting longitudinal network analysis to investigate the emergence of shared leadership
C. Emery

Using metrics to enable large-scale deliberation
M. Klein

Emerging expertise: Status and influence in electronic networks of practice
S. M.G. Otner

Temporal dimensions of organizational network stability: An example in the context of project teams
E. Quintane, P.E. Pattison, G.L. Robbins, J.M. Mol

Social learning and the dynamic cavity method
Y. Kanoria, A. Montanari, O. Tamuz

17:00 **Invited plenary presentation**
Prof. Alex "Sandy" Pentland, MIT
Introduced by Prof. Devavrat Shah, MIT

17:30 **End**

Talk Abstracts

METHODS FOR CLUSTERING MULTI-LAYER GRAPHS IN MOBILE NETWORKS

Xiaowen Dong †*, Pascal Frossard* †*, Pierre Vandergheynst* † *and Nikolai Nefedov* ‡

† Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
‡ Nokia Research Center (NRC), Lausanne, Switzerland
{xiaowen.dong, pascal.frossard, pierre.vandergheynst}@epfl.ch, nikolai.nefedov@nokia.com

1. INTRODUCTION

Clustering on graphs has been studied extensively for years due to its numerous applications. However, in contrast to the classic problems, clustering in mobile and online social networks brings new challenges. In these scenarios, it is common that observational data contains multiple modalities of information reflecting different aspects of human behavior and social interactions. These interactions may be represented by a multi-layer graph that share the same set of vertices representing users, while having different layers representing different relationships among users. Intuitively, each graph should contribute to a better understanding of the underlying clusters from its own angle. It may be expected that a proper combination of the multiple graphs could lead to a better unified clustering of users' behavior and their social interactions.

In this work we consider different methods to combine multi-layer graphs. In particular, we propose an efficient way to combine spectra of multiple graphs to form a "common spectrum". To verify the suggested approach we tested it using mobile datasets. Also we compare the proposed approach with community detection methods based on modularity maximization over single and multiple layer graphs.

2. GRAPH REGULARIZATION FRAMEWORK

The idea of working with the spectrum of the graph is inspired by the popular spectral clustering algorithm [1]. On a single graph, it applies eigen-decomposition of the graph Laplacian matrix and form a spectral embedding of the original vertices in a low dimensional space. This enhances the intrinsic relationship among vertices so that clustering based on this new representation is usually trivial. The problem is more complicated in case of multiple graph layers. As two recent examples, the authors of [2] use an unified matrix factorization framework to find a common low dimensional representation shared by the multiple graphs in the original space domain, while in [3] the authors propose a co-regularization framework to find such a representation in the graph spectral domain.

In this paper we generalize the one-layer spectral clustering to multiple graphs by finding a common low dimensional representation that captures the characteristics of all graph layers. More specifically, we propose first a graph regularization framework to combine the spectra of two graph layers. The key point is that we treat the eigenvectors of Laplacian matrix from one graph as functions defined on the vertices of another graph. By enforcing the "smoothness" of such functions on the second graph through a regularization framework, we capture the characteristics of both graphs and get a better unified clustering result than using single graphs separately. Moreover, our approach has several interpretations: it can be viewed as a propagation process of the cluster labels on the graph, as well as a framework to minimize a mismatch between the resulting partition and information from each individual graph. Next, we generalize this process to the case which involves more than two graphs.

3. MULTI-RESOLUTION COMMUNITIES DETECTION

To evaluate performance of the suggested approach above we compare it with modularity maximization [4] using fast greedy search algorithm [5]. Note that modularity maximization may give a different number of communities at different layers. On the other hand,

Table 1. MIT datasets: combination of phone-calls, BT and location layers. Evaluation of clustering performance using the proposed and the baseline methods. NMI and RI stand for normalized mutual information Rand index.

	NMI	Purity	RI
The proposed method	0.518	0.712	0.758
Sum of spectral kernels	0.486	0.673	0.729
Sum of norm. adj. matrices	0.484	0.685	0.753
Sum of adj. matrices	0.366	0.641	0.731

Table 2. NRC datasets: combination of phone-calls, BT and GPS layers. Evaluation of clustering performance using the proposed and the baseline methods.

	NMI	Purity	RI
The proposed method	0.395	0.539	0.708
Community detection $^{(*)}$	0.363	0.507	0.628
Sum of norm. adj. matrices	0.381	0.534	0.710
Sum of adj. matrices	0.278	0.475	0.650
Sum of spectral kernels	0.220	0.378	0.570

$^{(*)}$ graph formed by summation of adjacency matrices.

the ground truth data typically is clustered into a fixed number of groups. To obtain the same number of communities at different layers as in the ground truth data we apply random walk approach [6].

In general, the suggested framework of a "common spectrum" may be implemented using the community detection approach (to appear elsewhere).

4. APPLICATIONS TO MOBILE DATASETS

We evaluate performance of the proposed clustering methods on the mobile phone datasets collected by MIT Media Lab [7] and Nokia Research Center (NRC) Lausanne [8]. In particular, we consider graph layers formed by phone-calls, detected WLAN and bluetooth proximity, and GPS locations. Simulations show that our approach to combine graph layers improves reliability of clustering compared to a several base-line methods [2] (see Table 1 and Table 2).

Furthermore, the concept of a "common spectrum" is helpful in analysis of any multimodal data which can be conveniently modeled as multiple graphs. For instance, it would enable us to generalize the normal spectral analysis from one-dimensional to multi-dimensional cases.

5. REFERENCES

[1] J. Shi and J. Malik, "Normalized Cuts and Image Segmentation," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug 2000.

[2] W. Tang, Z. Lu, and I. Dhillon, "Clustering with Multiple Graphs," in *International Conference on Data Mining*, Miami, Florida, USA, Dec 2009.

[3] Abhishek Kumar, Piyush Rai, and Hal Daumé III, "Co-regularized Spectral Clustering with Multiple Kernels," in *NIPS Workshop: New Directions in Multiple Kernel Learning*, 2010.

[4] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, no. 066133, 2004.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and Etienne Lefebvre, "Fast unfolding of communites in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 1742-5468, no. 10, pp. P10008+12, 2008.

[6] R. Lambiotte, J.-C. Delvenne, and M. Barahona, "Laplacian dynamics and multiscale modular structure in networks," *arXiv:0812.1770v3*, 2009.

[7] N. Eagle, A. Pentland, and D. Lazer, "Inferring Social Network Structure Using Mobile Phone Data," in *Proceedings of the National Academy of Sciences*, 2009, vol. 106, pp. 15274–15278.

[8] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila, "Towards Rich Mobile Phone Datasets: Lausanne Data Collection Campaign," in *International Conference on Pervasive Services*, Berlin, Germany, Jul 2010.

Robustness and limited modular information in networks

James P. Bagrow[1,2,*], Sune Lehmann[3,4,†], and Yong-Yeol Ahn[1,2,‡]

[1]Center for Complex Network Research, Northeastern University, Boston, MA 02115.
[2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
[3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA
[4]College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA
[*]bagrowjp@gmail.com [†]sune.lehmann@gmail.com [‡]yongyeol@gmail.com

March 8, 2011

Many complex systems, from power grids and the internet, to the brain and society, can be modeled using modular networks. Modules, densely interconnected groups of elements, often overlap due to elements that belong to multiple modules. The elements and modules of these networks perform individual and collective tasks such as generating and consuming electrical load, transmitting data, or executing parallelized computations. We study the robustness of these systems to the failure of random elements. We show that it is possible for the modules themselves to become uncoupled or non-overlapping well before the network disintegrates. When modular organization is critical to overall functionality, networks may be far more vulnerable than predicted by the usual percolation transition.

Consider a system of interacting elements representing computers, power generators, neurons, etc. These elements perform tasks sufficiently complex that they must work together in densely interconnected modules. These tasks may be parallelized computations, protein biosynthesis, or higher-order neurological functions such as visual processing or speech production. In order for modules to communicate, they much share common elements, so that modules are coupled or overlapping, and the system functions properly only when modules can interact. We ask how these networks respond when a random fraction of elements fail: do the modules become uncoupled before the network loses global connectivity? Random failures provide a toy model of, e.g., a traumatic brain injury or degenerative disease. If enough elements fail, the modules can no longer communicate (higher brain functions are lost) even though the network may remain connected (simpler autonomic responses persist). Likewise, an individual module may fail if too many of its member elements cease to function.

Modular structure can be represented with a bipartite network (Fig. 1a) [1, 2] characterized by two degree distributions, r_m and s_n, governing the fraction of elements that belong to m modules and the fraction of modules that contain n elements, respectively. The average number of modules per element is $\sum_m m r_m \equiv \mu$ and the average number of elements per module is $\sum_n n s_n \equiv \nu$. We derive two networks from the bipartite graph by projecting onto either the elements or the modules: One is the network between elements, while the other is a network where each node represents a module and two modules are linked if they share at least one element. The giant component in the element network disappears when the network loses global connectivity; in the module network it vanishes when the modules become uncoupled (non-overlapping). Before projection elements fail with probability $1 - p$ and are removed from the network. Meanwhile, a module is unable to complete its collective task if fewer than a critical fraction f_c of its original elements remain. These failed modules are removed from the module network but any surviving member elements are not removed from the element network. See Fig. 1b.

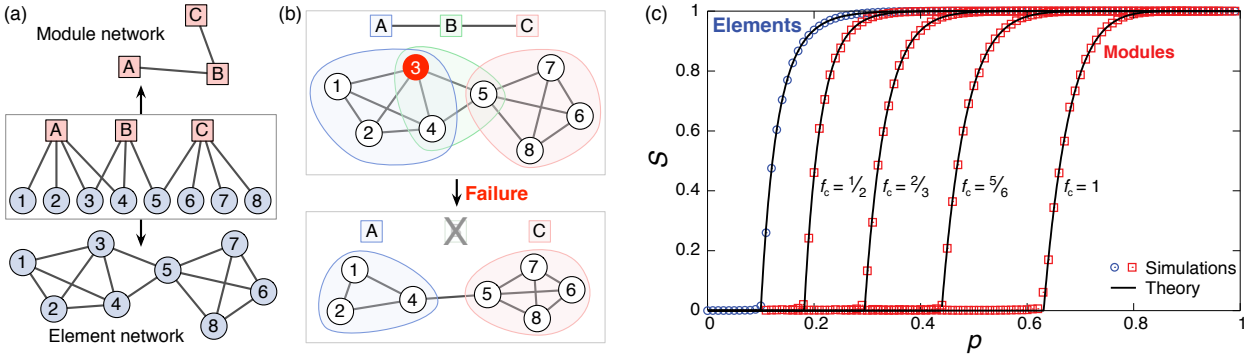


Figure 1: The modular network representation [1, 2]. (a) We obtain two networks by projecting onto elements or modules. (b) The failure of element 3 induces the failure of module B, uncoupling the remaining modules, even though the network itself remains connected. (c) The size of the giant component S for $r_m = \delta(m, \mu)$, $s_n = \delta(n, \nu)$, with $\mu = 3$ and $\nu = 6$. Theory and simulations confirm that the network undergoes a transition from coupled to non-overlapping modules well before it loses global connectivity. Symbols represent element (□) and module (○) networks.

We wish to determine $S(p)$, the fraction of remaining nodes within the giant component as a function of p, for both the element and module networks. We begin with four generating functions [1, 2]:

$$f_0(z) = \sum_{m=0}^{\infty} r_m z^m, \quad f_1(z) = \frac{1}{\mu} \sum_{m=0}^{\infty} m r_m z^{m-1},$$
$$g_0(z) = \sum_{n=0}^{\infty} s_n z^n, \quad g_1(z) = \frac{1}{\nu} \sum_{n=0}^{\infty} n s_n z^{n-1}. \tag{1}$$

These functions generate the probabilities for (f_0) a randomly chosen element to belong to m modules, (f_1) a random element within a randomly chosen module to belong to m other modules, (g_0) a random module to contain n elements, and (g_1) a random module of a randomly chosen element to contain n other elements. For this model, we derive both S and the condition for a giant component to exist in both the element and module networks. In Fig. 1c we show S for $\mu = 3$ and $\nu = 6$. The "robustness gap" between the element and module networks widens as the module failure cutoff increases, covering a significant range of p for the larger values of f_c.

Additionally, we discuss scale-free networks and verify the existence of the robustness gap empirically using a number of real-world datasets. This work can also help us to understand how empirical networks are affected by *missing information*, of critical importance when studying communities. Here p is the probability that a network element is successfully captured by an experiment, such as a high-throughput biological assay or web crawler. The robustness gap can explain how non-overlapping community methods may succeed in networks where overlap is expected: the network is sampled down to the intermediate regime where nodes are connected but modules are uncoupled.

[1] M.E.J. Newman. Properties of highly clustered networks. *Phys. Rev. E*, 68(2):026121, 2003.

[2] M.E.J. Newman and J. Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68(3):036122, 2003.

Impact of Context on Information Spreading

Dashun Wang[1,2,*], Zhen Wen[3], Hanghang Tong[3], Ching-Yung Lin[3], Chaoming Song[1,2], and Albert-László Barabási[1,2,4]

[1]CCNR, Dept. of Physics and Computer Science, Northeastern University, Boston, MA 02115, USA
[2]CCSB, Dana-Farber Cancer Institute, Harvard University, Boston, MA 02115, USA
[3]IBM T.J. Watson Research Center, Hawthorne, NY 10532, USA
[4]Dept. of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[*]dashunwang@gmail.com

March 11, 2011

Information spreading plays an essential role in numerous human interactions, including the spread of innovations, knowledge and information security management, social influence in marketing, and more. Thanks to the increasing availability of large-scale data, we have witnessed great advances in understanding how information propagates from person to person, ranging from incentivized word-of-mouth effects when recommending products, to understanding how a single piece of information forms internet chain letters on a global scale.

Despite recent studies in online social networks, it has been difficult to obtain detailed traces of information dissemination alongside relevant contextual data such as people's real social connections, their behavioral profiles, and job roles in organizations. Therefore, an important question is largely unanswered: *to what extent do spreading processes depend on the underlying social network and behavioral profiles of individuals*. Indeed, on one hand, information such as rumors, innovations and opinions diffuses through the underlying social networks. To whom and to how many people a user would pass such information is constrained by whom s/he connects to and how well s/he is connected in the social network, and the strength of those connections. On the other hand, the population-based heterogeneity in personal profiles coexists with complex connectivities between individuals, raising questions about to what degree the diverse profiles of individuals, from personal interests and expertise to communities and hierarchy, impact the information spreading process. Understanding the role of these features is of fundamental importance.

The lack of contextual information could change drastically, however, thanks to the pervasive use of email communications in well-documented settings, such as corporate work forces. Indeed, emails have become the most important communication method in various settings, unveiling detailed traces of social interactions among large populations. Previous studies have shown that email communications serve as a good indicator of social ties. *Forwarded emails*, written by someone other than the sender and sent to someone who was not included in the original email, serve as an ideal proxy for the information spreading process, where the single piece of information, the original body of the email, is passed through the social network.

We investigate the impact of context on spreading processes in two levels:

- At the *microscopic* level, we are interested in the behaviors of each individual in the spreading process, e.g. to whom and how fast does a user forward information?
- At the *macroscopic* level, we ask what are the structural properties of the spreading processes? And what is the best model for the observed structures?

At the microscopic level, we find that information spreading is indeed highly dependent on social context as well as the individuals' behavioral profiles. Macroscopically, however, we find that the tree structures observed in the spreading process can be accurately captured by a simple stochastic branching model, indicating the macroscopic structures of spreading processes, i.e., to how many people a user forwards the information and the overall coverage of the information, are largely independent of context and follow a simple reproducible pattern.

To the best of our knowledge, this work presents the first comprehensive analysis of the determining factors affecting information spreading processes. We believe our findings are of fundamental importance in developing prediction models for information flow, provide new insights towards the design of our social and collaborative applications, such as assisting users to disseminate information more efficiently, protecting digital information leakage, and promoting spreading strategies to achieve expected coverage.

Computing global social balance in large-scale signed social networks

G. Facchetti, G. Iacono, C. Altafini
SISSA-ISAS, International School for Advanced Studies,
via Bonomea 265, 34136 Trieste, Italy.
Corresponding author: altafini@sissa.it

March 9, 2011

1 Extended Abstract

Signed graphs have been used recently to model social networks of friendship/rivalry or trust/mistrust relationships [1, 4, 6]. Each of these pairwise relationships is represented as a +/- sign on the edge connecting the two agents involved. In terms of social balance theory [2], such bimodal relationships can be used to understand the structure and origin of tensions and conflicts in a community of linked agents.

It has been known for some time how to interpret social balance on such networks: the potential source of tensions are the cycles of the graph (i.e., the closed paths beginning and ending on the same node), notably those of negative sign (i.e., having an even number of negative edges). In particular, a signed graph is exactly balanced (i.e., conflict are completely absent) if and only if all its cycles are positive. The concept of balance is not related to the actual number of negative edges on the cycle but only to their parity, principle which is explained in a simple manner by the notion that "the enemy of my enemy is my friend".

For signed undirected networks, while verifying if the network is *exactly* balanced is an easy problem, which can be answered in polynomial time (e.g. through the calculation of the smallest eigenvalue of the associated Laplacian), computing global balance on a graph not exactly balanced is an NP-hard problem, equivalent to a series of well-known problems:

- computing the ground state of an Ising spin glass;
- solving a MAX-CUT problem;
- computing the distance to monotonicity of a dynamical system (for which the signed graph corresponds to the signature of the Jacobian linearization) [8, 3].

The equivalence with energy minimization of a spin glass has for example been highlighted recently in [1, 7]. In this context, a negative cycle is called a frustration, and frustrations are the trademark of complex energy landscapes, with many local minima whose structure and organization has been so far explored only in special cases, in which the graph is particularly simple. For istance the case studied in [7], the fully connected graph, is unrealistic for real social networks, which have usually heterogeneous connectivity degrees. In fact, for what concerns the real signed social networks currently available, only analysis of local, low-dimensional motifs have been carried out [4, 6]. This amounts essentially to the enumeration of the triangles (i.e., length-3 cycles) and to their classification into frustrated / not frustrated, see [6, 9]. An alternative approach is taken in [5], where spectral properties of the Laplacian are investigated. The

magnitude of the smallest eigenvalue of the Laplacian is indicative of how balanced a network is, i.e., of how much frustration is encoded in the cycles of the networks. Both approaches provide useful information in order to understand the balance of the social networks. Yet this information is partial and unsatisfactory. The small motif analysis, for example, only identifies the frustration on the smallest possible groups of interacting agents, but overlooks more longe range conflicts associated to longer cycles (and larger communities). These might be a considerable amount, particularly for sparse networks. The spectral approach, on the contrary, gives an idea of the overall amount of frustration of the network, but does not provide any information on which nodes are affected by the conflicts. What one would like to have is a method to estimate globally the level of balance and at the same time identify which residual conflicts are uneliminable in the best case, i.e., when the global optimum of the balance is found. In terms of spin glasses, one would like to compute the ground state(s) and, more generally, study the low-energy landscape.

In the proposed presentation we will provide the first efficient heuristic algorithm able to perform this task on large heterogeneous signed graphs. The examples which will be illustrated are drawn from the recent literature on social networks: (i) *Slashdots* [4]; (ii) *Epinions* [6]. These signed networks can be download from the Stanford Network Analysis Platform (http://snap.stanford.edu/). Their size is of the order of 10^5 nodes.

References

[1] T. Antal, P. L. Krapivsky, and S. Redner. Dynamics of social balance on networks. *Phys. Rev. E*, 72(3):036121, Sep 2005.

[2] F. Heider. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–122, 1946.

[3] G. Iacono, F. Ramezani, N. Soranzo, and C. Altafini. Determining the distance to monotonicity of a biological network: a graph-theoretical approach. *IET Systems Biology*, 4:223–235, 2010.

[4] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: Mining a social network with negative edges. In *18th International World Wide Web Conference*, pages 741–741, April 2009.

[5] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, D. L. E. W., and S. Albayrak. Spectral analysis of signed graphs for clustering, prediction and visualization. In *SDM'10*, pages 559–559, 2010.

[6] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, GA, April 2010.

[7] S. A. Marvel, S. H. Strogatz, and J. M. Kleinberg. Energy landscape of social balance. *Phys. Rev. Lett.*, 103(19):198701, Nov 2009.

[8] E. D. Sontag. Monotone and near-monotone biochemical networks. *Syst. Synth. Biol.*, 1(2):59–87, 2007.

[9] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31):13636–13641, 2010.

Stochastic Blockmodels with Growing Number of Classes [Extended Abstract]

David S. Choi*, Edoardo M. Airoldi, Patrick J. Wolfe
Harvard University

Statistical methods that were originally developed for social network analysis are being adopted in diverse fields, such as bioinformatics and marketing. In these new applications (and increasingly in the social sciences as well), the networks are often massive and are only partially observed. As a result, the possibility arises that models which work well in traditional settings may overfit. Using techniques from machine learning, we analyze this possibility for a basic model known as the stochastic blockmodel. This model is useful for identification of communities, which is a highly active research area [2, 3]. Moreover, under an exchangeability result analogous to de Fenetti's result for exchangeable sequences, the stochastic blockmodel model can be viewed as a basic building block for nonparametrics on random networks [1].

The main characteristic of the stochastic blockmodel is that it assigns to each actor a latent class representing his/her community, where the number of classes K is a fixed parameter. Previous results exist proving the asymptotic consistency of maximum likelihood estimation, under various conditions on: 1) the number of classes K 2) the number of edges M and actors N in the observed data. Using arguably simpler methods, we establish asymptotic consistency under more relaxed conditions than previously shown. A rough statement of the main result is that if: 1) K behaves as $\mathcal{O}(\sqrt{N})$, and 2) M behaves as $\omega(N \log^3 N)$, then the maximum likelihood class assignment contains a vanishing fraction of errors when the data is generated from the model. When the data is not assumed to be generated from the model, the analysis is still helpful in that it provides (conservative) confidence regions for estimates of the assortatitivity of the network. Simulation results suggest that the bounds are nearly tight; for example, Fig.

*dchoi@seas.harvard.edu, airoldi@fas.harvard.edu, patrick@seas.harvard.edu

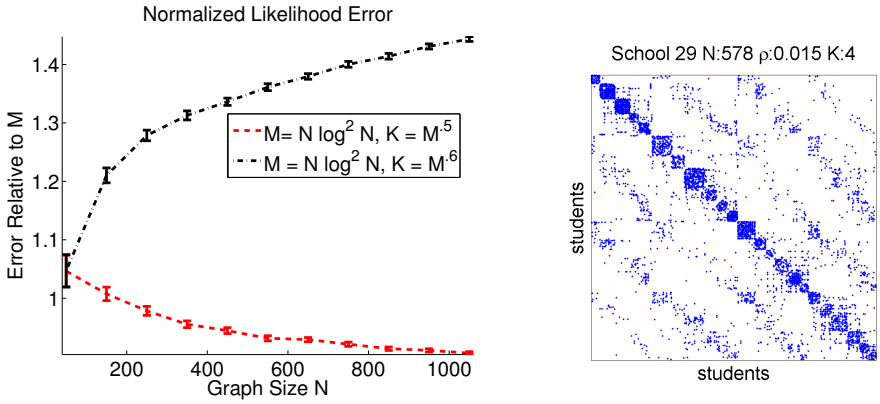


Figure 1: Left: Performance of maximum likelihood estimation for simulated data, showing convergence for $K \sim \sqrt{N}$ and divergence for $K \sim N^{.6}$. Right: Adjacency matrix showing community structure in a friendship network. The structure is statistically significant in that its assortativity is unlikely to be generated from grade, gender, or race differences alone.

1 shows that changing K from $\mathcal{O}(\sqrt{N})$ to $\mathcal{O}(N^{.6})$ in simulations causes a transition from consistency to divergence.

We fit the model to social network data collected from Facebook and the Add-Health study. For the Add-Health data set, the model detects pattens in student friendship networks that are more fine-grained than previously studied. Whereas generally no formal validation method exists for community detection, our analysis is able to validate these patterns as being statistically unlikely to be generated solely by differences in gender, grade, or race.

References

[1] P.J. Bickel and A. Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068, 2009.

[2] B. Karrer and MEJ Newman. Stochastic blockmodels and community structure in networks. *Arxiv preprint arXiv:1008.3926*, 2010.

[3] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577, 2006.

Fellows: Crowd-Sourcing the Evaluation of an Overlapping Community Model based on the Cohesion Measure

Adrien Friggeri, Guillaume Chelius and Eric Fleury
LIP / ENS de Lyon – DNET/INRIA
e-mail: FirstName.LastName@ens-lyon.fr

Although community detection has drawn tremendous amount of attention across the sciences in the past decades, no formal consensus has been reached on the very nature of what qualifies a community as such. Despite the lack of globally accepted analytical definition, all authors concur on the intuitive notion that a community is a relatively dense group of nodes which somehow features less links to the rest of the network. Unfortunately, this agreement does not extend to the specific formal meanings of *dense* and *less links*.

However, the past few years have witnessed a paradigm shift, as the idea of defining the nature of communities was progressively left aside. It has become apparent, and widely accepted, that it suffices to compare several sets of communities and choose the best obtained partition – relative to a given metric, often Newman's Q-modularity – in order to detect communities. In particular, there has been a growing interest in the study of *overlapping* communities; a distribution of the nodes across different groups which reflects more precisely what one might expect intuitively, namely that a given node might belong to different communities . For example, in a social network, an individual might simultaneously belong to a family, a group of friends and co-workers groups. Due to the historical evolution of the field, to this day, most methods used to detect overlapping communities are inspired by, or adapted from, existing counterparts for disjoint community detection.

A novel measure: the cohesion. In this work[1], we take an orthogonal approach by introducing a novel point of view to the problem of overlapping communities. Instead of quantifying the quality of a set of communities, we choose to focus on the intrinsic community-ness of one given set of nodes. In order to do so, we propose a general metric on graphs, the *cohesion*, inspired by sociological considerations. The *cohesion* is a purely local[2] metric based on the notions of weak ties[3] and triangles – triplets of pairwise connected nodes, instead of the classical view using only edge density. Let $G = (V, E)$ a graph and $U \subseteq V$ a subset of nodes, we define the cohesion:

$$\mathcal{C}(G, U) = \frac{\triangle_{\text{in}}(G, U)}{\binom{|U|}{3}} \frac{\triangle_{\text{in}}(G, U)}{\triangle_{\text{in}}(G, U) + \triangle_{\text{out}}(G, U)}$$

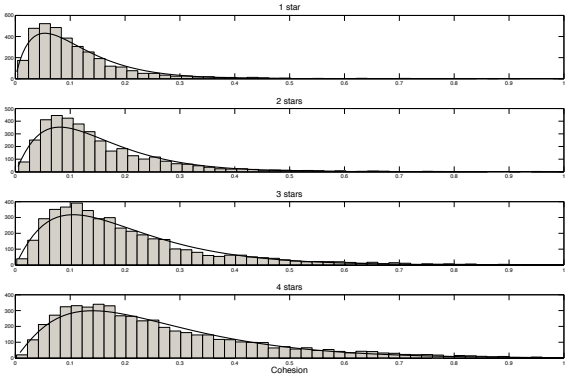
where $\triangle_{\text{in}}(G, U)$ is the number of triangles in U and $\triangle_{\text{out}}(G, u)$ is the number of outbound triangles from U (*i.e.* having two nodes in U and the third one in $V \setminus U$). The first factor is the *triangular density* of U and the second one denotes the quality of the boundaries of U. Intuitively, a community is a set of nodes having high cohesion, *i.e.* high density of triangles and "cutting" few triangles.

Validation *in situ*. In order to validate our model, building on the Facebook API, we have launched Fellows[4], a large-scale online experiment. Fellows is a web application which makes use of the *cohesion* to

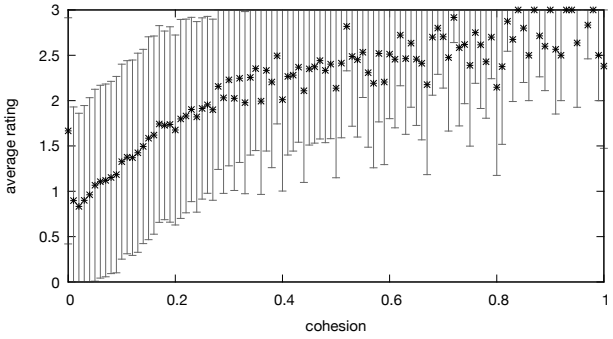
[1] Complementary material are available at http://hal.inria.fr/inria-00565336/en
[2] It only takes into account the considered set of nodes and their neighborhood
[3] Introduced by A. Rapoport in 1957 and developed by M.S. Granovetter in 1973
[4] http://fellows-exp.com

present the user with several highly cohesive groups – which we call *egomunities* – among their Facebook friends. For a given user u, the greedy algorithm used in Fellows focuses on his/her neighborhood $\mathcal{N}(G, u)$. The core idea is to group together neighbors in possibly overlapping egomunities, all containing u. We initialize an egomunity by selecting the node $v_0 \in \mathcal{N}(G, u)$ with highest degree to serve as *seed*. Thus the egomunity contains u and v_0. From that point we iterate and expand the egomunity by adding neighbors as long as it is possible to increase the cohesion. If there are several nodes which addition increases the cohesion, we choose to add the node v which addition maximizes the number of internal triangles $\triangle_{\text{in}}$ – and in the case more than one node satisfies this condition, we select the one which maximizes the number of outbound triangles $\triangle_{\text{out}}$. Once no more node can be added to the egomunity, we start over by selecting the highest degree seed from the sets of neighbors which haven't been assigned to an egomunity and repeat the process until all neighbors are in at least one egomunity.

The user is then asked to rate each egomunity from one to four stars, answering the question "*would you say that this list of friends forms a group for you?*". The user is also offered the opportunity to instantly create a *Friend List* on Facebook corresponding to that group. On the server side, we collect all egomunities with their cohesion and size. This data will allow us to statistically confront our cohesion model to the individual perception of egomunities.



(a) Density of cohesion for egomunities of rating 1 to 4.



(b) Average rating *vs.* cohesion.

Figure 1: Experiment early results

Although the Fellows's experiment is still ongoing at the time of writing, preliminary results are very promising. 1421 participants took the survey, rating 20805 egomunities and creating 7137 lists. On Figure 1(a) we group the rated egomunities by rating and represent the distributions of cohesion for each rating. This shows that on average egomunities with higher rating feature a higher cohesion. Conversely, in Figure 1(b), we plot the average rating obtained by egomunities grouped by cohesion in slices of width $1/100^{\text{th}}$. In turn, this shows that on average, egomunities with higher cohesion tend to obtain higher ratings. Hence we conclude that the *cohesion* metric provides an accurate quantification of an egomunity's quality, in terms of subjective perception.

Ongoing & futur works. Current and future works focus on other aspects of egomunities and cohesion. As *birds of a feather flock together*, we are currently exploring the aggregation of information from egomunities to infer traits of a user – *e.g.* their age, gender, Facebook Likes, etc. Another current work lies in the extension of cohesion to weighted (and/or oriented) networks, by introducing weighted triangles. In a simple unweighted model of social networks, when two people know each other, their is a link between them. In real life however, things are more subtle, as the relationships are not quite as binary: two close friends have a stronger bond than two acquaintances. In this case, weighted networks are a better model to describe social connections. For this reason, we deem necessary to introduce an extension of the cohesion to those networks. Finally, we are also considering the possibility of using greedy cohesion maximizing techniques to compute overlapping communities on whole graphs instead of locally inside the neighborhood of a node.

Monopoly Pricing in the Presence of Social Learning

Extended Abstract for WIDS 2011

Bar Ifrach* Costis Maglaras† Marco Scarsini‡

March 10, 2011

A monopolist will offer a new product to a market of heterogeneous consumers that differ in their willingness-to-pay for that product. Consumers do not know the true quality of the product, but estimate it based on a social learning mechanism and make their purchase decision accordingly. Buyers reveal information about their ex-post utility, specifically if it was positive or negative, but not their type. So we consider a monopolist that operates in a market whose beliefs about the offer product evolve dynamically through a social learning process. Questions of interest that we address are the following: do consumers learn the true quality of the product, and, if so, how fast? taking into account the dynamics of the learning process, what is the revenue maximizing price? can the seller use dynamic pricing to affect the learning process and maximize her discounted revenues?

The market. Agents (consumers) sequentially face the decision of purchasing a product with unknown quality, or choosing an outside option. Each consumer has a willingness-to-pay for the product that is the sum of an idiosyncratic component, which is assumed to be an iid draw from some known distribution, and an unknown component that is common across consumers and depends on the quality of the product.

Information. Agents do not know the true quality of the product. They observe the information reported by a sample -or potentially all- of the past agents, depending on their social network, and make an inference about the product quality. If their quality estimate is sufficiently high given their idiosyncratic valuation parameter, they purchase the product. If they purchase the product, they further "report" whether their ex-post utility was positive (but they do not report their valuation parameter). As such, new agents observe the purchase decisions of a sample of past agents, and for those who bought within that sample they also observe whether they had a positive ex-post utility, i.e., whether they "liked" or "disliked" the product.

Learning. It is typical to assume that fully rational agents update their beliefs for the unknown quality of the product through a Bayesian analysis, but this places an extraordinary analytical and computational onus on each agent that is hard to justify as a model of actual choice behavior. Instead, we will postulate a naive learning process that does not make use of such Bayesian analysis. There is a growing literature in economics that studies naive learning mechanisms that employ simpler and perhaps more plausible learning protocols by each agent, which is also related to the

*Columbia Business School, 4I Uris Hall, 3022 Broadway, NY, NY 10027. Contact author (bi2118@columbia.edu)
†Columbia Business School, 409 Uris Hall, 3022 Broadway, NY, NY 10027. (c.maglaras@gsb.columbia.edu)
‡LUISS (marco.scarsini@luiss.it)

growing body of engineering literature on sensor networks and decentralized algorithms. Almost exclusively the above works focus on the learning dynamics and do not study their interplay with pricing and revenue optimization in a market setting.

Our model postulates a fairly simple learning mechanism. Given their social network, agents observe the information reported from a sample of past agents that is summarized by the size of the sample, the number of people that purchased, and the number of people that purchased and liked the product. With this information agents form an estimate for the preferences of the marginal agent that purchased and liked the product, and from that they extract an estimate of the product quality. In doing so, our agent disregards the learning dynamics, and acts as if all prior agents made their decisions based on some common belief about the product quality, which is now easy to characterize.

Convergence of learning process. We first provide a negative result, specifically, that learning may fail in our model in the absence of experimentation; that is, consumers may not learn - even asymptotically- the true quality of the product. The underlying issue is that consumers get censored observations of the product quality, and learning may stop at an inferior quality estimate where all buyers report that they like the product, thus providing a censored quality estimate. On the positive side, we show that learning will eventually occur almost surely if consumers employ any degree of experimentation, or doubtfulness in the accuracy of their predecessors' decisions. Learning can be shown to occur for two types of social networks, the one where each agent observes the entire sequence of past agent decisions, and the one where each agent independently observes each of his predecessors with some probability but where the size of the sample grows large as the number of agents increases to infinity. Asymptotic learning also holds if agents weigh each of their predecessors' reports differently based on the predecessor's location on the arrival sequence. We characterize conditions on these weight trajectories that guarantee asymptotic learning.

Mean-field approximations, learning trajectories and pricing. The speed of convergence, and, better yet, the learning trajectory over time is essential in capturing the tradeoff between consumer learning and the monopolist's discounted revenue objective. We derive a mean filed (fluid model) approximation for the learning dynamics in an asymptotic regime, where the rate of arrival of new consumers to the system grows large, and where the mass of each individual consumer proportionally decreases – this type of scaling is often referred to as uniform acceleration. We show that the asymptotic learning trajectory is characterized by a differential equation, which in some special cases is solvable in closed form. We show how the learning speed depends on the weight function used by consumers, and finally explore the impact of these results on the seller's pricing decision.

Comparing and visualizing the social spreading of products on a large social network

Pål Roe Sundsøy (pal-roe.sundsoy@telenor.com),
Johannes Bjelland (Johannes.bjelland@telenor.com),
Geoffrey Canright (geoffrey.canright@telenor.com),
Kenth Engø-Monsen (Kenth.engo-monsen@telenor.com)
Corporate Development, Markets,Telenor ASA,Oslo, Norway

Rich Ling (rili@itu.dk)
IT University, Copenhagen Denmark and Corporate Development, Markets Telenor ASA Oslo, Norway

Abstract— By combining mobile traffic data and product adoption history from one of the markets of the telecom provider Telenor, we define and measure an adoption network—roughly, the social network among adopters. We study and compare the evolution of this adoption network over time for several products – the iPhone handset, the Doro handset, the iPad 3G and videotelephony. We show how the structure of the adoption network changes over time, and how it can be used to study the social effects of product diffusion. Specifically, we show that the evolution of the Largest Connected Component (LCC) and the size distribution of the other components vary strongly with different products. We also introduce simple tests for quantifying the social spreading effect by comparing actual product diffusion on the network to random based spreading models. As videotelephony is adopted pairwise, we suggest two types of tests: transactional- and node based adoption test. These tests indicate strong social network dependencies in adoption for all products except the Doro handset. People who talk together, are also likely to adopt together. Supporting this, we also find that adoption probability increases with the number of adopting friends for all the products in this study. We believe that the strongest spreading of adoption takes place in the dense core of the underlying network, and gives rise to a dominant LCC in the adoption network, which we call "the social network monster". This is supported by measuring the eigenvector centrality of the adopters. We believe that the size of the monster is a good indicator for whether or not a product is going to "take off".

Diffusion and Cascading Behavior in Random Networks

Marc Lelarge
INRIA - ENS*
Marc.Lelarge@ens.fr

Abstract submitted to:
WIDS Workshop on Information and Decision in Social Networks.
The full version of the paper is available at:
http://arxiv.org/abs/1012.2062

The spread of new ideas, behaviors or technologies has been extensively studied using epidemic models. Here we consider a model of diffusion where the individuals' behavior is the result of a strategic choice. We study a simple coordination game with binary choice and give a condition for a new action to become widespread in a random network. We also analyze the possible equilibria of this game and identify conditions for the coexistence of both strategies in large connected sets. Finally we look at how can firms use social networks to promote their goals with limited information. Our results differ strongly from the one derived with epidemic models and show that connectivity plays an ambiguous role: while it allows the diffusion to spread, when the network is highly connected, the diffusion is also limited by high-degree nodes which are very stable.

To illustrate our point, consider the basic game-theoretic diffusion model proposed by (Morris, 2000). Consider a graph G in which the nodes are the individuals in the population and there is an edge (i, j) if i and j can interact with each other. Each node has a choice between two possible behaviors labelled A and B. On each edge (i, j), there is an incentive for i and j to have their behaviors match, which is modeled as the following coordination game parameterised by a real number $q \in (0, 1)$: if i and j choose A (resp. B), they each receive a payoff of q (resp. $(1-q)$); if they choose opposite strategies, then they receive a payoff of 0. Then the total payoff of a player is the sum of the payoffs with each of her neighbors. Consider a network where all nodes initially play A. If a small number of nodes are forced to adopt strategy B and other nodes in the network apply best-response updates, then these nodes will be repeatedly applying the following rule: switch to B if enough of your neighbors have already adopted B. There can be a cascading sequence of nodes switching to B such that a network-wide equilibrium is reached in the limit. Most of the results on this model are restricted to deterministic (possibly infinite) graphs. In this work, we analyze the diffusion in the large population limit when the underlying graph is a random network $G(n, \mathbf{d})$ with n vertices and where $\mathbf{d} = (d_i)_1^n$ is a given degree (i.e. number of neighbors) sequence, similarly to (Jackson and Yariv, 2007).

*23 avenue d'Italie, 75013 Paris, France, tel:+33(0)1.39.63.55.33, fax:+33.(0)1.39.63.79.88

In this simple model, agents play a local interaction binary game where the underlying social network is modeled by a sparse random graph. First considering the deterministic best response dynamics, we compute the contagion threshold for this model, confirming the heuristic result of (Watts, 2002). We find that when the social network is sufficiently sparse, the contagion is limited by the low connectivity of the network; when it is sufficiently dense, the contagion is limited by the stability of the high-degree nodes. This phenomenon explains why contagion is possible only in a given range of the global connectivity (i.e. the average number of neighbors).

We identify the set of agents able to trigger a large cascade: the pivotal players, i.e. the largest component of players requiring a single neighbor to change strategy in order to follow the change. When contagion is possible, both in the low and high-connectivity cases, the number of pivotal players is low, resulting in rare occurences of cascades. However in the high-connectivity case, we found that the system displays a robust-yet-fragile quality: while the cascades are very rare, their sizes are very large. This feature makes global contagions exceptionally hard to anticipate.

Motivated by social advertising, we also consider cases where contagion is not possible if the set of initial adopters is too small, i.e. a negligible fraction of the total population, as in (Galeotti and Goyal, 2009). We compute the final size of the contagion as a function of the fraction of the initial adopters. We find that the low and high-connectivity cases still have different features: in the first case, the global connectivity helps the spread of the conatgion while in the second case, high connectivity inhibits the global contagion but once it occurs, it facilitates its spread.

We also analyze possible equilibria of the game and in particular, we find conditions for the existence of equilibria with co-existent conventions. Finally, we analyze a general percolated threshold model for the diffusion allowing to give different weights to the (anonymous) neighbors. This model allows us to study rigorously semi-anonymous threshold games of complements with local interactions on a complex network. Our general analysis gives explicit formulas for the spread of the diffusion in terms of the initial condition, the degree sequence of the random graph, and the distribution of the thresholds.

References

Galeotti, A. and S. Goyal (2009). Influencing the influencers: a theory of strategic diffusion. *RAND Journal of Economics 40*(3), 509–532.

Jackson, M. O. and L. Yariv (2007). Diffusion of behavior and equilibrium properties in network games. *The American Economic Review 97*(2).

Morris, S. (2000). Contagion. *Rev. Econom. Stud. 67*(1), 57–78.

Watts, D. J. (2002). A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci. USA 99*(9), 5766–5771 (electronic).

Analysis of Tipping Points in Social Networks for Diffusion of Innovations

Seulki Lee
KAIST
sklee19@kaist.ac.kr

Hyuna Kim
KAIST
hyunak@kaist.ac.kr

Kyomin Jung
KAIST
kyomin@kaist.edu

Tipping point phenomena (events that had rarely observed becomes suddenly common) for diffusion of innovations have received huge attention from academia and industry [4,6,12]. Understanding tipping point phenomena has numerous applications including viral marketing and minimizing the spread of contamination. Depending on the characteristics of the information and social network structures, the information either cascades globally or terminates quickly. For example, sometimes new technologies become widespread over the network (a global cascade), but in some cases they simply disappear in a short time. In this work, we identify conditions for the occurrence of tipping points for general classes of network structures and provide a novel proof for its correctness.

Various models of information spreading have been studied. These models are established based on the common assumption that the neighbors play significant roles for the spread of information. The SIR (Susceptible-Infected-Recover) model is one of those popular models applied to the cases when accepting the information requires low costs, such as the epidemics of contagious diseases [1, 2, 8]. Under the SIR model, some sufficient conditions for a global cascade have been studied [2, 3, 9]. On the other hand, for the diffusion of new technologies or innovations which requires relatively high costs to adopters, the linear threshold model is widely used [7, 12, 13]. However, general conditions for a global cascade under the linear threshold model are known for restricted cases.

In the linear threshold model, individuals make their decisions based on the decisions of their neighbors. Each node has its own threshold value and if the fraction of neighbors who have already adopted the innovation is greater than the threshold, it will adopt the innovation. The mechanism of this model is originated from the utility maximization of individuals in game theory.

A tipping point is defined as the number of initial adopters x so that the cascade size becomes suddenly large as the number increases from x to $x + \delta$ for a small δ. Under the linear threshold model, the mechanism how a tipping point arises in a complete graph is well known [5]. When the thresholds of all nodes are homogeneous, the average cascade size and the number of initial adopters that triggers a global cascade have been predicted in the case of Erdös-Rényi random graph networks [13]. However, it is known that distributions of thresholds usually follow diverse unimodal distributions such as the normal distribution [10, 11].

In this work, we consider any distributions of thresholds and assume that each node takes its threshold value from the distribution independently. We first analyze that in a social networks including Facebook and Myspace, tipping points occur almost always if certain conditions on the distribution of thresholds are met. We provide a novel proof that under those conditions, a tipping point occurs almost surely for

any graphs whose nodes' degrees are $\omega(\log n)$, where n is the number of nodes. Our proof can be applied to any distributions of thresholds such as the uniform, the normal and homogeneous distributions, and it works for any class of graphs with reasonably high degrees. We also numerically analyze that in graphs having nodes with $O(\log n)$ degrees, the similar result holds.

Secondly, we conducted extensive experiments on real world social networks such as Facebook and Myspace, and synthetic network graphs including Erdös-Rényi random graphs, generalized random graphs with expected degree sequences, and scale-free networks generated by the preferential attachment process. We discover that tipping points indeed appear in these graphs if similar conditions on threshold distributions are met. In order to investigate properties of tipping points, we performed experiments on various network structures with regard to the network size, degree distributions and their community structures. We obtain strikingly similar results from several independent network graphs and conclude that even though some properties of network structures can affect tipping points to some extent, the distribution of thresholds are much more relevant to them.

References

[1] F. Brauer and C. Castillo-Chavez. Mathematical models in population biology and epidemiology. *Springer*, 2001.

[2] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *The European Physical Journal B*, 10, 2008.

[3] F. Chung and L. Lu. The volume of the giant component of a random graph with given expected degrees. *The SIAM Journal on Discrete Mathematics*, 20:395–411, 2006.

[4] M. Gladwell. The tipping point: How little things can make a big difference. *Back Bay Books*, 2000.

[5] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83:1420–1443, 1978.

[6] S. Lohmann. The dynamics of informational cascades: The monday demonstrations in leipzig, east germany, 1989-91. *World Politics*, 47:42–101, October 1994.

[7] D. Lopez-Pintado. Contagion and coordination in random networks. *International Journal of Game Theory*, 34(3):371–381, 2006.

[8] N. Madar, T. Kalisky, R. Cohen, D. Ben-avraham, and S. Havlin. Immunization and epidemic dynamics in complex networks. *The European Physical Journal B*, 38:269276, 2004.

[9] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review*, 64, 2001.

[10] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion, political hashtags, and complex contagion on twitter. *World Wide Web*, 2011.

[11] T. W. Valente. Social network thresholds in the diffusion of innovations. *Social Networks*, 18:69–89, 1996.

[12] D. Watt. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:5766–5771, 2002.

[13] D. Whitney. Cascades of rumors and information in highly connected networks with thresholds. *Second International Symposium on Engineering Systems*, 2009.

Controllability of Complex Networks

Yang Liu[1,2], Jean-Jacques Slotine[3,4], & Albert-László Barabási[1,2,5]

[1]*Center for Complex Network Research and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA*

[2]*Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA.*

[3]*Nonlinear Systems Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA*

[4]*Department of Mechanical Engineering and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, USA*

[5]*Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA.*

yangliu61@gmail.com

jjs@mit.edu

alb@neu.edu

The ultimate proof of our understanding of natural or technological systems is reflected in our ability to control them. While control theory offers mathematical tools to steer engineered and natural systems towards a desired state, we lack a general framework to control complex self-organized systems, like the regulatory network of a cell or the Internet. Here we develop analytical tools to study the controllability of an arbitrary complex directed net- work, identifying the set of driver nodes whose time-dependent control can guide the system's dynamics. We apply these tools to several real networks, finding that the number of driver nodes is determined mainly by the network's degree distribution. We show that sparse in- homogeneous networks, which emerge in many real complex systems, are the most difficult to control, but dense and homogeneous networks can be controlled

via a few driver nodes. Counterintuitively, we find that in both model and real systems the driver nodes tend to avoid the hubs. We show that the robustness of control to link failure is determined by a core percolation problem, helping us understand why many complex systems are relatively insensitive to link deletion. The developed approach offers a framework to address the controllability of an arbitrary network, representing a key step towards the eventual control of complex systems.

Information Flow and Active Social Influence in Social Networks

Georgios C. Chasparis* Jeff S. Shamma†

March 10, 2011

When individuals in a social network exchange information, beliefs or opinions through their immediate connections, the following questions naturally emerge:

1. *What are the social networks which most likely form* when individuals are concerned with the efficient and effective dissemination of *endogenous* information through the network?
2. Given a network of connections, *what is the optimal targeting policy* for which an *exogenous* belief can be adopted to the largest extent by the network?

The above questions, although different, overlap to a large degree. On the one hand, we recognize that individuals are dynamically changing their links to search for efficient information flow through the network. In this case, we are interested to know what are the networks which most likely are going to form. On the other hand, when information or beliefs are exogenously implanted to the network, adoption of these beliefs will highly depend on which individuals are initially targeted and what is their influence to the network (i.e., their centrality measure). In the following discussion, we analyze these two questions independently.

The first part of this discussion is motivated by the current research on social network formation [1, 2] and how social networks form when individuals have discretion over the links they establish or sever. We model the problem as a noncooperative game, where each individual makes decisions based on myopic considerations, i.e., so that its own utility is maximized. Links are assumed unidirectional, which model phenomena such as web links, observations of others, citations, etc. [2]. The utility considered for each individual reflects the ability to disseminate information efficiently through the network similarly to [3, 4].

Several models for endogenous network formation have been proposed that are based on game theoretic formulations. These include *static models*, [3], where agents play an one-stage game, with actions corresponding to network links. These studies characterize networks in terms of the Nash equilibria of the associated game, called *Nash networks*. The processes under which such equilibria emerge are proposed via *dynamic* or *evolutionary* models [4, 5, 6]. In these models, players adaptively form and sever links in reaction to an evolving network, and in some models, their decisions are subject to small random perturbations.

Our approach is also concerned with dynamic or evolutionary models, and is mostly related to the papers of [4, 5]. Our contributions are the following: i) We discuss the case where nodes can form links only with a subset of the other nodes (i.e., neighborhood structures), as opposed to the entire network; ii) We introduce utility functions that are distance-dependent variations of the *connections model* of [3] and guarantee that Nash networks exist; iii) We introduce state-dependent utility functions that can model dynamic phenomena such as *establishment costs*; iv) We derive a learning process that guarantees convergence to Nash equilibria for the state-based extension of weakly acyclic games; and v) We employ *payoff-based* dynamics for convergence to Nash networks based on a reinforcement learning scheme and drop the typical assumptions that nodes have knowledge of the full network structure and can compute optimal link decisions.

The second part of this discussion is concerned with the derivation of optimal targeting policies for the diffusion of beliefs in a social network. Equivalently, we may think of the targeting policies as advertising strategies and the

*G. Chasparis is with the Department of Automatic Control, Lund University, 221 00-SE Lund, Sweden; E-mail: georgios.chasparis@control.lth.se; URL: http://www.control.lth.se/chasparis.

†J. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: shamma@gatech.edu. URL: http://www.prism.gatech.edu/˜jshamma3.

individuals as customers. Contrary to the first part of the discussion, here the network is assumed constant and the customers' preferences are affected by both their neighbors and the incentives provided through advertising. Our contribution lies in the inclusion of three important factors in the derivation of an optimal advertising strategy: i) dynamic network effects in the formation of preferences, ii) possible misspecifications/uncertainties in the assumed model of evolution of preferences, and iii) uncertainty in the intentions of a competitive firm that also tries to influence the network.

Prior work has focused on i) the derivation of dynamic models which capture the sales response to advertising, and ii) the computation of an optimal policy of advertising as a function of the sales. Those models which capture the effect of advertising on sales, usually assume the following behavior: i) advertising effects persist over the current period but diminish with time [7], ii) marginal advertising effects diminish or remain constant with the size of advertising [8], iii) advertising effects diminish with the size of sales [7], iv) advertising effects diminish with the size of competitive advertising [9], and v) advertising effects are affected by word-of-mouth communication (or excess advertising) [10].

Our model is related to the sales response models [7] (which capture the evolution of the rate of sales) and diffusion models [11] (which capture the market growth). It exhibits diminishing returns with time in the absence of advertising effort, constant marginal returns with the size of advertising, and diminishing returns with the size of competitive advertising. It emanates from traditional advertising models by also considering the effect of word-of-mouth communication through a network of interactions similarly to [12]. The difference here is that the dynamics of preferences become part of the optimization. We derive analytically optimal advertising strategies and relate them to centrality measures usually considered in sociology [13]. This result also establishes a connection with the first part of our discussion, since nodes of high centrality measure can be provided through an analysis of endogenous network formation.

We also consider the possibility that we are uncertain of the accuracy of the model of preferences' update, instead of assuming a deterministic update. This form of uncertainty is usually neglected in prior work on optimal advertising. We derive optimal policies which are robust to a norm-bounded uncertainty. We show that the model exhibits a certainty equivalence property, since the optimal policy for the perturbed model coincides with the optimal policy for the unperturbed model. Finally, we consider the possibility that a competitive firm also tries to influence the network, introducing a second form of uncertainty. In this case, we compute robust optimal policies through the notion of Stackelberg and Nash solutions.

References

[1] M. Jackson, *Social and Economic Networks*. Princeton University Press, 2008.

[2] S. Goyal, *Connections: An Introduction to the Economics of Networks*. Princeton, NJ: Princeton University Press, 2007.

[3] M. Jackson and A. Wolinsky, "A strategic model of social and economic networks," *Journal of Economic Theory*, vol. 71, pp. 44–74, 1996.

[4] V. Bala and S. Goyal, "A noncooperative model of network formation," *Econometrica*, vol. 68, no. 5, pp. 1181–1229, 2000.

[5] B. Skyrms and R. Pemantle, "A dynamic model of social network formation," *Proc. of the National Academy of Sciences of the USA*, vol. 97, pp. 9340–9346, 2000.

[6] M. Jackson and A. Watts, "The evolution of social and economic networks," *Journal of Economic Theory*, vol. 106, no. 2, pp. 265–295, 2002.

[7] M. L. Vidale and H. B. Wolfe, "An operations research study of sales response to advertising," *Operations Res.*, vol. 5, pp. 370–381, 1957.

[8] G. Leitmann and W. E. Schmitendorf, "Profit maximization through advertising: A nonzero sum differential game approach," *IEEE Transactions on Automatic Control*, vol. 23, no. 4, pp. 646–650, 1978.

[9] J. Case, *Economics and the competitive process*. New York, NY: New York University Press, 1979.

[10] S. Jørgensen, "A survey of some differential games in advertising," *Journal of Economic Dynamics and Control*, vol. 4, pp. 341–369, 1982.

[11] F. M. Bass, "New product growth model for consumer durables," *Management Science*, vol. 15, pp. 215–227, 1969.

[12] P. K. Dubey, B. D. Meyer, and R. Garg, "Competing for customers in a social network," Cowles Foundation, Tech. Rep. 1591, Nov. 2006.

[13] P. Bonacich, "Power of centrality: A family of measures," *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

Structural Analysis of Information Dissemination in Large-Scale Networks

Victor M. Preciado and Ali Jadbabaie
Department of Electrical and Systems Engineering
University of Pennsylvania
*Philadelphia, PA 19104 USA**

During the last decade, the complex structure of many large-scale networked systems has attracted the attention of the scientific community. The availability of massive databases describing these networks allow researchers to explore their structural properties with great detail. Statistical analysis of empirical data has unveiled the existence of multiple common patterns in a large variety of network properties, such as power-law degree distributions, or the small-world phenomenon. Aiming to replicate these structural patterns, a rich variety of synthetic network models has been proposed in the literature, such as the classical Erdös-Rényi random graph and its generalizations, the preferential attachment model proposed by Barabási and Albert, or the small-world network proposed by Watts and Strogatz.

Synthetic network models are useful to analyze the performance of communication protocols, as well as to predict and control network evolution, and to study network reliability and survivability. In this direction, a fundamental question is to understand the impact of a particular structural property in the performance of the network. The most common approach to address this question is to use synthetic network models in which one can prescribe the structural property under study. The impact of structural features, such as degree distributions, clustering, correlations, or hierarchy, has been widely studied in the literature via synthetic network models. Although this approach is very common in the literature, it presents two major flaws:

1. Synthetic network models implicitly induce many structural properties that are not directly controlled and can be relevant to the network dynamical performance. Therefore, it is difficult to isolate the role of a particular structural property in the network performance using synthetic network models.

2. There is no systematic technique to decide what structural properties are most relevant to the network dynamical performance.

In this paper, we propose an alternative approach that overcomes the two issues mentioned above. Our approach is based on studying how structural properties of the network, such as the distribution of degrees, triangles and other substructures, impose bounds on performance metrics of the network. In our analysis, we exploit the close relationship between the eigenvalues of a network and the performance of many dynamical processes taking place in the network, in particular, the performance of viral spreading processes. Our work builds on algebraic graph theory and convex optimization to find optimal bounds on relevant spectral properties of networks from structural information. We illustrate our approach by studying epidemic-style processes of information dissemination in networks. Our results are relevant in many real scenarios, from rumor spreading in online social networks, to malware propagation in computer networks, or information dissemination in communication networks.

'Friendship-based' Games

PJ Lamberson*

Abstract

This paper analyzes a model of games played on a social network that employs a 'friendship-based' approach, which allows for local correlation in players' strategies. The model is applied to two specific classes of games, games of strategic complements and strategic substitutes. We also examine the dependence of diffusion on network clustering – the probability that two individuals with a mutual friend are friends of each other – which is not possible in previous frameworks. We find that clustering has a negative impact on welfare in games of substitutes, but has a positive impact in games of complements. These results imply that clustering may lead to redundant overprovision of public goods, but can help beneficial new technologies break into a market that is dominated by a less preferred standard.

People rarely make decisions in isolation. Often, the choices and experiences of friends, family, and acquaintances shape our beliefs and behavior. This observation motivates a stream of recent research that addresses the effects of social network structure on behavior (Jackson and Yariv, 2005, 2007; Jackson and Rogers, 2007; López-Pintado, 2008; Lamberson, 2009, 2010; Galeotti et al., 2010). All of these articles employ a "mean-field analysis" borrowed from physics to understand how equilibrium diffusion levels depend on the structure of social interactions.

The mean-field approach requires one of two interpretations. Either, the analysis is thought of as an approximation to the true diffusion dynamics, or agents are assumed to have limited information about their social contacts: the agents act as if the behavior of their neighbors matches the behavior of the population as a whole. In order to gain analytic tractability, the method discards much of the connectivity information of an actual network, retaining only the underlying degree distribution. In particular, correlation in neighboring agents' behavior is lost. For example, in an epidemic sick people are more likely to be connected to other sick people, and when a new technology spreads technology adopters are more likely to be connected to other adopters. These effects are lost in the mean-field approach.

*MIT Sloan, E62-441, Cambridge, MA 02139. pjl@mit.edu.

This paper develops an alternative framework that keeps track of local correlations using a "pair approximation" (Matsuda et al., 1992; Keeling et al., 1997; Morris, 1997; Van Baalen, 2000). In this model, the focus is on friendship ties rather than individual actors, so we refer to it as a 'friendship-based' game. Our paper is most closely related to the "network games" framework of Galeotti et al. (2010). However, the partial information structure of the network games framework corresponds with a mean-field approximation rather than the pair approximation that we employ. From an information perspective the friendship-based game allows us to capture a richer information structure – rather than assume that an agent expects her neighbors to play like the population as a whole (conditional on her degree), we assume that the agent expects her neighbors to play like the population *conditional on her own behavior*. If we think of both the network games and friendship-based games as approximations to a process occurring in a fixed social network, simulations demonstrate that the additional information regarding local correlations included in the friendship-based model lead to more accurate results.

We apply the model to two classes of games, games of strategic complements and games of strategic substitutes. We find that games of strategic complements tend to exhibit multiple adoption equilibria separated by a tipping point analogous to the "epidemic threshold" in disease spread models. Locally, agents rapidly split into clusters using competing strategies. Games of strategic substitutes tend towards a unique equilibrium. Agents strategies are locally dissociative. For example, if we think of the model as capturing provision of a public good, a few agents serve the role of local providers of the good while their neighbors free ride. For substitutes, the friendship-based framework predicts a more efficient outcome than the network games framework because agents take more information into account in their decision.

We also extend the framework to take network clustering into account. In many empirical settings, two agents that share a mutual friend are likely to be friends of each other (Newman and Park, 2003; Watts, 2004). This feature, known as clustering or triadic closure, cannot be modeled using a mean-field approach. After adjusting the friendship-based model to incorporate clustering we examine the dependence of predicted equilibria in games of complements and substitutes on clustering as quantified by the clustering coefficient of the network. We find that in games of substitutes clustering decreases efficiency. In the case of public goods provision, this implies that clustered networks require more agents to provide the public good than non-clustered networks. In contrast, in games of complements clustering increases total welfare by helping new beneficial technologies break into a market that is dominated by a less preferred standard. Intuitively, clustering protects the new technology by shielding tightly coupled communities of early adopters from the current standard.

On Global Games of Regime Change in Networks with Non-Binary Payoffs

M. Dahleh, A. Tahbaz-Salehi, J. Tsitsiklis, and S. Zoumpoulis*

Global games of status quo subversion — coordination games of incomplete information in which a status quo is abandoned once a sufficiently large fraction of agents attacks it — have been used to study crisis phenomena such as currency attacks (e.g., Morris and Shin [1998]), debt crises (e.g., Morris and Shin [2004]), bank runs (e.g., Goldstein and Pauzner [2005]), and political regime change (e.g., Edmond [2005]).

To the best of our knowledge, all existing applications of such games to crises assume a continuum of agents and a private (and possibly, in addition, a public) noisy signal of the fundamentals for each agent (there are no complex patterns of communication among the agents). In this work, we propose a model involving a *discrete* number of agents, interconnected through an underlying *network*.

We study a game of regime change with a finite number of agents, in which each agent receives and shares noisy signals concerning the strength of the status quo (i.e., the fundamentals) according to her position in a social network. She can then either attack the status quo or not attack. Attacking can net a positive or negative payoff and is thus a risky action. Not attacking nets 0 payoff and is thus a safe action. In contrast to most of the literature on regime change, we assume a non-binary payoff structure. It is common in the relevant literature (e.g., Angeletos et al. [2007]) to model games of regime change so that payoffs incur a discrete change when the regime changes. In those models, the outcome of a collective attack against the regime is determined by the relative strength of the collective attack and the regime; once the outcome is determined, individual payoffs for attackers depend merely on the (binary) outcome, not on the relative strength of the collective attack and the regime. We consider a variation in which payoffs are not discrete: individual payoffs for attackers depend directly on the relative strength of the collective attack and the regime.

Our game admits a variety of interpretations and applications, in all of which beliefs have the

*All authors are with the Laboratory of Information and Decision Systems, at MIT. Their emails are {dahleh, alirezat, jnt, szoumpou}@mit.edu.

same self-fulfilling nature. Prominent examples are currency attacks (when a large speculative attack forces the central bank to abandon the peg), bank runs (when a large number of bank customers withdraw their deposits because they believe the bank is, or might become, insolvent), debt crises (when a country/company fails to coordinate its creditors to roll over its debt and is hence forced into bankruptcy), and political protests (when a large number of citizens decide whether or not to take actions to subvert a repressive dictator or some other political establishment).

In this work we seek to quantify the connection between the topology of the social network and the predictability of individual behavior in large networks, as well as the connection between the topology of the social network and individual attitude towards risk. We provide an algorithm for the characterization of strategies that survive IESDS for any finite network that is a union of disconnected cliques. We prove that for each agent, all the information about the strength of the status quo can be summarized in a one-dimensional statistic, the average of the observations: in any strategy profile that survives IESDS, each agent chooses the risky action (attack) if the average of her observations is less than a threshold t_R, and chooses the safe action (not attack) if the average of her observations is greater than some threshold t_S; in addition, any strategy profile that satisfies these two conditions survives IESDS. For the special case of cliques of equal size, we provide a characterization involving closed-form analytical expressions.

In a network consisting of finitely many disconnected agents, there is a unique strategy profile that survives iterated elimination of strictly dominated strategies, and therefore a unique Bayesian Nash equilibrium; we argue that for a finite network, a single link suffices to induce multiplicity. Of more interest is the asymptotic regime (as the number of agents grows large), in which some non-trivial network topologies guarantee uniqueness; we obtain sufficient conditions on the network topology for uniqueness in the asymptotic regime.

As Angeletos and Werning [2006] put it, "it is a love-hate relationship: economists are at once fascinated and uncomfortable with multiple equilibria." In this work, we identify the social network topology as the determining factor with respect to the dichotomy between multiplicity and uniqueness. In the economic literature, common knowledge of the fundamentals leads to the standard case of multiple equilibria due to the self-fulfilling nature of agents' beliefs. Morris and Shin [1998, 2000] and others propose that multiplicity vanishes once the economy/society is perturbed away from the perfect-information benchmark. We show that perturbation may or may not induce uniqueness in the context of a social network of discrete agents, depending on how the noisy signals are communicated, in other words depending on the topology of the social network.

References

G.-M. Angeletos and I. Werning. Crises and prices: Information aggregation, multiplicity, and volatility. *American Economic Review*, 96(5), December 2006.

G.-M. Angeletos, C. Hellwig, and A. Pavan. Dynamic global games of reime change: Learning, multiplicity, and timing of attacks. *Econometrica*, 75(3):711–756, May 2007.

C. Edmond. Information manipulation, coordination and regime change. working paper, NYU Stern School of Business, 2005.

I. Goldstein and A. Pauzner. Demand deposit contracts and the probability of bank runs. *Journal of Finance*, 60(3):1293–1328, 2005.

S. Morris and H. S. Shin. Unique equilibrium in a model of self-fulfilling currency attacks. *The American Economic Review*, 88(3):587–597, 1998.

S. Morris and H. S. Shin. Rethinking multiple equilibria in macroeconomics. *NBER Macroeconomics Annual*, 2000.

S. Morris and H. S. Shin. Coordination risk and the price of debt. *European Economic Review*, 48: 133–153, 2004.

A Differential Games Framework for Consensus in Social Networks: From Nash Equilibrium to Mean-Field Equilibrium*

Quanyan Zhu† and Tamer Başar†

†Coordinated Science Laboratory & Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 1308 W. Main St., Urbana, USA, 61820
E-mail: {zhu31, basar1}@illinois.edu

March 11, 2011

Abstract

In social networks, agents typically seek to achieve a task with some knowledge of their neighbors or immediate friends. Consensus is one of the fundamental and pivotal problems in social sciences involving a large number of distributed agents reaching consensus in their opinions, resources, security, etc. In this paper, we use a differential game-theoretic approach to model the dynamic interactions among a large number of consensus-seeking agents in social networks. Such an approach to consensus provides a theoretical basis for incentive mechanism design and for the construction of optimal defense strategies in an adversarial environment. In this framework, each agent aims to find a local optimal control to reach an agreement with its neighbors with minimum control effort. The model we adopt is one of linear-quadratic nonzero-sum differential games defined on an infinite horizon with discounted cost, which we study under different information structures and with a view to consensus. We characterize the open-loop (OL) and strongly time-consistent closed-loop (STC CL) Nash equilibrium (NE) strategies for finite population and large population regimes. For the finite population game, the STC CL NE strategy of each agent is affine in the states of its neighbors and consensus is achieved depending on the initial states of the agents. For a large homogeneous population, the STC CL NE requires the solution of a nonlinear PDE that describes the state evolution of the population, which is coupled with a set of coupled algebraic Riccati equations. We study the relationship between OL and STC CL as the population or the neighborhoods grow.

*Submitted to May 2011 MIT LIDS Workshop on Decision Making in Social Networks.

In the paper, we also inject into the network a population of malicious agents who selfishly force the social network to reach consensus at their own target value. We study the effect of such malicious agents on the consensus process and investigate the two-population interactions when the number of malicious agents is large. We propose a sub-optimal solution for a simplified analysis of consensus in an adversarial environment and a dynamic trust management mechanism for agents to defend themselves against malicious agents. Presentation of some simulation results wraps up the paper.

We provide below a list of selected relevant bibliography.

References

[1] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994

[2] H. Tembine, Q. Zhu, and T. Başar, "Risk-sensitive mean field stochastic games", in Proceedings of IFAC 2011, August 2011 (to appear).

[3] D. Bauso, L. Giarre and R. Presenti, "Consensus for networks with unknown but bounded disturbances," *SIAM Journal on Control and Optimization*, **48**(3):1756-1770, 2009.

[4] M. Huang, P.E. Caines, and R.P. Malhame. "Large-population cost-coupled LQG problems with non-uniform agents: individual-mass behavior and decentralized epsilon-Nash equilibria," *IEEE Transactions on Automatic Control*, **52**(9):1560-1571, Sept. 2007.

[5] J.-M. Lasry and P.-L. Lions, "Mean Field Games," *Japanese Journal of Mathematics*, **2**:229-260, 2007.

Iterative Learning from a Crowd

David R. Karger*, Sewoong Oh†, and Devavrat Shah‡
Department of EECS, Massachusetts Institute of Technology
Email: {karger*, swoh†, devavrat‡}@mit.edu

I. Introduction

Crowdsourcing systems, such as Amazon Mechanical Turk [Mec], establishes a marketplace where small tasks are distributed through an open call to a large and undefined group of people called a 'crowd'. A typical crowdsourcing begins with a requester broadcasting a large number of simple tasks to the crowd. The workers respond by submitting solutions to the tasks. The requester then verifies the solutions to make a decision whether to reject or approve each of the solutions. Workers are typically only rewarded for the solutions that are approved.

The kind of problems well-suited for crowdsourcing are problems which involve large data sets and can be easily broken down into a large number of small tasks. Examples include product categorization, document labeling, and image and video annotation. These simple tasks are routinely completed on general purpose crowdsourcing systems like Amazon Mechanical Turk [Mec] or Crowd Flower [Cro], often at lower prices than in-house or traditional outsourcing solutions. Further, these platforms can be integrated to perform specialized and more complex tasks such as transcription [Cas] and proofreading [BLM+10], [Soy].

Consider the following document labeling example [IPW10]. Each worker is given a set of web sites and has to decide if there is any adult content on each site. While some workers are diligent and give accurate labels to each sites, there might be spammers who give random labels. Given these labels with limited accuracy, a requester needs to achieve two goals. First, she needs to estimate the correct labels of the sites. If each site is assigned to a single worker, sites labeled by spammers have no chance of being corrected. Next, she needs to correctly identify who the spammers are and reward those that are diligent. Frequent incorrect identifications might lead to rejecting diligent workers and eventually ruin the reputation of the requester. However, oftentimes the cost of verifying a submitted solution is comparable to that of solving the task. A common solution to both of these challenges is to introduce redundancy: assigning each task to multiple workers and each worker to multiple tasks. Since any intervention from the requester is costly, we want a fully unsupervised and automated procedure to (i) design which tasks should be assigned to which workers; and (ii) infer the quality of workers and the solution to the tasks.

A naive approach to exploiting the redundancy is to use majority voting to identify the correct solutions. However, as we will see in the following sections, majority voting can be significantly improved upon. To fully exploit redundancy, we need to infer the quality of the workers simultaneously while inferring the solutions of the tasks. Dawid and Skene [DS79] proposed an iterative algorithm for inferring the solutions and quality of workers, based on expectation maximization. The algorithm first estimates the quality of the workers by comparing the submitted solutions to the estimated solutions. Then, the solutions are estimated based on the submitted solutions and the estimated quality of the workers. The algorithm iterates these two steps until convergence.

In the following, we first introduce a simple and effective model to describe the interactions between the tasks and the workers. Then, we propose a novel and efficient message passing algorithm to infer the solutions of the tasks, inspired by belief propagation. We show that when each task is assigned to a fixed number l of random workers (and each worker is assigned a fixed number r of random tasks), the probability of making an incorrect estimation decays exponentially in the redundancy: the number of workers assigned to each task. It proves that it is possible to iteratively infer the quality of the workers and the solution of the tasks to achieve performance significantly better than majority voting. Further, the computational complexity of the proposed algorithm is linear in the problem dimension m.

II. Model definition

As problems involving a large number of small tasks are well-suited for crowdsourcing, we model a crowdsourcing system as a set of m tasks t_i for $i \in \{1, \dots, m\}$ associated with 'correct' answers $s_i \in \{\pm 1\}$ and a set of n workers w_a for $a \in \{1, \dots, n\}$. We characterize the quality of each worker with a single parameter $p_a \in [0, 1]$. When task t_i is assigned to a worker w_a, the worker submits an answer $A_{ia} \in \{\pm 1\}$. We assume that the event that an answer is correct, $A_{ia} = s_i$, happens with probability p_a independent of any other event.

Let $[N] = \{1, \dots, N\}$ denote the set of first N integers. The submitted answers can be represented by a weighted bipartite graph $G\big(\{t_i\}_{i\in[m]}, \{w_a\}_{a\in[n]}, E, A\big)$. $E \subset [m]\times[n]$ is the set of edges where nodes t_i and w_a are connected if task t_i is assigned to a worker w_a. To each edge (i, a), we assign a weight A_{ia}. With a slight abuse of notations, we use

$A \in \{0, \pm 1\}^{m \times n}$ to denote the weighted adjacency matrix of the graph G.

We assign tasks according to a (l, r) random regular graph. Let the degree of a node denote the number of neighbors of the node in the graph. We wish to construct graphs with regular fixed degrees. Given a degree pair (l, r) such that $lm = rn$. Among all graph realizations with regular (l, r) degree, we choose one uniformly at random. This is called the *configuration model* in a random graph literature [RU08], [Bol01].

Throughout this paper, we use boldface characters to denote random variables and random matrices. We assume that the quality of workers, represented by random variables $\{\mathbf{p}_a\}$, are independent and identically distributed according to a probability measure μ on $[0, 1]$.

III. Algorithm

The algorithm operates on a set of messages $x_{i\to a}^{(k)}, y_{a\to i}^{(k)} \in \mathbb{R}$ associated with the edges in the bipartite graph $G\big(\{t_i\}_{i\in[m]}, \{w_a\}_{a\in[n]}, E, A\big)$. First, the messages $\{y_{a\to i}^{(0)}\}$ are initialized as independent copies of a Gaussian random variable with mean one and variance one. The algorithm is not sensitive to a specific initialization as long as the distribution has non-zero mean and is independent of the problem size m. At each iteration k the messages are updated according to the following rule:

$$\begin{aligned} x_{i\to a}^{(k)} &= \sum_{b\in\partial i\setminus a} A_{ib} y_{b\to i}^{(k-1)} \ , \\ y_{a\to i}^{(k)} &= \sum_{j\in\partial a\setminus i} A_{ja} x_{j\to a}^{(k)} \ , \end{aligned}$$

where ∂i and ∂a are the sets of neighbors of nodes t_i and w_a respectively, and $\partial i \setminus a$ is the set of all neighbors of node i excluding node a. After a pre-defined number of iterations k, an estimation of the correct s_i is made according to $\mathrm{sign}(\hat{x}_i^{(k)})$, where

$$\hat{x}_i^{(k)} = \sum_{b\in\partial i} A_{ib} y_{b\to i}^{(k-1)} \ .$$

When $\hat{x}_i^{(k)} = 0$ for some task t_i, we flip a fair coin to make a decision.

IV. Main results

When an edge (i, a) is chosen uniformly at random, the distribution of the messages $x_{i\to a}^{(k)}$ and $y_{a\to i}^{(k)}$ are characterized by the following evolution of random variables $\mathbf{x}^{(k)}$ and $\mathbf{y}_p^{(k)}$. Since we initialize the messages $\{y_{a\to i}^{(0)}\}$ using a Gaussian distribution, we initialize $\mathbf{y}^{(0)}$ with the same Gaussian distribution: $\mathbf{y}_p^{(0)} \sim \mathbf{N}(1, 1)$. Let $\stackrel{d}{=}$ denote that the random variables are equal in distribution. Then, for $k \in \{1, 2, \ldots\}$,

$$\mathbf{x}^{(k)} \stackrel{d}{=} \sum_{i=1}^{l-1} \mathbf{z}_{\mathbf{p}_i,i}^{(k)} \mathbf{y}_{\mathbf{p}_i,i}^{(k-1)} \ , \quad \mathbf{y}_p^{(k)} \stackrel{d}{=} \sum_{j=1}^{r-1} \mathbf{z}_{p,j}^{(k)} \mathbf{x}_j^{(k)} \ , \tag{1}$$

where $\mathbf{x}_j^{(k)}$ for $j \in \{1, \ldots, r-1\}$ are independent copies of $\mathbf{x}^{(k)}$, $\mathbf{p}_i$ for $i \in \{1, \ldots, l-1\}$ are independent copies of $\mathbf{p}$ which is distributed as according to μ, $\mathbf{y}_{p,i}^{(k)}$ for $i \in \{1, \ldots, l-1\}$ are independent copies of $\mathbf{y}_p^{(k)}$, and $\mathbf{z}_{p,i}^{(k)}$'s and $\mathbf{z}_{p,j}^{(k)}$'s are independent copies of $\mathbf{z}_p$. $\mathbf{z}_{\mathbf{p}_i,i}^{(k)}$ and $\mathbf{y}_{\mathbf{p}_i,i}^{(k-1)}$ are independent conditioned on $\mathbf{p}_i$, $\mathbf{z}_{p,j}^{(k)}$ and $\mathbf{x}_j^{(k)}$ are independent random variables, and

$$\mathbf{z}_p = \begin{cases} +1 & \text{with probability } p \ , \\ -1 & \text{with probability } 1-p \ . \end{cases}$$

For a task node t_i chosen uniformly at random, the decision variable $\hat{x}_i^{(k)}$ is characterized by

$$\hat{\mathbf{x}}^{(k)} \stackrel{d}{=} \sum_{i=1}^{l} \mathbf{z}_{\mathbf{p}_i,i}^{(k)} \mathbf{y}_{\mathbf{p}_i,i}^{(k-1)} \ . \tag{2}$$

To simplify the notations, let $\hat{l} \equiv l-1$, $\hat{r} \equiv r-1$, $q_0 \equiv \mathbb{E}[2\mathbf{p}-1]$, and $q \equiv \mathbb{E}[(2\mathbf{p}-1)^2]$. Define

$$\begin{aligned} \sigma_k^2 &\equiv 5\hat{l}^2(\hat{l}\hat{r})^{k-1} + {q_0}^2\hat{l}^3\hat{r}(4+q\hat{r})(q\hat{l}\hat{r})^{2k-2}\frac{1-(1/q^2\hat{l}\hat{r})^{k-1}}{(q^2\hat{l}\hat{r}-1)\hat{l}\hat{r}} \ , \\ &= \frac{{q_0}^2\hat{l}^3\hat{r}(4+q\hat{r})(q\hat{l}\hat{r})^{2k-2}}{(q^2\hat{l}\hat{r}-1)\hat{l}\hat{r}} + o\big((q\hat{l}\hat{r})^{2k-2}\big) \ , \end{aligned}$$

for $q^2\hat{l}\hat{r} > 1$. Then we can show the following bound on the probability of making an error.

Theorem IV.1. *Assume $\hat{l}\hat{r}q^2 \neq 1$. Then,*

$$\mathbb{P}\big(\hat{\mathbf{x}}^{(k)} < 0\big) \le \exp\left\{-\frac{1}{2}\frac{l^3{q_0}^2(\hat{l}\hat{r}q)^{2k-2}}{(l-1)\sigma_k^2}\right\} \ . \tag{3}$$

As the number of iterations k grows large, the above upper bound converges to a non-trivial limit when the degrees are large enough: $q^1\hat{l}\hat{r} > 1$.

Corollary IV.2. *For $\hat{l}\hat{r}q^2 > 1$,*

$$\lim_{k\to\infty} \mathbb{P}\big(\hat{\mathbf{x}}^{(k)} \le 0\big) \le \exp\left\{-\frac{1}{2}\frac{l^3(\hat{l}\hat{r}q^2-1)}{(l-1)^3(4+\hat{r}q)}\right\} \ . \tag{4}$$

References

[BLM+10] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, *Soylent: a word processor with a crowd inside*, Proceedings of the 23nd annual ACM symposium on User interface software and technology (New York, NY, USA), ACM UIST, 2010, pp. 313–322.

[Bol01] B. Bollobás, *Random Graphs*, Cambridge University Press, January 2001.

[Cas] *Casting Words*, http://castingwords.com/.

[Cro] *Crowd Flower*, http://crowdflower.com.

[DS79] A. P. Dawid and A. M. Skene, *Maximum likelihood estimation of observer error-rates using the em algorithm*, Journal of the Royal Statistical Society. Series C (Applied Statistics) **28** (1979), no. 1, 20–28.

[IPW10] P. G. Ipeirotis, F. Provost, and J. Wang, *Quality management on amazon mechanical turk*, Proceedings of the ACM SIGKDD Workshop on Human Computation (New York, NY, USA), HCOMP '10, ACM, 2010, pp. 64–67.

[Mec] *Amazon Mechanical Turk*, http://www.mturk.com.

[RU08] T. Richardson and R. Urbanke, *Modern Coding Theory*, Cambridge University Press, march 2008.

[Soy] *Soylent*, http://projects.csail.mit.edu/soylent/.

Asymptotic Learning on Social Networks

Elchanan Mossel*, Allan Sly† and Omer Tamuz‡

March 14, 2011

In his seminal Agreement Theorem, Aumann (1976) showed that two rational agents who repeatedly share their beliefs must converge to identical opinions and cannot "agree to disagree". This result was later extended to apply to a group of agents connected by a social network (see, e.g., Geanakoplos (1992)).

The eventual common belief depends on the different pieces of information that were initially available to the individuals. It is natural to ask how well this information is aggregated; how well do the agents learn from each other? We show that asymptotically learning is indeed efficient.

We consider a finite number of agents connected in a social network, with $S \in \{0,1\}$ a binary state of the world which the agents are interested in knowing. The agents are initially provided with noisy (but informative) independent signals regarding S. They proceed to iteratively share their belief regarding the state of the world, telling their neighbors in each iteration what they think the probability of the event $S = 1$ is, and thus learning from each other.

The Agreement Theorem and its extensions guarantee that the agents converge to a common belief. We show that as the size of the network tends to infinity this limiting common belief becomes more precise: given $\epsilon > 0$ and assuming $S = 1$, for large enough networks the agents will converge to a belief that assigns the event $S = 1$ a probability that is larger than $1 - \epsilon$.

Following Gale & Kariv (2003) we additionally consider a more restricted communication model. Here each agent must in each iteration choose an action in $\{0,1\}$, where the optimal action equals the unknown state of the world. The agents learn by observing their neighbors' actions, and so gain far less information regarding each others' beliefs than in the previous model. Here too we show that under weak conditions, for large networks, the agents eventually take the correct action with probability arbitrarily close to one. This is in contrast to other results (on slightly different models) which exhibit "herd behavior" or "information cascades" (e.g., Banerjee (1992), Bikhchandani, Hirshleifer & Welch (1992)), where it is possible that the wrong action is taken with probability bounded away from zero.

*UC Berkeley and the Weizmann Institute of Science. Supported by a Sloan fellowship in Mathematics, NSF awards DMS 0528488 and DMS 0548249 (CAREER), and ONR grant N0014-07-1-05-06.

†Microsoft Research.

‡Weizmann Institute of Science. Supported by ISF grant 1300/08.

Opinion fluctuations and persistent disagreement in social networks

Daron Acemoglu,*Giacomo Como †, Fabio Fagnani‡, and Asuman Ozdaglar §

March 11, 2011

Disagreement among individuals in a society, even on central questions that have been debated for centuries, is the rule; agreement is the rare exception. How can disagreement of this sort persist for so long? Most existing models of communication and learning, based on Bayesian or non-Bayesian updating mechanisms, typically lead to consensus provided that communication takes place over a strongly connected network. (See, e.g., Smith and Sorensen '00, Banerjee and Fudenberg '04, Acemoglu et al. '10, Bala and Goyal '98, Gale and Kariv '03, De Marzo et al. '03, Golub and Jackson '10, Acemoglu et al. '11) These models are thus unable to explain persistent disagreements, and belief fluctuations.

In this work, we propose a tractable model that generates long-run disagreements and persistent opinion fluctuations. Our model involves a stochastic gossip model of continuous opinion dynamics in a society consisting of two types of agents: *regular agents*, who update their beliefs according to information that they receive from their social neighbors; and *stubborn agents*, who never update their opinions and might represent leaders, political parties or media sources attempting to influence the beliefs in the rest of the society. When the society contains stubborn agents with different opinions, the belief dynamics never lead to a consensus (among the regular agents). Instead, beliefs in the society almost surely fail to converge, the belief profile keeps on oscillating in an ergodic fashion, and it converges in law to a non-degenerate random vector.

*Economics Department, Massachusetts Institute of Technology, daron@mit.edu.

†Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, giacomo@mit.edu

‡Dipartimento di Matematica, Politecnico di Torino, Italy, fabio.fagnani@polito.it

§Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, asuman@mit.edu

controlled matrices, indexed by the vector $u := [u_1, \cdots, u_d]^T$. The per step reward is $\sum_{i=1}^{m} x_k(i) := \mathbf{1}^T x_k$ where $\mathbf{1}$ is the vector of all 1's, and the per step cost is $\sum_{i=1}^{d-m} c_2(i) f(\nu_k(i)) + \sum_{i=1}^{m} c_3(i) g(u_k(i))$ for prescribed f, g. Let $f(\nu) := [f(\nu_1), \cdots, f(\nu_d)]^T$ for $\nu := [\nu_1, \cdots, \nu_d]^T$, where $\nu_j = 0$ for $j \leq m$ by convention. Define $u, g(u)$ analogously, with $u_i =$ a prescribed u_0 for $i > m$. We consider the stationary problem:

Maximize over (u, ν) the quantity

$$\text{Max}_{(u,\nu)} \quad \mathbf{1}^T x - c_2^T f(\nu) - c_3^T g(u)$$

subject to the constraints

$$x = P^u x + \bar{P}^u \nu.$$

We map this problem to a somewhat non-classical Markov decision process and use the techniques of Markov decision theory to propose algorithms for its resolution.

More generally, one can consider the transition matrix $\tilde{P}$ indexed by two controls, say $\tilde{P}^{u,v} := [[p(j|i, u_i, v_i)]]$, where the v_is are controlled by an adversary. This maps to a stochastic game problem which can also be addressed in a similar manner.

OPINION FORMATION UNDER PEER PRESSURE

Vivek S. Borkar

School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai 400005
email: borkar.vs@gmail.com

Aditya Karnik

Enterprise Analytics Group, General Motors R&D India Science Lab., Bangalore 560 066
email: aditya.karnik@gm.com

'I get by with a little help from my friends.'

We consider a stylized model of opinion formation in which an individual agent pursues her interest, but also takes into account what the peers say. Specifically, the agent in question, say the ith out of d of them, holds an 'opinion' $x_k(i) \in \mathcal{R}$ at time instant k and polls a peer j with probability $p(i,j)$, this being the (i,j)th element of a stochastic matrix P. Let $\xi_k(i)$ denote the (random) identity of the peer who has been polled. She then updates her opinion incrementally according to

$$x_{k+1}(i) = x_k(i) + \gamma[\alpha_i(\sum_j I\{\xi_k(i) = j\}x_k(j) - x_k(i)) + (1-\alpha_i)f_i(x_k(i))], \quad (1)$$

where $0 < \alpha_i < 1$ is the weight she attaches to 'peer pressure' while attaching weight $(1-\alpha_i)$ to her own 'inclination' $f_i(x_k(i))$, where the f_is are bounded continuous. As an example of the latter, consider, e.g., $f_i := \nabla g$ where g represents a common 'payoff landscape' the agents share. Here $\gamma > 0$ is a stepsize ensuring the incremental nature of the learning process. We consider the case $f_i = f \ \forall i$. Let $F(x_1, \cdots, x_d) := [f(x_1), \cdots, f(x_d)]^T$, $A :=$

diag$(\alpha_1, \cdots, \alpha_d)$. Then one can view (1) as a constant stepsize stochastic approximation algorithm with the o.d.e. limit

$$\dot{x}(t) = A(P - I)x(t) + (I - A)F(x(t)). \tag{2}$$

Using the Hirsch theorem for cooperative o.d.e.s, one can show that this generically converges. We consider the case where the scalar o.d.e.

$$\dot{z}(t) = f(z(t)) \tag{3}$$

converges to one of finitely many equilibria for any initial condition. If $x^* \in \mathcal{R}$ is one such equilibrium, then $[x^*, \cdots, x^*]^T$ is an equilibrium for (2). The latter, however, can also have other, 'mixed' equilibria. Assuming that P is irreducible, one necessary condition for any $\hat{x} \in \mathcal{R}^d$ to be an equilibrium of (2) is that

$$\sum_i \pi_i f(\hat{x}(i)) = 0, \tag{4}$$

where $\pi = [\pi_1, \cdots, \pi_d]^T$ is the unique stationary distribution under P. We investigate the role of α_is in bringing about consensus or disagreement of opinions.

We also consider the case of 'opinion manipulation' where some, say $m < d$ agents fix their opinions to some prescribed equilibrium x^* of (3) for all k. Let $\tilde{x}_k$ denote the opinions of the remaining agents. One can then show that they *cannot* converge to $[x', \cdots, x']^T$ for some equilibrium x' of (3) other than x^*, whereas if α_is are close enough to 1 (i.e., everyone succumbs to peer pressure), $[x^*, \cdots, x^*]^T$ is the only equilibrium, i.e., a consensus on the desired opinion is obtained. As α_is decrease, one can get mixed equilibria. The situation is even more complex if we replace the scalar valued f by a vector valued function. We explore this phenomenon through numerical experiments.

Discovery and Security in Social Network Models: Graph-Theoretic Characterizations

Sandip Roy
Associate Professor
School of Electrical Engineering and Computer Science
Washington State University
sroy@eecs.wsu.edu

Mengran Xue
Graduate Research Assistant
School of Electrical Engineering and Computer Science
Washington State University

Abstract

The incredible inter-personal connectivity offered by modern social networking tools facilitates a fundamental human need for interaction, and in doing so can enact a profound impact on our professional, political, and social lives. At the same time, however, the essential connectivity afforded by these tools by its very nature impacts our privacy and re-defines our notions of trust. The consequent grave concerns regarding information security in modern social networks badly require study, and in fact both a philosophical and technical literature on security of social networking capabilities has developed recently. While much of the technical research has focused on allowing users to define trust and protect information in particular social network tools, our viewpoint is that the extent of possible information discovery (and, conversely, information security) in a social network is an intrinsic consequence of the network's connectivity. We thus believe that fundamental relationships between the *graph topology* of a social network and the discovery/security of information can be found.

In this talk, we will draw on several recent graph-theoretic characterizations of estimator performance in complex dynamical networks (developed by our group and others), to obtain graph-theoretic characterizations of security and discovery in abstract models for social network interactions. We will focus on two simple models for social interactions: 1) a classical linear model describing distributed consensus among networked agents, in which agents update their personal opinions based on interactions with neighbors as well as multiple local storage variables (memory variables) so as to quickly reach a fair consensus; 2) a stochastic automaton model known as the influence model, that has been used to represent such diverse phenomena as human conversation patterns, discrete-valued consensus algoriths, idea propagation, and self-grouping of network agents. For both models, we will imagine an observer—whether an adversary or a benign player—as accessing noise-corrupted state statistics of particular agents in the network. We will explore how the the graph-topology of the network, together with specifics of local dynamics and the observer's sensing capability, impact the observer's ability to estimate the full network state or important statistics defined thereof (for instance, initial

opinions of certain important hidden network components or agents).

The graph-theoretic characterizations of security and discovery for social networks that we will present derive from recent graphical analyses of dynamical-network estimation. These graphical analyses of network estimation, which were originally motivated by sensor-design needs in infrastructure networks, are based on combining classical estimation theory with algebraic graph theory constructs. Precisely, the graphical analysis is initiated from classical algebraic expressions of state/topology estimators and of the estimation error. With some effort, these algebraic expressions can be phrased in terms of the spectrum of a matrix representing the graph topology, together with the specifics of the observation paradigm and network-component model. In turn, algebraic graph theory constructs can be used to translate the spectral conditions to explicit graph-theoretic characterizations of estimator structure and performance, and to compare estimator performance for different graph classes (e.g., random graphs vs. small-world or coherently-structured graphs). We will apply these graphical analyses to understand the connection between social-network connectivity and information discovery/security, in particular elucidating the role of 1) graph coherency structures, 2) connection density, and 3) observation locations in information discovery and security.

We kindly ask the reader to see the author's web page, www.eecs.wsu.edu/˜sroy, for a list of publications and foundational project work related to network estimation.

On the Unpredictability of Elections using Social Media Data

Daniel Gayo-Avello, Panagiotis T. Metaxas, Eni Mustafaraj

Departamento of Informatica, Universidad de Oviedo and Department of Computer Science, Wellesley College
dani@uniovi.es, pmetaxas@wellesley.edu, emustafa@wellesley.edu

Predictions using data from social networks and Web searches have attracted a lot of attention lately. By following what people are blogging about or what they are searching about can give us some intuition on the collective psyche and lead us to understand what is currently happening in society before it has actually happened. Sometimes people refer to this phenomenon as the "wisdom of the crowd", that is, taking into account the opinions of the society as a whole, instead of the opinions of the expert.

Facebook and Twitter are services where users can share personal opinions, and the APIs provided by these companies make it relatively easy to extract and process them. Google search volume trends, also accessible by APIs, can give an indication on what people are currently searching for, thus providing some insight into their current needs and worries. Theoretically, these data, if used correctly, can lead to predictions of currently occurring events influenced by human behavior. In fact, Choi & Varian (2009) have coined the term "predicting the present" to describe this phenomenon, while Asur & Huberman (2010) simply refer to it as "predicting the future".

Being able to make predictions based on publicly available data would have numerous benefits to areas such as health (e.g. predictions of flu epidemics –Ginsberg *et al.* 2009, Lampos *et al.* 2010), business (e.g., prediction of box-office success of movies –Asur & Huberman 2010, Mishne & Glance 2006; and product marketability – Shimshoni *et al.* 2009), economics (e.g., predictions on stock market trends and housing market trends –Bollen *et al.* 2010, Choi & Varian 2009, or Gilbert & Karahalios 2010), and politics (e.g., trends in public opinion – O'Connor *et al.* 2010; and predictions of election results – Geek Blog 2010, Tumasjan *et al.* 2010, or Tweetminster 2010).

One would expect that, following the previous research literature (e.g. O'Connor *et al.* 2010; Tumasjan *et al.* 2010), and given the high utilization that the Web and online social networks have in the US (Smith 2011), Twitter volume should have been able to predict consistently the outcomes of the US Congressional elections. But is it so? In this presentation we examine the instances and methods that have been used in the past in the claims of electoral results predictions and discuss their predictive power. We then argue that Social Media cannot predict elections and we will give a range of reasons inherent in the use of Social Media that undercut the predictability of elections.

Claims that Social Media Data could have predicted the elections

Because of the promising results achieved by many of the projects and studies mentioned in the previous section there is a relatively high amount of hype surrounding the feasibility of predicting electoral results using social media. It must be noted that most of that hype is fueled by traditional media and blogs, usually bursting prior and after electoral events. For example, shortly after the recent 2010 elections in the US, bold statements made it to the news media headlines. From those arguing that Twitter is not a reliable predictor (e.g. Goldstein & Rainey 2010) to those claiming just the opposite, that Facebook and Twitter were remarkably accurate (e.g. Carr 2010).

We point out that all of such statements were issued after the elections were over and the final results were disclosed. Moreover, the degree of accuracy of these "predictions" was usually assessed in terms of percentage of correctly guessed electoral races – e.g., the winners of 74% for the House and 81% for the Senate races were predicted, according to Facebook (2010) – without further qualification. This is of vital importance since many races were won by very tight margins. They were also not always compared against traditional ways of prediction, such as the professional polling results or the simple predictions based on "incumbency" (the fact that those who are already in office are far more likely to be re-elected in the US).

Though not as bold as the news reports, scholarly research does tend to support a positive opinion on the predictive power of social media as a promising line of research, while exposing some of the caveats of the methods. Thus, according to Williams & Gulati (2008), the number of Facebook fans for election candidates had a measurable influence on their respective vote shares. These researchers assert that *"social network support, on Facebook specifically, constitutes an indicator of candidate viability of significant importance [...] for both the general electorate and even more so for the youngest age demographic."*

A study of a different kind was conducted by O'Connor *et al.* (2010). They analyzed the way in which simple sentiment analysis methods could be applied to tweets as a tool of automatically pulsing public opinion. These researchers correlated the output of such a tool with the temporal evolution of different indices such as the index of Consumer Sentiment, the index of Presidential Job Approval, and several pre-electoral polls for the US 2008 Presidential Race. The correlation with the first two indices was rather high but it was not significant for the pre-electoral polls. According to the findings of this paper, sentiment analysis on Twitter data seems to be a promising field of research to replace traditional polls although, in their words, it's not there quite yet.

Finally, the work by Tumasjan *et al.* (2010) focused on predicting elections from social media. Indeed, one of the research questions their study aimed to answer was whether Twitter can serve as a prediction of electoral results. In that paper, a strong statement is made about predictability, namely that *"the mere number of tweets mentioning a political party can be considered a plausible reflection of the vote share and its predictive power even comes close to traditional election polls."* Moreover, these researchers found that co-occurrence of political party mentions accurately reflected close political positions between political parties and plausible coalitions.

Unpredictability of Elections using SM

In a recent study, we examined how the algorithms that have claimed to predict elections would have performed in several instances in the 2010 US elections. This was important to establish because a wider set of test cases was needed to base any claims of predictability of elections through Social Media.

For our study, two data sets related to elections that took place in the US during 2010 were collected. Predictions were calculated based on Twitter chatter volume, as in (Tumasjan *et al.* 2010), and then based on sentiment analysis of tweets, in a way similar to (O'Connor *et al.* 2010). These predictions were compared to the actual results of the elections. Out of six senatorial races, each method predicted correctly the win/lose outcome of only three. The predictions based on Twitter volume had a mean average error (MAE) of 17.1%, while the predictions based on sentiment analysis had a MAE of 7.6%.

To examine whether sentiment-based analysis actually performed better, we deepened our analysis in this direction. Without getting into details, we report that based on three experiments, we found that the accuracy of lexicon-based sentiment analysis when applied to political conversation is quite poor. When compared against manually labeled tweets it seems to just slightly outperform a random classifier; it fails to detect and correctly assign the intent behind disinformation and misleading propaganda; and, finally, it's a far cry from being able to predict political preference.

Predicting elections with accuracy should not be a matter of luck or post-processing adjustment. It should not be supported without having some clear understanding why it works. Instead, it should be a matter of correctly identifying likely voters and getting an un-biased representative sample of them. That's what professional pollsters have been doing for the last 80 years, with mostly impressive results. But that's something that today's Social Media cannot do. Let us examine why.

To make our point clear, two pieces of evidence will be provided in our presentation. First, we will describe the complexity of professional polling and explain the reasons why their methods cannot be duplicated by **sampling** Social Media data. Next, we will discuss the **manipulation** of Social Media by spammers and propagandists, since they can shape the data so that they do not reflect the true intentions of the users, even if they happened to be representative of the whole population.

Our research has revealed that data from social media did only slightly better than chance in predicting election results in the last US congressional elections. We argue that this makes complete sense: So far, only a very rough estimation on the exact demographics of the people discussing elections in social media is known, while according to the state-of-the-art polling techniques, correct predictions requires the ability of sampling likely voters randomly and without bias. Moreover, answers to several pertinent questions are needed such as the actual nature of political conversation in social media, the relation between political conversation and electoral outcomes, and the way in which different ideological groups and activists engage and influence online social networks.

In addition to that, further research is needed regarding the flaws of simple sentiment analysis methods when applied to political conversation. In this sense it would be very interesting to understand the impact of different lexicons and, even more important, to go one step farther by using machine learning (such as in the work by Asur & Huberman 2010); or looking for a deeper understanding of the dynamics of political conversation in social media following the work of Somasundaran & Wiebe (2010).

Acknowledgements

The Twitter data for the November election was courtesy of the University of Indiana. This work was partially supported by a Brachman-Hoffman grant.

References

Asur, S.; Huberman, B.A. 2010. *Predicting the Future With Social Media* (Technical Report). HP Labs, Palo Alto.

Blumenthal, M. 2004. *The Why and How of likely Voters* (Blog Post). Appeared online at: http://www.mysterypollster.com/main/2004/10/the_why_how_of_.html

Bollen, J.; Mao, H.; Zeng, X.J. 2010. *Twitter mood predicts the stock market* (Technical Report).

Carr, A. 2010. Facebook, Twitter Election Results Prove Remarkably Accurate. *Fast Company*. Appeared online at: http://www.fastcompany.com/1699853/facebook-twitter-election-results-prove-remarkably-accurate

Choi, H.; Varian, H. 2009. *Predicting the Present with Google Trends* (Technical Report). Google Inc.

Cohen, J. 1988. Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates, Inc.

Esuli, A.; Sebastiani, F. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation.*

Facebook, 2010. *Snapshot: The Day After Election Day* (Press Release). Appeared online at: http://www.facebook.com/notes/us-politics-on-facebook/snapshot-the-day-after-election-day/448930025881

Gayo-Avello, D. 2011. A warning against converting Social Media into the next Literary Digest. To appear in *Communications of the ACM*.

Geek Blog, The. 2010. *Twitter Analysis - Belgian 2010 Elections: Party with most Twitter coverage also wins elections* (Blog Post). Appeared online at: http://geekblog.eyeforit.be/component/content/article/18-news/20-twitter-analysis-belgian-2010-elections-party-with-most-twitter-coverage-also-wins-elections.html

Gilbert, E., Karahalios, K. 2010. Widespread Worry and the Stock Market. In *Proceedings of ICWSM'2010.*

Ginsberg, J.; Mohebbi, M.H.; Patel, R.S.; Brammer, L.; Smolinski, M.S.; Brillian, L. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457(7232):1012–1014.

Golbeck, J.; Hansen, D. 2010. *Computing Political Preference among Twitter Followers* (Technical Report).

Goldstein, P.; Rainey, J. 2010. The 2010 elections: *Twitter isn't a very reliable prediction tool* (Blog Post). Appeared online at: http://latimesblogs.latimes.com/the_big_picture/2010/11/the-2010-midterms-twitter-effect-not-a-very-reliable-election-prediction-tool.html

Lampos, V.; De Bie, T.; Cristianini, N. 2010. Flu detector - Tracking epidemics on Twitter. In *Proceedings of ECML PKDD 2010.*

Lui, C.; Metaxas, P.T., Mustafaraj, E. 2011. On the Predictability of the U.S. Elections through Search Volume Activity. To appear in *Proceedings of e-Society Conference 2011.*

Mishne, G.; Glance, N. 2006. Predicting movie sales fromblogger sentiment. In *Procedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs.*

Mustafaraj, E.; Metaxas, P. 2010. From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line.* http://journal.webscience.org/317/

O'Connor, B.; Balasubramanyan, R.; Routledge, B.R.; Smith, N.A. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of ICWSM'2010.*

Shimshoni, Y.; Efron, N.; Matias, Y. 2009. *On the predictability of Search Trends* (Technical Report). Google Israel Labs.

Smith, A. 2011. *Twitter and Social Networking in the 2010 Midterm Elections.* Pew Internet and American Life publication available online at: http://pewresearch.org/pubs/1871/internet-politics-facebook-twitter-2010-midterm-elections-campaign

Somasundaran, S.; Wiebe, J. 2010. Recognizing Stances in Ideological On-Line Debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.*

Tumasjan, A.; Sprenger, T.O.; Sandner, P.G.; Welpe, I.M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of ICWSM'2010.*

Tweetminster. 2010. *Is word-of-mouth correlated to General Election results?* (Technical Report). Appeared online at: http://www.scribd.com/doc/31208748/Tweetminster-Predicts-Findings

Williams, C.B.; Gulati, G.J. 2008. The Political Impact of Facebook: Evidence from the 2006 Midterm Elections and 2008 Nomination Contest. *Politics & Technology Review* 1:11–21.

Wilson, T.; Wiebe, J.; Hoffman, P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Capturing unobserved correlated effects in diffusion in large virtual networks: distinguishing individual preferences, social connections and cultural discourse influence on the adoption of Twitter clients

Elenna R. Dugundji*, Ate Poorthuis, Michiel van Meeteren

Universiteit van Amsterdam
Department of Geography, Planning and International Development Studies
Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, Netherlands
e.r.dugundji@uva.nl, atepoorthuis@gmail.com, michielvanmeeteren@gmail.com

With the onset of Internet and phone-based technology, people leave numerous traces of their social behavior in – often publicly available – data sets. In this paper we look at a virtual community of independent ('Indie') software developers for the Macintosh and iPhone that use the social networking site Twitter as one of their platforms of choice. Using Twitter's publicly available API, we collect longitudinal data on both network connections and the use of Twitter client software over a period of five weeks. We use this data of the virtual community of 'Indie' developers to analyze the adoption of Twitter client software. (See Fig 1 and Fig 2.) Within this community, four prominent software developers have developed Twitter clients that compete for adoption by users in the same community. Apart from these 'Indie' Twitter clients, members of the virtual community can choose from a range of clients that are developed outside of the Indie community. Generally, social networks and social capital are considered to be important variables in explaining the adoption and diffusion of behavior. However, it is contested whether the actual social connections, cultural discourse, or individual preferences determine this adoption and diffusion. Using discrete choice analysis applied to longitudinal data, we are able to distinguish between social network influence on one hand and cultural discourse and individual preferences on the other hand. Our analysis shows that, within the virtual community, social connections are generally of greater importance for the adoption and diffusion of Indie clients than they are for the adoption of clients that are developed outside of the Indie community. In addition, we present a method using readily available software to estimate the size of the error due to unobserved correlated effects. This is critically important to test for in any application of multinomial logistic regression where social influence variables and/or other network measures are used as explanatory variables, since their use poses a classic case of endogeneity. We show that even in a seemingly saturated model, the log likelihood can increase dramatically by accounting for unobserved correlated effects. Furthermore the estimated coefficients in the uncorrected model can be significantly biased beyond standard error margins. (See Table 1.) Failing to correct for unobserved correlated effects can yield potentially highly misleading policy interpretations.

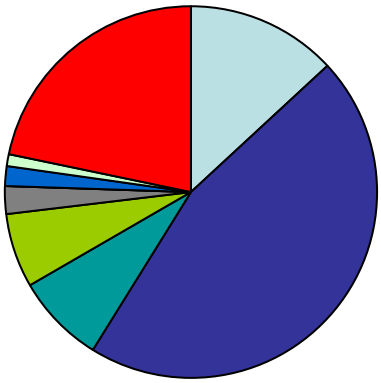


Fig 1. Twitter Client Market Share in Indie Community, 630633 Tweets

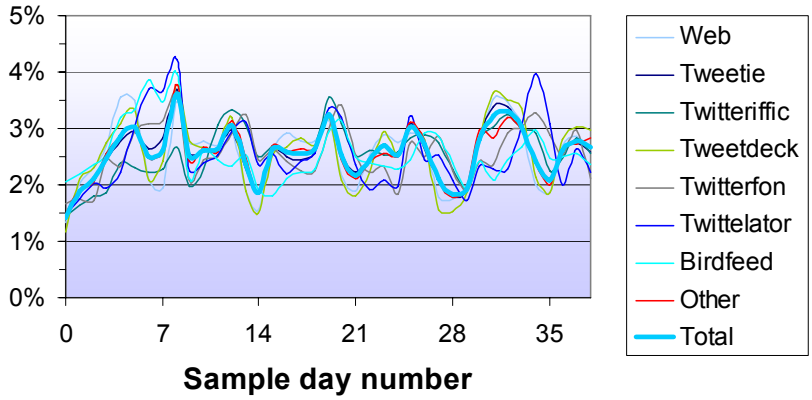


Fig 2. Norm. Tweets per Day per Twitter Client in Indie Community
Obs. period: Sunday 9 August to Wednesday 16 September 2009

* Corresponding author

TABLE 1. Estimated Parameters for MNL Model versus Mixed MNLP Model with Correlated Effects

		Multinomial Logit				Mixed Multinomial Logit for Panel data								
										Robust	Robust t	Robust	Absolute	Bias > SE
Nr.	Twitter client	Value	Std err	t-test	p-val	Value	Std err	t-test	p-val	Std err	test	p-value	Bias	MNL
	Alternative specific constants: "cultural discourse" on Twitter client													
1	Web	0,14	0,07	2,01	0,04	2,55	0,12	21,24	0	0,35	7,22	0	2,41	Yes
2	Tweetie	-2,06	0,07	-28,98	0	-5,69	0,23	-24,31	0	1,98	-2,87	0	3,63	Yes
3	Twitterrific	-3,22	0,10	-31,66	0	-5,67	0,18	-32,22	0	0,74	-7,67	0	2,45	Yes
4	Tweetdeck	-1,10	0,10	-10,62	0	0,07	0,19	0,37	0,71 *	0,78	0,09	0,93 *	1,17	Yes
5	Twitterfon	-2,65	0,16	-16,86	0	-1,85	0,22	-8,29	0	0,92	-2,02	0,04	0,80	Yes
6	Twittelator	-5,43	0,22	-25,10	0	-6,01	0,26	-23,25	0	0,89	-6,74	0	0,58	Yes
7	Birdfeed	-8,51	0,23	-36,35	0	-11,4	0,31	-36,30	0	1,10	-10,35	0	2,89	Yes
	Individual preferences (7-day cumulative lag): "stickiness" of Twitter client													
8	Web	3,40	0,013	256,8	0	1,36	0,022	61,65	0	0,070	19,38	0	2,04	Yes
9	Tweetie	3,59	0,010	358,0	0	1,35	0,017	79,71	0	0,053	25,39	0	2,24	Yes
10	Twitterrific	4,75	0,018	259,7	0	2,30	0,028	81,59	0	0,077	29,73	0	2,45	Yes
11	Tweetdeck	5,26	0,020	260,6	0	2,67	0,036	73,43	0	0,150	17,82	0	2,59	Yes
12	Twitterfon	6,04	0,030	200,7	0	4,55	0,055	82,29	0	0,280	16,24	0	1,49	Yes
13	Twittelator	7,05	0,042	166,9	0	6,06	0,064	94,37	0	0,293	20,67	0	0,99	Yes
14	Birdfeed	6,37	0,047	135,2	0	4,40	0,068	64,70	0	0,258	17,02	0	1,97	Yes
	Network influence of choice behavior of user's Friends in community (7-day cumulative lag): "viralness" of Twitter client													
15	Web	0,134	0,022	6,23	0	0,130	0,034	3,84	0	0,091	1,42	0,16 *	0,004	No
16	Tweetie	-0,296	0,015	-19,91	0	-0,133	0,022	-6,05	0	0,059	-2,27	0,02	0,163	Yes
17	Twitterrific	0,608	0,029	20,74	0	0,816	0,043	19,16	0	0,123	6,66	0	0,208	Yes
18	Tweetdeck	0,429	0,039	11,12	0	0,817	0,058	14,09	0	0,197	4,15	0	0,388	Yes
19	Twitterfon	0,411	0,097	4,23	0	1,120	0,149	7,49	0	0,948	1,18	0,24 *	0,709	Yes
20	Twittelator	1,240	0,138	8,97	0	1,240	0,185	6,69	0	0,607	2,04	0,04	0,000	No
21	Birdfeed	2,310	0,113	20,46	0	2,850	0,142	20,09	0	0,376	7,58	0	0,540	Yes
	User "follows" Opinion-maker John Gruber: contextual effect on Twitter client choice													
22	Web	0,074	0,011	6,54	0	-0,114	0,022	-5,24	0	0,084	-1,36	0,18 *	0,188	Yes
23	Tweetie	0,177	0,009	19,70	0	0,333	0,031	10,88	0	0,259	1,29	0,2 *	0,156	Yes
24	Twitterrific	0,234	0,015	15,41	0	0,757	0,032	23,47	0	0,164	4,63	0	0,523	Yes
25	Tweetdeck	-0,068	0,018	-3,82	0	-0,445	0,034	-13,07	0	0,177	-2,51	0,01	0,377	Yes
26	Twitterfon	-0,056	0,025	-2,24	0,03	-0,134	0,036	-3,75	0	0,155	-0,86	0,39 *	0,078	Yes
27	Twittelator	-0,071	0,038	-1,87	0,06 *	-0,268	0,047	-5,70	0	0,138	-1,94	0,05 *	0,197	Yes
28	Birdfeed	0,386	0,034	11,31	0	0,360	0,046	7,79	0	0,173	2,08	0,04	0,026	No
	User "follows" Twitter client developer: contextual effect on Twitter client choice													
29	Tweetie	0,237	0,010	23,71	0	0,795	0,036	21,96	0	0,304	2,61	0,01	0,558	Yes
30	Twitterrific	0,142	0,020	7,12	0	0,743	0,043	17,45	0	0,174	4,26	0	0,601	Yes
31	Tweetdeck	0,300	0,054	5,59	0	1,390	0,104	13,36	0	0,465	2,99	0	1,090	Yes
32	Twitterfon	0,532	0,116	4,59	0	1,620	0,256	6,34	0	2,200	0,74	0,46 *	1,088	Yes
33	Twittelator	0,479	0,213	2,25	0,02	1,640	0,230	7,15	0	0,405	4,05	0	1,161	Yes
34	Birdfeed	0,296	0,045	6,64	0	0,721	0,060	11,99	0	0,203	3,56	0	0,425	Yes
	Frequency of tweets sent by user during observation period: "power" user effect on Twitter client choice													
35	Web	-0,1210	0,0017	-69,81	0	-0,1370	0,0032	-42,35	0	0,0114	-12,07	0	0,0160	Yes
36	Tweetie	-0,1050	0,0014	-72,84	0	-0,0806	0,0040	-20,09	0	0,0285	-2,83	0	0,0244	Yes
37	Twitterrific	-0,0990	0,0025	-39,26	0	-0,0504	0,0044	-11,45	0	0,0182	-2,77	0,01	0,0486	Yes
38	Tweetdeck	-0,1010	0,0028	-35,99	0	-0,0714	0,0049	-14,63	0	0,0174	-4,10	0	0,0296	Yes
39	Twitterfon	-0,0570	0,0040	-14,14	0	-0,0489	0,0052	-9,40	0	0,0146	-3,36	0	0,0081	Yes
40	Twittelator	-0,0624	0,0056	-11,21	0	-0,0792	0,0067	-11,77	0	0,0181	-4,38	0	0,0168	Yes
41	Birdfeed	-0,0357	0,0057	-6,23	0	-0,0087	0,0075	-1,16	0,25 *	0,0234	-0,37	0,71 *	0,0270	Yes
	User eigenvector centrality in community: centrality effect on Twitter client choice													
42	Web	1,21	0,12	9,89	0	2,77	0,20	14,18	0	0,55	5,02	0	1,56	Yes
43	Tweetie	-0,91	0,10	-8,81	0	-2,70	0,27	-9,90	0	2,02	-1,34	0,18 *	1,79	Yes
44	Twitterrific	-0,86	0,16	-5,34	0	-4,66	0,26	-18,29	0	1,06	-4,39	0	3,80	Yes
45	Tweetdeck	1,79	0,21	8,58	0	4,17	0,33	12,65	0	1,26	3,30	0	2,38	Yes
46	Twitterfon	-1,24	0,31	-3,96	0	-2,57	0,42	-6,14	0	1,50	-1,71	0,09 *	1,33	Yes
47	Twittelator	-0,30	0,43	-0,71	0,48 *	0,73	0,54	1,35	0,18 *	1,72	0,42	0,67 *	1,03	Yes
48	Birdfeed	-2,31	0,29	-7,96	0	-3,68	0,36	-10,24	0	1,12	-3,27	0	1,37	Yes
	User "closeness" in community: centrality effect on Twitter client choice													
49	Web	-3,04	0,19	-15,71	0	-8,74	0,34	-25,71	0	1,00	-8,73	0	5,70	Yes
50	Tweetie	3,58	0,21	16,99	0	13,20	0,70	18,85	0	6,00	2,20	0,03	9,62	Yes
51	Twitterrific	2,27	0,29	7,95	0	5,70	0,47	12,04	0	1,92	2,97	0	3,43	Yes
52	Tweetdeck	-2,20	0,29	-7,48	0	-6,52	0,56	-11,75	0	2,25	-2,90	0	4,32	Yes
53	Twitterfon	-1,05	0,46	-2,31	0,02	-4,58	0,65	-6,99	0	2,70	-1,69	0,09 *	3,53	Yes
54	Twittelator	4,79	0,61	7,89	0	4,91	0,73	6,77	0	2,54	1,93	0,05 *	0,12	No
55	Birdfeed	9,89	0,62	15,98	0	12,50	0,81	15,33	0	2,85	4,38	0	2,61	Yes
	Ratio of user's Friends in community to user's total Friends in Twitter universe: extended "in-degree" (w.r.t. Tweet flow) effect on Twitter client choice													
56	Web	0,365	0,052	6,98	0	0,391	0,100	3,91	0	0,349	1,12	0,26 *	0,026	No
57	Tweetie	0,766	0,043	17,94	0	1,220	0,124	9,82	0	0,930	1,31	0,19 *	0,454	Yes
58	Twitterrific	1,360	0,068	19,97	0	2,620	0,117	22,43	0	0,383	6,86	0	1,260	Yes
59	Tweetdeck	-1,310	0,094	-14,00	0	-4,050	0,188	-21,51	0	0,875	-4,63	0	2,740	Yes
60	Twitterfon	-0,027	0,121	-0,22	0,83 *	0,104	0,183	0,57	0,57 *	0,658	0,16	0,87 *	0,131	Yes
61	Twittelator	-0,728	0,187	-3,90	0	-0,515	0,233	-2,21	0,03	0,648	-0,80	0,43 *	0,213	Yes
62	Birdfeed	1,630	0,148	10,98	0	3,520	0,200	17,58	0	0,667	5,27	0	1,890	Yes
	Ratio of user's Followers in community to user's total Followers in Twitter universe: extended "out-degree" effect on Twitter client choice													
63	Web	0,047	0,053	0,88	0,38 *	0,055	0,100	0,55	0,58 *	0,358	0,15	0,88 *	0,009	No
64	Tweetie	0,587	0,043	13,60	0	0,960	0,109	8,83	0	0,572	1,68	0,09 *	0,373	Yes
65	Twitterrific	0,164	0,068	2,43	0,02	0,304	0,122	2,50	0,01	0,402	0,76	0,45 *	0,140	Yes
66	Tweetdeck	0,152	0,093	1,64	0,1 *	0,454	0,183	2,48	0,01	0,812	0,56	0,58 *	0,302	Yes
67	Twitterfon	0,773	0,121	6,38	0	1,220	0,170	7,20	0	0,580	2,11	0,03	0,447	Yes
68	Twittelator	0,235	0,182	1,29	0,2 *	-0,558	0,236	-2,37	0,02	0,708	-0,79	0,43 *	0,793	Yes
69	Birdfeed	0,629	0,149	4,22	0	0,229	0,200	1,14	0,25 *	0,669	0,34	0,73 *	0,400	Yes
	Estimated user-specific error: unobserved correlated effects on Twitter client choice													
70	Web					1,77	0,0129	136,66	0	0,0451	39,22	0		
71	Tweetie					2,29	0,015	153,36	0	0,0728	31,54	0		
72	Twitterrific					2,39	0,0198	120,34	0	0,0686	34,82	0		
73	Tweetdeck					2,36	0,0241	97,71	0	0,0961	24,50	0		
74	Twitterfon					1,61	0,0325	49,62	0	0,184	8,77	0		
75	Twittelator					1,32	0,0478	27,70	0	0,244	5,42	0		
76	Birdfeed					1,74	0,0349	49,77	0	0,13	13,42	0		

Emergence of Superstars in Online Social Networks

Devavrat Shah and Tauhid Zaman
devavrat@mit.edu, zlisto@mit.edu
LIDS, Department of EECS
Massachusetts Institute of Technology

Abstract

Information or influence spreading in online social networks is a network growth process. The popular "preferential attachment" network growth model possesses power law degree distributions observed in infrastructure networks like the World Wide Web and the power-grid [1]. In contrast, we find that information-propagation networks such as the micro-blogging site Twitter exhibit non-local phenomena in addition to power law degree distributions. Specifically, there is always a "superstar" node with extremely high degree that is not predicted by preferential attachment. Here we show that local network growth models such as preferential attachment are not sufficient to describe information propagation phenomena that occur in real online social networks such as Twitter. We propose a new class of network growth model which incorporates the global structure of the network. One special instance of this model is equivalent to preferential attachment, while another special instance of this model accurately describes the power law and superstar phenomena seen in Twitter networks.

Empirical Observations: In Twitter users post messages known as tweets which can be forwarded in what are known as retweets. We collected retweets from Twitter about several different live events, such as sporting events and musical performances [2]. For example, we collected any retweet with the phrase "World Cup" during the 2011 World Cup opening ceremony. Because each retweet represents an edge between the tweet source and the retweeter, we can connect these retweet edges to form a network for each event. Each of these retweet networks contained many small connected components, and one large connected component. Throughout this work, we focus on these largest connected components, which are shown in Fig. 1a.

The empirical degree distributions for the retweet networks are shown in Fig. 1b. It can be seen that the distributions follow a power law, but there is another interesting phenomenon occurring. Each retweet network has one very high degree node, which we call a *superstar*. This superstar node has degree that scales like $\Theta(N)$, where N is the network size. In contrast, the popular preferential attachment model has maximum degree that scales like $\Theta(N^{1/2})$. Therefore, this superstar phenomenon requires a new network growth model.

Data Model: We propose the following model for these retweet networks. We start with a single node, and then at each time step a new node arrives and connects with an existing node with probability proportional to the node's *importance* function. The key question then becomes what *importance* function to use. If one uses the node's degree, then this model reduces to preferential attachment. However, we require an importance function that can incorporate the *global* network structure. We find that if one uses the network centrality measure known as rumor centrality [3], then we get excellent agreement with the Twitter data, as shown in Fig. 1b. For each of the different retweet networks, rumor centrality not only reproduces the superstar degree very closely, but also the exponent of the power law. Therefore, it seems that rumor centrality possesses the correct global structural information to describe the growth of these networks. Rumor centrality was designed as an estimator for finding rumor sources in networks [3]. Specifically, a node's rumor centrality is related to the likelihood of it being the rumor source.

Model Analysis: We now explain why rumor centrality leads to a power law and superstar. To do so, we claim the following "fixed point" structure for a rumor centrality network with N nodes when N is large. There is one superstar node with degree $N/2+o(N)$, i.e. roughly half of the nodes neighbor the superstar. Also, there is a positive integer c such that no more than c neighbors of the superstar are roots of subtrees with $\Theta(N^{1/2})$ nodes. The remaining nodes of the network are roots of subtrees with $o(N^{1/2})$ nodes. Finally, there are only $o(N^{1/2})$ nodes more than two hops from the superstar.

To understand the behavior of the superstar node, consider a simple star network with a central superstar node as the hub with N neighbors. Using results from [3] it can be shown that the rumor centrality of the superstar is $N!$ and the rumor centrality of every other node is $(N\text{-}1)!$. Using these values, the attachment probability of the superstar is 1/2. We prove that in the fixed point network, the superstar's attachment probability is slightly perturbed from 1/2 to 1/2-$o(1)$. Thus, each new node has roughly a 1/2 chance of joining the superstar, and so the degree of the superstar should be close to $N/2$, where N is the network size.

The power-law degree distribution can be understood if we look at the ratio of rumor centralities of non-superstar nodes. It was shown in [3] that the ratio of rumor centralities of any two nodes u and v which neighbor the superstar v^* is given by

$$\frac{R(v,G)}{R(u,G)} = \frac{T_v^{v^*}}{T_u^{v^*}} \frac{N - T_u^{v^*}}{N - T_v^{v^*}} \approx \frac{T_v^{v^*}}{T_u^{v^*}} \approx \frac{d_v}{d_u} \qquad (1)$$

where the variable $T^{v^*}_v$ is the size of the subtree rooted at node v and pointing away from node v^* and d_v is the degree of node v. For large N, because each subtree has size at most $\Theta(N^{1/2})$, we will have $N >> T_v^{v^*}$ for any node $v \neq v^*$ in the network, so this ratio is dominated by the $T_v^{v^*} / T_u^{v^*}$ term. Also note that because there are only $o(N^{1/2})$ nodes more than 2 hops from the superstar in the entire network, the subtree size of these nodes is very close to their degree. Therefore, the ratio of the attachment probabilities is approximately equal to the ratio of the degrees, just as in preferential attachment. This explains the near equivalence of the power law exponent with that of preferential attachment.

Conclusion: We have shown that with rumor centrality we obtain a network growth model that accurately describes retweet networks in Twitter. It may be that the reason why rumor centrality works for these networks is that it correctly quantifies the importance of a node in a retweet network in terms of how likely it was to have been the source or followed. When viewed this way, rumor centrality could be used for other applications such as predicting the spread of information or ranking influential people in these networks.

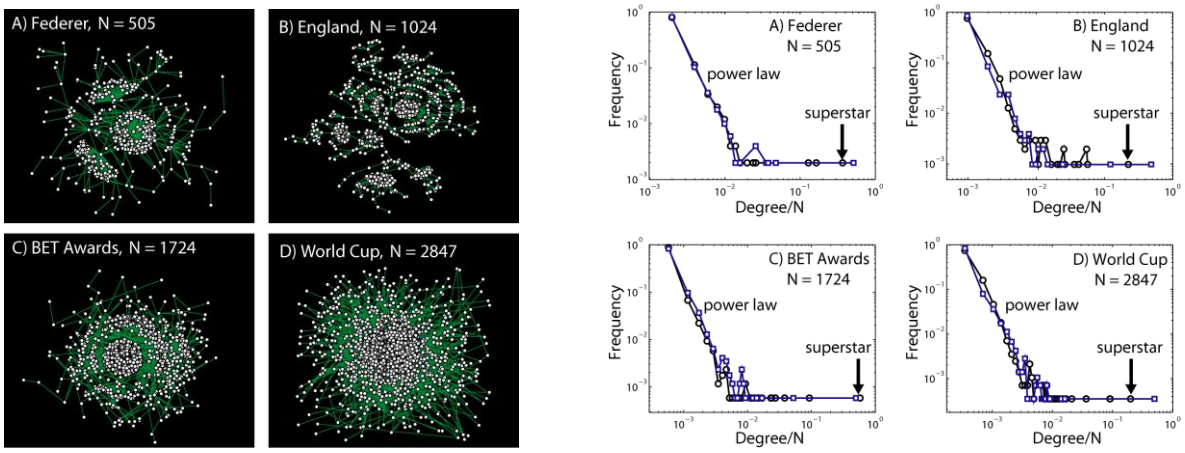


Fig. 1. (a) The largest connected component for different retweet networks in Twitter (N is the network size). **(b)** The degree distribution for the largest connected component of the retweet networks and corresponding equal size simulated networks using rumor centrality as an importance function. The curves correspond to the retweet network (black circles) and rumor centrality network (blue squares). The x-axis is the node degree normalized by the network size N. The event key phrases for the networks are (**A**) Federer, (**B**) England, (**C**) BET Awards, and (**D**) World Cup.

References and Notes

[1] A. Barabasi, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509 (1999).
[2] Data courtesy of Microsoft Research Cambridge.
[3] D. Shah, T. Zaman, Detecting sources of computer viruses in networks: theory and experiment. *Proc. ACM Sigmetrics* (2010). (Longer version available at http://arxiv.org/abs/0909.4370).

Bias in Social and Mainstream Media

Yu-Ru Lin[1,2,*], James P. Bagrow[3,4,†], and David Lazer[1,2,5,‡]

[1]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA
[2]College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA
[3]Center for Complex Network Research, Northeastern University, Boston, MA 02115.
[4]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA.
[5]Department of Political Science, Northeastern University, Boston, MA 02115, USA.
[*]yuruliny@gmail.com [†]bagrowjp@gmail.com [‡]davelazer@gmail.com

Social media, such as blogs, are often seen as democratic entities that allow more voices to be heard than the conventional mainstream media as well as a balancing force against the arguably slanted elite media. A systematic comparison between social and mainstream media is necessary but challenging due to the scale and dynamic nature of modern communication. We propose empirical measures to quantify the extent and dynamics of social (blog) and mainstream (news) media bias. We focus on a particular form of bias—coverage quantity—as applied to stories about the 111th US Congress. We compare observed coverage of Members of Congress against a null model of unbiased coverage, testing for biases with respect to political party, popular front runners, regions of the country, and more. Our measures suggest distinct characteristics in news and blog media. A simple generative model, in agreement with data, reveals differences in the process of coverage selection between the two media.

The extent of media bias determines the information available to the public and can affect public opinion and decision-making. Social media, powered by the growth of the Internet and related technologies, is envisioned as a form of grassroots journalism that blurs the line between producers and consumers and changes how information and opinions are distributed. Indeed, social media can be used by underprivileged citizens, promising a profound impact and a healthy democracy.

Do social media exhibit more or less bias than mass media and, if so, to what extent? Identifying media bias is challenging for a number of reasons. First, bias is "in the eyes of the beholder" and hence not easy to observe, e.g., conservatives tend to believe that there is a liberal bias in the media while liberals tend to believe there is a conservative bias [1]. Second, the assessment of bias usually implies knowing what "fairness" would be, which may not be available or consistent across different viewpoints. Third, Internet-based communication promises easy, inexpensive, and instant information distribution, which not only increases the number of online media outlets, but also the amount and frequency of information and opinions delivered through these outlets. The scale and dynamic nature of today's communication should be accounted for.

Our major contribution is that we propose empirical measures to quantify the extent and dynamics of "bias" in mainstream and social media (hereafter referred to as *News* and *Blogs*, respectively). Our measurements are not normative judgment, but examine bias by looking at the attributes of those being mentioned, against a null model of "unbiased" coverage. We focus on the number of times a member of the 111th US congress was *referenced*, and study the distribution and dynamics of the references within a large set of media outlets. We demonstrate bias measures for slants in favor of specific political parties, popular front-runners, or certain geographical regions.

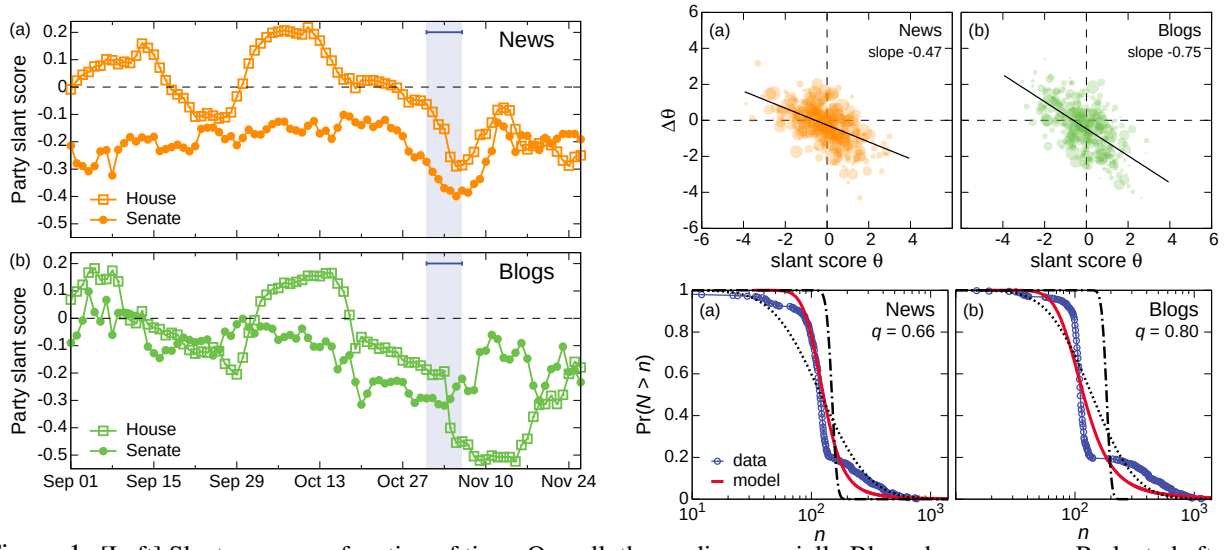


Figure 1: [Left] Slant score as a function of time. Overall, the media, especially Blogs, become more R-slanted after the 2010 election. [Top-right] Media outlets are slightly shifting towards the other side after election. The majority of news outlets become slightly more R-slanted. For blogs, originally D-slanted blogs become more R-slanted. Each point represents a media outlet. [Bottom-right] The generative model for the distribution of references n per legislator. The larger value of q for Blogs indicates that they are more driven by the rich-get-richer mechanism than News (both distributions are heavy-tailed). Dashed lines indicated fitted poisson and log-normal distributions, for comparison.

Consider a news or blog outlet's biased coverage of two political parties. We quantify bias of an outlet i by a *slant score* θ_{ik} which is defined as $\theta_{ik} = \log(\text{odds-ratio}) = \log\left(\frac{n_{ik}/(n_i - n_{ik})}{p_k/(1-p_k)}\right)$, where n_{ik} be number of times an outlet i references legislators in group k, n_i is the total references of i, and p_k is the *baseline probability* that i refers to k. The advantage of having such a baseline probability is that "fairness" become configurable, e.g., one can consider fairness as a 50-50 chance to reference either party (i.e. $p_{\mathrm{D}} = p_{\mathrm{R}} = 0.5$), or define $p_{\mathrm{D}} = 0.6$ since roughly 60% of the Congress are Democrats. In this two-party case, we take $\theta_i \equiv \theta_{ik}$, with $k = \mathrm{D}$, and $\theta_i > 0$ means outlet i is more likely to be D-slanted and 0 simply means no bias w.r.t that baseline. To characterize the overall bias within a media, we derive a media-wide *collective slant score*, Θ, which is defined as $\Theta \equiv \theta^*$, where θ^* is the asymptotically unbiased estimator for θ based on a *random effect* model. The dynamics of media bias can be measured as a function of time (Fig. 1 [left]).

We extend such dichotomous-outcome measures to multi-outcome bias measures such as front-runner slant. Using these measures to examine newly collected data, we have observed distinct characteristics of how News and Blogs cover the US congress. Our analysis of party and ideological bias indicates that Blogs are not significantly less slanted than News. However, their slant orientations are more sensitive to exogenous factors such as national elections (Fig. 1 [top-right]). In addition, blogs' interests are less concentrated on particular front-runners or regions than news outlets.

To better understand the distinctive slant structures between the two media, we propose to use a simple "wealth allotment" model to explain how legislators gain attention (references) from different media. The results about blog media's inclination to a rich-get-richer mechanism indicates they are more likely to echo what others have mentioned (Fig. 1 [bottom-right]). This observation does not contradict our measures of bias – compared with news media, blogs are weaker adherents to particular parties, front-runners or regions but are more susceptible to the network and exogenous factors.

[1] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly J. of Economics*, 120(4):1191–1237, 2005.

The Effects of Just-In-Time Social Networks on People's Choices in the Real World

Kwan Hong Lee
MIT Media Lab
Cambridge, MA 02139
kwan@media.mit.edu

Andrew Lippman
MIT Media Lab
Cambridge, MA 02139
lip@media.mit.edu

Alex S. Pentland
MIT Media Lab
Cambridge, MA 02139
sandy@media.mit.edu

ABSTRACT

We address the question of how online social networks affect real world just-in-time decisions. The question is significant due to the pervasiveness of mobile devices in our just-in-time decisions and the way we are connected to our social networks at various scales, across time and space, through these mobile communication channels. An empirical inquiry on mobile social influence and how these social networks impact our decisions will provide a framework for utilizing these virtual social influences to build persuasive mobile interfaces and provide timely decision aids that can help with our personal and social goals in the real world. We approach this problem through a real world experiment where we deploy mobile digital menus (Social Menu) in a restaurant and capture people's dish choices in real time. Results show that the modality in which social information is presented affects people's decisions in the dimensions of taste, time and price.

METHOD AND RESULTS

With wide scale adoption of smart phones world wide, the mobile devices are increasingly becoming influential as they are utilized for location sensitive and time limited decisions such as shopping, eating, meeting people, and traveling. During the past holiday season over 60% of mobile phone users used their phones to pre-shop before they went to the stores. In the physical world, unlike online transactions, people face cognitively limited situations where decisions have to be made on-the-go within limited time and space. We try to plan our schedule but many times we are faced with many options and have to make choices on the spot. In the case of grocery shopping, up to 70% of purchases are decided in the store[12]. The difference between an online and offline purchase scenario is that there is a cost to not making a decision in the offline world. If one needs to purchase at a store later, one needs to travel back to the store to make the purchase while online purchases can be done at anytime and anywhere.

In our Social Menu study, we investigate the impacts of social proof and mental shortcuts due to mobile social information by instrumenting real people in the real world. Mobile devices not only allow people to connect with other people at varying scales at anytime, from anywhere, but also allow us to capture and share people's economic activity in real time. Systems like SmartRestaurant made lunch menus of a local campus restaurant accessible over the phone and allowed people to pre-order their lunch for pick up and make payments through the phone[9]. The resulting transactions are a valid proxy to the economic decisions and by instrumenting the choice architecture we can understand the impacts of augmented social information. Our goal is to understand the impacts of virtually mediated social influences on people's decisions and how it relates to time, taste, price and the social environment.

We approached this problem by creating a digital menu mobile application we call Social Menu, that was used directly in people's economic decisions. 250 people participated in the study by dining at the restaurant. Prior research in this area has focused on clever experiments with individuals and trained confederates that influence the research subjects in lab and real world settings. Other experiments involved imagined situations and surveys. Our approach allows us to observe people in different social networks in the real world in a microscopic manner with large groups of people. Therefore it allows us to capture multi-source, multi-target social influences over time and at scale, which requires further real world studies[10]. We measured behavioral traces by capturing the categories and dishes they browsed, and the time to making their choices. The work was evaluated by analyzing the data from different experimental groups.

The results show:

1. Diversion of choices: about 50.2% of people made on the spot decisions that did not include any of their favorite dishes from the online pre-survey menu. This implies that people's choices can be changed by the current context and is in agreement with studies on shopping where people make 70% of purchase decisions in the store. Therefore, mobile guided just-in-time decision systems could have significant influence on people's choices.

2. Second degree friends have experienced as many common dishes as the first degree friends indicating that friends of two degrees of separation can provide people with reinforcement in their choices.

3. Scale of influence: empirical data indicates that across time accessibility and scale of virtual social information may affect 2 to 10 times more people in their considerations compared to physical social influence by co-present party on the table.

4. Time of engagement: friend's names (individuals) on the menu made people spend longer time to decide indicating that seeing other people's choices encourages one to

spend more time to evaluate choices before making a decision.

5. Price factor: average price comparison between different experimental groups show that anonymous group of friends had strongest influence in pulling people to choosing cheaper items.

6. Summary: Individual friends increase engagement, group of friends affect price choice and popularity serves as shortcuts to decision making.

BACKGROUND

Recent research with mobile phones have allowed us to capture in detail and understand our communication patterns, mobility patterns and to deduce how people behave in aggregate in the real world. Researcher have been using mobile probes[6] to capture and understand people's shopping behavior. Bluetooth scanning and location based information from mobile phones have been used to capture people's social relationships in the real world, their patterns of activity and their habits[3]. AT&T study showed that people in New York City travel larger distances compared to people in LA[7]. In contrast, Barabasi's work showed how people in a city in Europe regularly do not leave the 3 mile radius during their daily life. The communication patterns based on frequency of incoming and outgoing calls also allow the tie strengths of customers to be identified[11]. Instead of focusing on mobility patterns and tie strengths, we capture the choices people make through the mobile phones and inject social information to understand how just-in-time choices are affected when certain social signals are published from the social network.

When people are seated together at a table and are choosing their dishes, the sequential order creates social influence that makes people to choose differently from preceding dish choices by others. Ordering patterns in a Chinese restaurant were investigated to understand group ordering behavior and the results show that on average people's dish choices diverge from other's choices seated on the same table[1]. More importantly, peer impacts and normative influences that are most situationally similar can affect the outcome of people's decisions[5].

Decades of research has been done in how people make decisions in varying social contexts, however with the wide adoption of smartphones, social interactions we now engage in has become a mix of virtual and physical interactions. Lab experiments and field studies of social scientists have shown how social influence causes people to make irrational decisions and how such forces can be identified, managed and utilized for the benefit of achieving certain goals of persuasion[2]. As people in the US are spending over 20 billion hours a year on Facebook both online and mobile, it is unprecedented how such networks might impact people's choices in the real world. We attempt to further the understanding of interaction of virtual social networks in the physical world by investigating the impacts of social information in particular decision making scenario that is constrained by space and time. People also are prone to deal with uncertainty by following other people's choices. In the context of the restaurant, research have been done on tipping, menu choices and how price plays a role on people's choices. We are extending these bodies of research by investigating the effects of mobile mediated influences as mobile phones prevail in our lives.

Most recently, influence of social networks on social networking sites and cultural markets have revealed the effects of status in purchase behavior and resulting unpredictability and inequality. We extend these studies by engaging and measuring different types of social influence (peers, peers anonymous and popularity) in a real world setting with real decisions.

Analysis of 208 users in the most popular social networking site Cyworld in Korea, shows that there are three different groups of users with very different purchase behaviors[8]. The low status group of about 48% are not affected by social influence because they are not well connected and show limited interaction with others in the social network. The middle status group of about 40% are moderately connected and are influenced to generate 5% higher revenue. Finally the 12% of high status group are very active on the site and are negatively impacted by their friend's purchases.

Salganik et al.[13] investigated the role of social influence in the inequality and unpredictability of success in an artificial cultural market. They were able to create an artificial music download site to experiment music selection by real people. They separated the world into 9 different worlds with 1 world being the independent condition. The other 8 worlds were socially influenced worlds (showing download popularity) that were independent from each other. The socially influenced worlds were shown the number of downloads next to the songs. The socially influenced worlds showed consistently higher *inequality*, popular songs are more popular and unpopular songs are less popular and higher *unpredictability* of success of good quality content. Inequality is measured by the average difference in market share between all pair of songs and unpredictability is measured by the Gini coefficient.

Fogg[4] iterates how mobile phones can be used to for opportune intervention to improve individual and social behaviors. People are wedded to these devices where many people spend more time with their mobile devices than any other human being. The nature of mobile phones being always available and responsive allows it to be a continual channel of influence. Experiments performed with mobile applications in encouraging better eating habits, recycling behaviors and healthy activity have shown positive outcomes in encouraging behavioral changes. They also document that connecting with people who are enacting on similar behavioral changes strengthens the effectiveness of the application due to power of social comparison. We investigate these social influences more in detail in the context of just-in-time setting to see how different modalities of social information affects people's choices.

REFERENCES

1. Ariely, D., and Levav, J. Sequential choice in group settings: Taking the road less traveled and less enjoyed. *Journal of Consumer Research 27*, 3 (2000), 279–290.

2. Cialdini, R. B. The science of persuasion. (cover story). *Scientific American 284*, 2 (February 2001), pp. 76–.

3. Eagle, N., Pentland, A. S., and Lazer, D. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences 106*, 36 (2009), 15274–15278.

4. Fogg, B. J. *Persuasive Technology*. Morgan Kaufmann, San Francisco, CA, 2003.

5. Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research 35*, 3 (2008), 472–482.

6. Hulkko, S., Mattelmäki, T., Virtanen, K., and Keinonen, T. Mobile probes. In *Proceedings of the third Nordic conference on Human-computer interaction*, NordiCHI '04, ACM (New York, NY, USA, 2004), 43–51.

7. Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Rowland, J., and Varshavsky, A. A tale of two cities. In *HotMobile '10: Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications*, ACM (New York, NY, USA, 2010), 19–24.

8. Iyengar, R., Han, S., and Gupta, S. Do Friends Influence Purchases in a Social Network? *SSRN eLibrary* (2009).

9. Lukkari, J., Korhonen, J., and Ojala, T. Smartrestaurant: mobile payments in context-aware environment. In *ICEC '04: Proceedings of the 6th international conference on Electronic commerce*, ACM (New York, NY, USA, 2004), 575–582.

10. Mason, W. A., Conrey, F. R., and Smith, E. R. Situating Social Influence Processes: Dynamic, Multidirectional Flows of Influence Within Social Networks. *Personality and Social Psychology Review 11*, 3 (2007), 279–300.

11. Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences 104*, 18 (2007), 7332–7336.

12. Packard, V. *The Hidden Persuaders*. Pocket Books, 1957.

13. Salganik, M. J., Dodds, P. S., and Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science 311*, 5762 (February 10 2006), 854–856.

Localizing Externalities in Social Networks: Inducing Peer Pressure to Enforce Socially Efficient Outcomes

Ankur Mani, Iyad Rahwan, Alex (Sandy) Pentland
amani@mit.edu, irahwan@acm.org, pentland@mit.edu

Abstract

Can we use the value in social relationships to promote cooperative behavior and responsible consumption? We consider the example of energy consumption. It is an activity that produces negative externalities that affects the whole society. This is a well-known *free rider* problem, where free riders carry less than their fair cost of consumption. The equilibrium outcome is always sub-optimal. There are two traditional institutional solutions to reducing such externalities: *quotas*, which impose caps on consumption, and (Pigouvian) *taxation* [1], which impose a cost to such consumption. In this paper, we propose a new approach, which exploits the value of social relationships, and the ability of peers to exert pressure on one another. We create a suitable reward structure in which peers of an individual get rewarded for the individual's reduced consumption. This creates localized externalities in the social network and induces peer pressure on the individuals to reduce consumption. By localizing the externalities produced by an individual on his/her social peers, we can use the value in social relationships to promote cooperative behavior and responsible consumption. We show that in most cases, our approach requires a smaller budget to implement than Pigouvian subsidies.

1 Extended Abstract

We propose a novel approach of *localized externalities*, which exploits the value of social relationships, and the ability of peers to exert pressure on one another. Localizing the externalities involves amplifying them, by punishing/rewarding peers disproportionately based on the individual's behavior. Thus the global externalities are exposed to the immediate peers (peers) of an individual in the social network through the local externalities. The externality produced by an individual on the whole society also has a magnified effect on his peers by punishing/rewarding them (see Figure 1) and hence the peers exert high pressure on the individual on behalf of the whole society. By localizing the externalities produced by the individual to his/her social peers, we can use such social relationships as means for enforcing more efficient outcomes.

The effect of peer pressure in social networks has been studied by Calvo-Armengol and Jackson [2]. In their model, individuals impose externalities over their peers through their actions, and these peers can impose social pressure to influence those actions. In contrast, we are interested in scenarios where an individual's action imposes externalities on the entire society, but only peers are able to impose pressure as discussed in the dorm example. Hence, we study a model in which, peers are *induced* to exercise their ability to impose pressure on behalf of the whole society.

We now discuss the key results. We consider a network of individuals connected via social ties. Each individual in this network produces, through his/her consumption, a (negative or positive) externality that affects the rest of society. Individuals choose the amount of consumption and the pressure to exert on each of their peers. We analyze the two stage game. In the first stage, individuals decide the amount of peer pressure to exert on their peers. In the second game, each individual observes the peer pressure on herself and decides the amount of consumption. An individual derives utility from her own consumption and receives a reward depending upon the consumption of her peers. An individual also experiences a cost proportional to her consumption due to peer pressure from her peers in the social network. Individuals also experience cost of exerting pressure on their peers and the cost is proportional to the amount of pressure they exert.

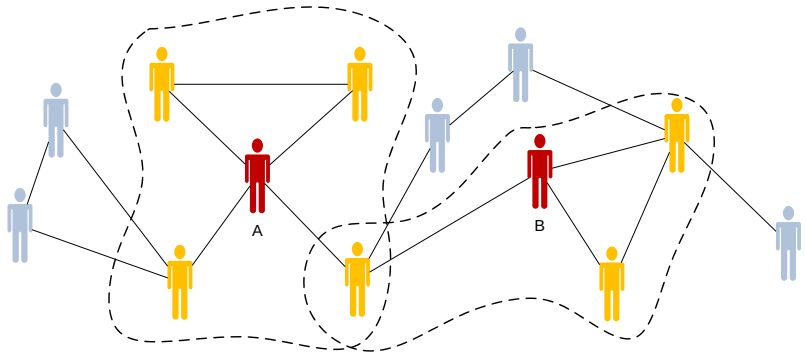


Figure 1: Scenario with localized externality: The peers of individuals A and B receive amplified (and more salient) punishment/reward for the negative/positive externalities produced by these individuals.

We study a subgame perfect equilibrium of the game where an individual experiences equal pressure from all her peers. We show that:

1. In the second stage of the game, there exists a peer pressure profile such that the equilibrium consumption is equal to optimal social consumption. In this peer pressure profile, each individual experiences a net marginal peer pressure equivalent to the net marginal externality she imposes on the whole society. Let us call this the optimal peer pressure profile.

2. We characterize the reward structure such that in equilibrium the peer pressures profile chosen in the first stage of the game is the optimal peer pressure profile. At the optimal consumption profile, the marginal reward to the peers of an individual balances the marginal global externality on the peers and the marginal reduction in cost of exerting pressure to reduce the individual's consumption for the peers. When the individual's consumption is lower than the optimal consumption, the marginal reward is higher causing the peers to reduce pressure and when the individual's consumption is higher, the marginal reward is lower, causing the peers to increase pressure.

3. We also study the constraints on the marginal cost of exerting peer pressure such that the optimal peer pressure can be induced with the limited budget for the reward. We show that when the marginal cost of exerting pressure is below a threshold, then the optimal consumption profile can be achieved using a budget lower than the budget otherwise used to subsidize low consumption. In fact, if the network is to large as compared to the degree of any node, then this threshold is very high and hence the reward structure is better than Pigouvan subsidies.

4. When the individuals are rewarded both for their peers' consumption as well as the pressure their peers' exert, then with a suitable reward function, all peers of an individual exert equal pressure on the individual and there can be no coalitions that are incentive compatible. The budget for rewards in this case is still lower than the budget for Pigouvian subsidies.

References

[1] W. J. Baumol. On taxation and the control od externalities. *American Economic Review*, 1972.

[2] A. Calvó-Armengol and M.O. Jackson. Peer pressure. *Journal of the European Economic Association*, 8(1):62–89, 2010.

[3] Garrett Hardin. Tragedy of the commons. *Science*, pages 1243–1248, 1968.

Spread of Influence in Cellular Social Networks

Abhik Das, Suriya Gunasekar, Sujay Sanghavi and Sriram Vishwanath
Department of ECE, University of Texas at Austin, USA
{akdas,suriya,sanghavi,theory}@mail.utexas.edu

Social networks have always been central to human interaction and behavior; recent years have seen a surge in interest in their study – primarily due to the large-scale and fine-grained access provided by online networks like Facebook® and Twitter®. This work concerns the study of a similar large-scale and finely-sampled social network, arguably as important but far less studied: the cellular social network. In particular, cell-phone calls and text messages reflect an underlying social network; this work leverages a unique data-set we have obtained to address several fundamental social network questions. *We note that this work is in its formulation stage and only some of the early explorations are discussed here in this abstract.*

The source of data is the 30 months (from November 2006 – April 2009) record of cell-phone usages of about 20 million users of an Asian telecom giant. This is a rich database which has monthly records of various services like calls/SMS's/MMS's made by users, data/web usage etc, along with information such as call timings, location ID and applicable discount schemes. It also has the data of complaints registered and billing information. Separate cross-referenced tables record other demographic data about users such as their age, gender, marital status, date of birth, call plan etc. The size of this database is $\sim$4 TB.

The nature and massive scale of this database provides unique opportunities and algorithmic challenges. The richness of the data arises from the wide range of demographic and user meta-data provided and the fact that the data represents "real" contact patterns. Some of the interesting problems which can be explored in this framework include learning network structures (e.g. the friend/colleague circle of a group of users), finding most influential users, detection and prediction of cascades/large-scale events (e.g., outbreak of viral MMS or malware). The main challenges associated with these problems concern with pre-processing required on the data to extract essential information pertaining to specific objectives. Working with the social network as the fundamental unit, defining the notion of "friends" is highly subjective. But our prime challenge arises from the massive size of the data which poses scalability and computational issues.

We look into the problem of finding influential users in a cellular network – given the social network topology, how to choose the initial set of individuals to seed the influence, in order to maximize the word-of-mouth/viral spread in the network within a given time frame. *Crucially, we evaluate our schemes empirically using the actual call timings and durations.* The problem of influence maximization, with applications to marketing, was studied by Domingos and Richardson [1]. Kempe, Kleinberg and Tardos [2] rephrased influence maximization as a discrete optimization problem and developed a greedy algorithm for choosing the initial seed set (the individuals to influence), which gives a 63%-approximate solution with respect to the optimal one. However, the greedy algorithm is not scalable to social networks with large number of nodes and edges. As a starting point, we construct a social network graph from one month of the call record. Our long term goal is to develop an efficient and scalable algorithm to select the best initial seed set of individuals. We intend to make the model more robust by incorporating the data from longer time duration, demographic information and SMS/MMS patterns to get a better estimate of the edge strengths.

We consider the social network of mobile phone calls made by users during the month of January 2009,

and represent this by an undirected graph $G = (V, E)$, where V is the set of users. An undirected edge in E is constructed between two users in V, if atleast 4 calls are exchanged between them in the entire month. This graph results in a giant component $G' = (V', E')$ with ∼820000 users in V' and ∼5 million edges in E'. *Indeed, one of our first main challenges is devising algorithms that remain tractable at this scale.* We simulate the spread of influence in G' using a version of the *independent cascade* model for influence dissemination. We assume that whenever a call is established between an influenced user A and a neutral user B within 10 days since A got influenced, with some fixed probability p, the neutral user becomes influenced. The weights are assigned to edges in E' as follows: if m calls are exchanged between two users during the entire month, we assign an edge weight of $1 - (1 - p)^{m/3}$, which is the probability of success in influencing a neutral user within 10 days (we assume calls are made uniformly over the month). For a given seed size k, we choose k users for influence maximization as per the following criteria:
1. *Highest (weighted) degree*: we choose k users with the highest (weighted) degrees,
2. *Selective highest degree*: we choose k users as follows – in every step, we choose 10 highest degree users and eliminate some of their neighbors who have a high chance of getting influenced,
3. *Shortest paths*: we choose $2k$ users with highest degrees (we assume influence spreads only via shortest paths) and choose k users in order of expected spread each user can produce via shortest paths.

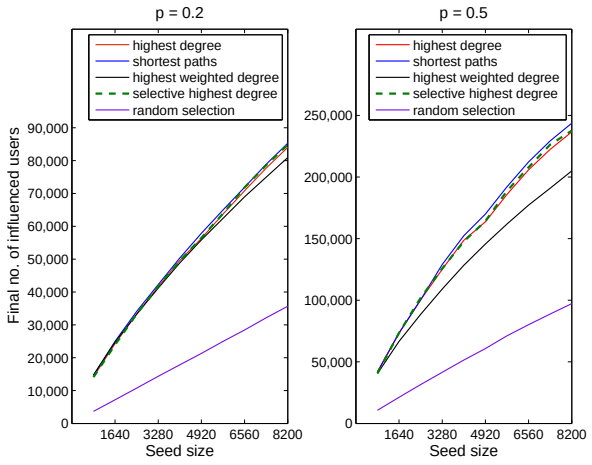


Figure 1: Performance of seed selection algorithms

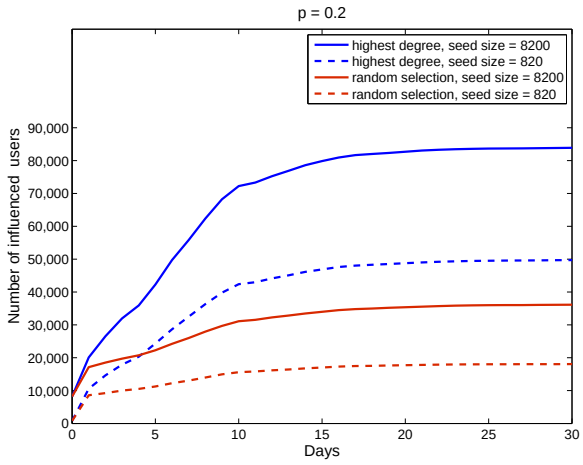


Figure 2: Time evolution of influence spread

The plots depicting performances of the seed selection algorithms are presented in Figure 1 for $p = 0.2, 0.5$. The time evolution of the number of the influenced users across the month is presented in Figure 2. We are continuing work on development of optimal seed selection algorithms for influence maximization. The hope is to obtain close-to-optimal yet scalable techniques for cellular large social networks.

References

[1] P. Domingos and M. Richardson, "Mining the network value of customers," *Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining*, 2001.

[2] D. Kempe, J. Kleinberg and E. Tardos, "Maximizing the spread of influence through a social network," *Proceedings of the 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.

Social networks and research output

Lorenzo Ductor* Marcel Fafchamps† Sanjeev Goyal‡
Marco J. van der Leij§

March 2011

*University of Alicante. E-mail: lductor@merlin.fae.ua.es
†University of Oxford. Email: marcel.fafchamps@economics.ox.ac.uk
‡University of Cambridge. E-mail: sg472@econ.cam.ac.uk
§E-mail: mvanderleij@gmail.com

Good recruitment requires an accurate prediction of a candidate's potential future performance. Sports clubs, academic departments, and business firms routinely use past performance as a guide to predict the potential of applicants and to forecast their future performance. A football club looks at goals scored or passes made, an academic department looks at published papers, while an investment bank looks at past bonus earnings of applicants.

In this paper the focus is on researchers. Social interaction is an important aspect of research activity: researchers discuss and comment on each other's work, they assess the work of others for publication and for prizes, and they join forces to coauthor publications. Scientific collaboration involves the exchange of opinions and ideas and facilitates the generation of new ideas. It follows that the characteristics of one's collaborators and the general structure of the collaboration network may reveal useful information about future productivity. We also expect that access to new and original ideas helps researchers be more productive. So we would expect that, other things being equal, highly connected individuals or individuals who are 'central' in the network are more likely to be productive in the future.

Centrality and proximity themselves arise out of links created by individuals and so they reflect their individual characteristics – e.g., ability, sociability, and ambition. For instance, collaboration with highly productive coauthors may reveal that these coauthors find such collaboration worthwhile. Since the ability of a researcher is imperfectly known, the existence of such ties may by itself be informative. Moreover, if this is the case then, as individual performance gradually reveals information about a person's potential over time, we expect network information to be more useful for young researchers and less so for older researchers. In contrast, we expect access to new ideas to be valuable for all researchers. Consequently, we expect the role of networks as facilitating flows of ideas to remain relatively constant over an author's entire career. These observations form the basis of our empirical strategy.

The empirical work presented here first asks whether social network measures help predict future research output *over and above* the information contained in individual past performance. We then investigate which specific network variables are informative and how their informativeness varies over a researcher's career.

Our first set of findings are about the information value of networks. We find that incorporating information about coauthor networks leads to an improvement in the accuracy of forecasts on individual output, *over and above* what we can predict based on the knowledge of past individual output. The effect is significant but modest, e.g., the root mean squared error in predicting future productivity falls from 0.677 to 0.663, while the R^2 increases from 0.417 to 0.442. We also observe that several network variables – such as productivity of coauthors, closeness centrality, and the number of coauthors – have predicting power. Of those, the productivity of coauthors is the most informative network statistic among those we examine.

The predicting value of network information varies over a researcher's career: it is more powerful for young researchers but declines systematically with career time. By contrast, information on recent past output remains a strong predictor of future output over an author's entire career. As a result, fifteen years after the onset of a researcher's publishing career, the prediction accuracy with and without network variables is very similar.

Our second set of findings are about the relative importance of signaling versus flow of ideas in explaining the predictive power of network variables. To ascertain their relative

roles, we examine how the information content of specific network variables changes across an individual's career. This comparison builds on the idea that coauthor productivity contains more information about the type of author while topological variables contain more information about flow of ideas. We find that coauthor productivity is initially informative but this informativeness declines significantly over time. By contrast, topological variables such as centrality and degree are more modestly informative at the start of a career but retain their informativeness across time.

Our third set of findings is about the relation between author ability and the prediction value of networks. We partition individual authors in terms of past productivity and examine the extent to which network variables predict their future productivity. We find that the predictive value of network variables is non-monotonic with respect to past productivity. Network variables do not predict the future productivity of individuals with below average initial productivity. They are also uinformative for individuals in the highest past productivity quantiles. But they are informative about individuals in between. Taken together, these results predict that academics recruiters would benefit from gathering and analyzing information about the coauthor network of young researchers, especially for those who are relatively productive.

This paper is a contribution to the empirical study of social interactions. Traditionally, economists have studied the question of how social interactions affect behavior across well defined groups, paying special attention to the difficulty of empirically identifying social interaction effects. For an overview of this work, see for instance Moffitt (2001) and Glaeser and Scheinkman (2002). Identification of network effects is difficult as links in a network are endogenous and may be correlated with unobservable characteristics of individuals and links. In this paper we take an alternative route: we focus instead on the *predictive power* of social networks in terms of future research output.

We believe that the predictive value of social connections arises because they reveal information about individuals that is not apparent from their past output. Knowledge about a researcher's recent coauthors – e.g., their number, seniority, and own productivity, – constitutes a potentially valuable signal about the individual in question. This is because these coauthors have privileged information about the individual. The fact that they have coauthored with this individual indicates that this information is on average positive. Finding that an individual with high quality coauthors has higher predicted output in the future should not be interpreted as evidence of network effects in the traditional causal sense. Nonetheless, this knowledge can potentially be used by an academic department in making recruitment decisions.

References

[1] Glaeser, E. J. Scheinkman (2002), Non-market interactions. *Econometric Congress: advances in economic theory and econometrics.* edited by M. Dewatripont, L.P Hansen and S. Turnovsky. Cambridge University Press.

[2] Moffitt R. (2001), Policy interventions, low-level equilibria, and social interactions, in: S. Durlauf and P. Young (eds.), *Social Dynamics*, Cambridge: MIT Press.

Title: Automated Extraction of Social Networks from Meeting Transcripts

Authors: David A. Broniatowski, MIT Engineering Systems Division Research Affiliate, Synexxus Quantitative Analytics Senior Research Scientist, david@mit.edu

Many important technical and policy decisions are made by committees of experts. Society relies on these committees to fairly combine information from multiple perspectives in order to reach a decision that one person could not make alone. Although the research on group decision-making is vast, analysis of committees of experts in a real-world setting have been relatively scarce. This may largely be ascribed to difficulty in gathering data (e.g., because it might be proprietary or simply not recorded) and the absence of a corresponding methodology. The advent of the internet has made much text data available. Furthermore, regulations such as the U.S. Federal Advisory Committee Act (FACA) of 1997 ensure that transcripts of real-world expert committee meetings are available online or are available upon request. Finally, recent innovations in machine learning and computational linguistics have enabled the analysis of large sources of text data in a repeatable and consistent fashion. These methods have yet to be applied to the analysis of social data on a large scale. There is therefore an opportunity to apply some of these methods to enable a deeper empirical understanding of decision-making by committees of technical experts. Furthermore, some of these methods may be extended using signal processing techniques. Ultimately, these methods may help to generate quantitative insight into committee decision processes, perhaps enabling better decision outcomes. Using a methodology presented in (Broniatowski and Magee, 2010) we quantify information flows between members of a set of the 37 FDA Circulatory Systems Devices advisory panel meeting transcripts from meetings held from 1997-2005, in which the panel voted on recommendations for device Pre-Market Approval. These panels were selected because of the required participation of multiple experts from different specialties who must nevertheless come to a decision regarding data. In addition, each meeting transcript contains a voting record that might be used as a data source.

Data Acquisition and Coding

Each voting member was coded according to his or her medical specialty (e.g., surgeons, cardiologists, radiologists, electrophysiologists, statisticians, etc.) and according to how s/he voted in each of the 37 meetings in our sample, as recorded in the meeting transcript. Linguistic data consisted of the text of each transcript. Non-content-bearing words, identified using a standard list, were automatically removed. The frequency of each remaining word was then counted and assigned to each speaker in the meeting.

Generation of Social Networks

Social networks were generated from the same set of 37 transcripts using a method based on the Author-Topic model (Rosen-Zvi et al. 2004). Briefly, for each meeting transcript, non-content-bearing words (e.g., “is”, “the”, “and”, etc.) were removed using a standard list. The remaining content-bearing words were then parsed into a “word-document matrix”, where each row represents a unique word, and each column represents an utterance spoken by a speaker in the panel meeting. Each entry in the matrix is therefore a count of the number of times a given word occurs in a given utterance. The Author-Topic

model was then used to fit each transcript's words to a fixed number of topics, and to isolate the words that were specific to each voting member on each committee. Speakers who frequently shared the same topics were considered to be linked in a social network. This procedure was then repeated several times for each transcript in order to average across whatever probabilistic noise might exist in the Author-Topic model fit. Speakers who were frequently linked across multiple Author-Topic model fits were considered to be linked in the social network associated with that transcript. Full details of the methodology used to generate these social networks may be found in the article by Broniatowski and Magee (2010).

Network Analysis Metrics

For each network, we defined *specialty cohesion* as the total proportion of links between committee members who had the same medical specialty. This value was then compared to the specialty cohesion for 1000 random graphs with the same density and with committee members holding medical specialties as in the original network. *Specialty cohesion percentile* is defined as the proportion of random graphs with lower specialty cohesion than the network generated from the transcript associated with that meeting. 1000 random graphs were used because this number was empirically determined to generate stable results. Similarly, we defined *vote cohesion* and *vote cohesion percentile* the same way, only substituting panel members' votes for their specialties for the subset of 11 meetings in which there was a voting minority of at least two members. Subsequent results analyze the distributions of specialty cohesion percentile and vote cohesion percentile, and the relationships between these distributions.

Directed Graphs

We extend this preliminary method using cross-correlation techniques in order to generate directed social networks illustrative of the flow of influence within these meetings. Furthermore, we may quantify the impact that the committee chair has upon the meeting by determining, for each graph, which proportion of edges is part of a cycle. This is a metric of the hierarchy in the graph. The difference in this metric between graphs with and without the chair therefore quantifies impact of the chair on the meeting.

Findings include insights into the impact of professional specialty upon decision-making, identification of different leadership styles on these committees and possible indications of panels in which panel members may have learned from one another in order to reach a consensus decision.

Broniatowski, D. & Magee, C.L. Analysis of Social Dynamics on FDA Panels using Social Networks Extracted from Meeting Transcripts. *2nd IEEE Conference on Social Computing, 2nd Symposium on Social Intelligence and Networking* (2010).

Rosen-Zvi, M. et al. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence* 487-494(2004)

The Effects of Corruption on Organizational Networks and Individual Behavior

Brandy Aven
Carnegie Mellon University
Tepper School of Business
aven@cmu.edu

March 11, 2011

This paper analyzes the social networks of both non-corrupt and corrupt projects within an organization and the effects of these two information types on communication patterns. The results of this multi-method study show that the type of information communicated between organizational members affects both behavior and network characteristics. By examining email data taken from Enron between 1998 and 2002, I find that in contrast to non-corrupt information networks, corrupt project networks are less connected, less reciprocal, and communication is less frequent. These different patterns appear to be the aggregate effects of individual level behaviors. When communicating information about corruption, individuals' communications are also less symmetrical and less transitive. In other words, individuals are less likely to reciprocate communications or to introduce their alters when sharing corrupt information. The central claim of this research is that when information is meant to be kept secret as with corrupt information, the processes by which people share or discuss the information require different behaviors and strategies than if the information is public. These strategies then in turn alter the network form.

In the first study, I look at the topological implications of information on six different project networks – three non-corrupt and three corrupt. Initially for each relation within a project, I calculated the tie-strength based on the frequency of emails exchanged and performed a paired t-test, with project information (corrupt/non-corrupt) as the grouping variable. Communication frequency is a common means by which to operationalize tie-strength (Granovetter 1973). The mean frequency for corrupt information was significantly lower 2.787 than for non-corrupt information 7.454 (t = -3.6025; df = 4,957, $p < .001$). This reduced traffic found in corrupt information networks may be explained by the desire of the participants to minimize the possibility of detection where each communication increases the risk of discovery. Next, I apply of a classic method of data normalization, z-score transformation, to provide a way of standardizing the data across a range of networks independent of their size and density (Robins and Alexander 2004). The z-scores for each project were based on 100 simulated random networks, conditioned on size and degree distribution. For corrupt communications, the network structure is sparse with comparably low connectivity between actors (z-score = -37.0402) as compared to non-corrupt networks (z-score = -20.6273). These lower levels of connectivity may serve as organizational buffers, sealing off groups of members from each other (Simmel 1950; Goffman 1970). Such fractioned networks also decrease the likelihood identifying all the participants. Finally, corrupt communication structures are overwhelmingly asymmetrical (z-score =-275.423), meaning that there are few reciprocal communications. By contrast, non-corrupt communication structures are far more reciprocal (z-score =-49.2616). A structure that contains a high number of asymmetrical or nonreciprocal ties delineates differences in information acquisition, with information generally flowing to the most powerful or highest status members in the network (Brass, Butterfield, and Skaggs 1998). Hierarchical structures such as these where power is centralized are common to illegal enterprises (Baker and Faulkner 1993). In sum, in the first study I find evidence that the group level communications of corrupt networks differ from the non-corrupt networks along three dimensions that are consequential for enterprises.

In the second study, I explore how corrupt and non-corrupt information influences individual communication strategies. Here 106 individuals were identified who participated in both the corrupt and non-corrupt projects I observe in study 1, totaling 198 observations across the networks. For these individuals, I analyze three different dependent variables to determine the effect of information on the member's communication network. These measures are the

egocentric analogues to the sociocentric measures used in study 1. In order to control for the individual-level differences and correct for non-independence common to network samples, I employ fixed-effects estimates. The findings indicate that individuals systematically channel non-corrupt and corrupt information differently through their ego-networks. First, for individuals the corrupt information is not significant for tie-strength (0.603; $p < 0.279$), however it is positive, as we would expect given the findings in the first study. Next, corrupt information does increase asymmetry in the egocentric networks (0.323; $p < 0.108$), meaning the communication is less reciprocal for corrupt email communication. Two primary benefits arise from reciprocity in communication. The first is simply instrumental by allowing individuals to exchange and clarify information (March and Simon 1958). Second, reciprocity helps to engender trust between parties (Molm, Schaefer, and Collett 2007). Finally, corruption also reduces transitivity for individuals in corrupt projects (-0.328; $p < 0.152$). Transitivity refers to the tendency of two individuals, who both share a connection to a common third person, to also become tied to each other (Davis, 1963; Feld, 1981; Holland and Leinhardt 1971). Generally, transitivity is optimal for sharing information effectively, encouraging cooperation, and reducing conflict because information can easily be relayed from A to B and then to C. Thus, individual decisions about what information to share and with whom to share it appear to be influenced by content.

In this study, I show the effects of information on both egocentric and sociocentric structures. Specifically, the results support the role of content specification in social network research. Given these findings, it is clear that disaggregating networks by information content presents new opportunities to better understand the link between social structure and individual behavior. By investigating networks by information type, future research will gain insight into the individual-level dynamics of how ties form and how networks are reproduced. To date social network research has examined social structure as a plumbing system, containing and directing the flow of information, rather than the natural watercourses that carve the riverbeds and canyons of social relations.

REFRENCES

Brass, D., Butterfield, K., & Skaggs, B. 1998. Relationships and unethical behavior: A social network perspective. ***Academy of Management Review,*** 23(1): 14-31.

Davis, J. A. 1963. Structural Balance, Mechanical Solidarity, and Interpersonal Relations. ***American Journal of Sociology***, 68: 444-462.

Feld, S. 1981. The focused organization of social ties. ***American Journal of Sociology***, 86(5): 1015-1035.

Goffman, E. 1970. ***Strategic Interaction***.

Granovetter, M. 1973. The strength of weak ties. ***American Journal of Sociology***, 78(6): 1360-1380.

Holland, P., & Leinhardt, S. 1976. Local Structure in Social Networks. ***Sociological Methodology***, 7(1): 1-45.

G. March, J., & Simon, Alexander H. 1958. ***Organizations***.

Molm, L., Schaefer, D., & Collett, J. 2007. The Value of Reciprocity. ***Social Psychology Quarterly***, 70(2): 199-217.

Robins, G., & Alexander, M. 2004. Small worlds among interlocking directors: Network structure and distance in bipartite graphs. ***Computational & Mathematical Organization***, 10(1).

Simmel, G. 1950. ***The Sociology of Georg Simmel***.

Social & Spatial Network Dynamics in the U.S. House of Representatives

Clio Andris[1], Frank Hardisty[2]

[1] MIT, Department of Urban Studies and Planning & Sense*able* City Lab, Cambridge, MA
[2] The Pennsylvania State University, Department of Geography & GeoVISTA Center, University Park, PA

[clio@mit.edu], [hardisty@psu.edu]

1 INTRODUCTION

Noting a lack of effective methods for simultaneously analyzing datasets with *social* and *spatial* components, we build a system (Social/Spatial, or *S/S*) for the interactive exploration of dynamic networks as they are linked with map representations. We illustrate the uses of our software environment with an example analysis which uncovers hidden patterns in the spatial, social and temporal aspects of the U.S. legislative system, via a social network of Congressional Roll Call Votes in the U.S. House of Representatives.

Roll call votes have been analyzed for statistical patterns in previous studies. [1] In two different studies, Porter et al employ network science techniques like hierarchical clustering and modularity to model the system of House committees and subcommittees, in terms of shared members, and in terms of party majorities on each committee. [2] [3] We approach the U.S. Congressional Representative social network in a less constrained manner, as we do not concentrate on formal groupings, but informal connections visible through similar voting patterns. Some have noted that this 'informality' though seemingly unconstrained, does not show autonomy in representative decision making, but that representatives are driven by affiliated party, and the level of party alignment (as shown by voting records) has varied over time. [4] [5]

While literature on congressional social ties all seem to point to clusters of relationships in the U.S. House of Representatives, few studies have attributed these relationships to similarities in geography or in the demographic make-up of constituents. One such study showed that funding for statewide projects in congress favored the states with many representatives, but did not mention a pattern of intra-state members having closer relationships. [6] In terms of constituent demographics—an inherently topological feature—district representatives and their friendships are rarely causally linked to the similar nature of their constituent district demographics. One example is study of partisanship links low-home ownership to certain representatives. [7] There are certainly complex interactions that drive decision-making and relationships in the House, which represents the largest social network in our three branches of federal government; We can only imagine the social "balancing act" between pleasing constituents, sponsoring bills, interacting with lobbyists, following party agendas, creating trust networks for communication, collaboration, shared ideas and initiatives, negotiating provisions, and maintaining one's own sense of ethics and orthopraxy.

Harkening back to the agenda set by [2] [3], we build a social network of congressional representatives based on their voting preferences, in order to analyze the holistic network behavior with 'new' methods for complex network analysis. [8] Our study is novel in that is brings together a social network and the geographic districts and demographic constituencies that these network actors represent—so that their decisions and behavior can be analyzed in terms of hidden correlations in friendship, geographic clustering and demographic homophily.

To find these correlations, we use a custom tool, *Social/Spatial,* which links a force-directed, agent-based dynamic network with an interactive U.S. Congressional District map (Fig. 1), enabling instant visual knowledge discovery. This software package draws on previous development efforts, using the open source *S/S* (for geographic visualization and analysis) and JUNG (for network visualization and analysis) libraries as core components of *S/S*. Normally, we assume that nearby places have similar ideologies, but using *S/S* we are able to see the friendships, loyalties, and disagreements between people in places that we may have never thought had anything in common.

2 DATASET AND NETWORK

We create a network based on the number of agreements between two unique congress people

during the first session of the 111th Congress, in 2008, with 445 total members. Edges are created by tallying 985 "Roll Call" votes, where an agreement is constituted as two *Yays* or two *Nays* between two agents. A disagreement is recorded if an agent voted opposite another agent, or if one or both of the agents declined to vote. (The latter make up a relatively small percentage of the dataset.)

The weighted edge distribution of our network depicts a bimodal distribution that is most probably driven by polarization of the Democratic and Republican Party affiliation in the House. (Fig. 2) It is clear that there are pairs of people where their voting pattern and ideologies can be considered either 'clashing' or 'cooperating'. Therefore we mark clashing pairs with a 1, and cooperating pairs with a 5 in our network. Henceforth, the term ***friendship*** (and ***friend***) will be used to describe the relationship between congresspeople who show aligning interests: specifically, those in the 660 – 985 agreeing votes range. We operate on the theoretical basis that stronger ties indicate a relationship where information is transferred faster, ideologies are shared, trust is deeper, opinions are heard and there might be a stronger ability to convince each another to vote a certain way—whether implicitly or explicitly.

3 RESULTS

We incorporate the following deterministic metrics into the system. The following social network measures are implemented under four major themes:

Theme A is comprised of a group of *community detection* metrics like cliques (maximally complete subgraphs), hierarchical clusters (groups that form within larger groups), [9-11] modularity measures (the propensity of a member of a partition to talk to people in the partition more than to another singular partition).[12,13] **Theme B** incorporates *popularity* measures like, degree centrality (The number of nodes adjacent to a given node), betweenness centrality (the number of times a node occurs on all system geodesic paths), flow betweenness (the contribution of a node to all possible maximum flows), [14-16] and hubs & authorities (A high hub node connects to many good authorities and a high authority node receives from a number of strong hubs)**.** [17, 18] **Theme C** models *spreading processes*, like time-step propagation, resistance to percolation, cascades and rule-based diffusion—like voter, gossip, SIS, SIR and SISR models. [19-21] **Theme D** represents *attachment behavior*, namely, homophily (how much an ego's attaches to its alters based on a specified attribute, or the correlation between ego attributes and alter attributes) [22] and system clustering coefficients (the density of an agent's open neighborhood). [21]

These network measures are not the end of our quantitative evaluative statistics, as we have not mentioned temporal or spatial trends. First, the congressional data lends itself to temporal dynamic network, as a mostly static number of districts and representatives act on different issues each year. By using the *New York Times* Congressional Roll Call Vote API, we are able to feed multiple years into the system for exploring temporal change.

Secondly, the geographic trends are quantified in two ways. First is the spatial position of competing or cooperating districts, their geographic adjacency, proximity and propensity to form cohesive regions. [23] The tightness of these district clusters are measured with spatial statistics *Moran's I* or *Geary's C* for statistically significant spatial clusters of behavior groups [24, 25], Hot & Cold Spot Detection, and LISA [26]. In addition to measuring proximal regions, demographic (feature) clustering shows the correlation between certain constituent social features and U.S. Census information like of Income, Urban Areas, Racial Percentages, or areal designations like Economic Development Regions, or Statewide Alliances. These underlying features of a geographic district or socio-demographic constituency are correlated with the relative party-loyal, maverick, authority, cliquey, or neutral behavior of the congressman in the wide congressional social network.

4 CONCLUSIONS

The overarching goal of *S/S* is to provide a workspace for the interactive exploration of dynamic networks as they are linked with map representations. Our system is robust enough to accommodate temporal changes in network agents and relationships for archives of past Congresses. With access to each congressperson's friends, alters, social group, propensity to agree with other certain agents, or to unconventionally float on the periphery of the network, we can see how the 435 unique social groups in the United States different economic, demographic, and localities are relating to one another through their representative. We enumerate some other examples of how social networks and spatial relationships can illustrate each other, and in particular, how spatial

relationships can illustrate friendship and agreement behaviors that would otherwise be difficult to observe.

5 REFERENCES

[1] J. Clinton, S. Jackman and D. Rivers, The statistical analysis of roll call data, *American Political Science Review* **98** (2004).

[2] M. A. Portera, P.J. Muchab, M.E.J. Newman, A.J. Friendd, Community structure in the United States House of Representatives, *Physica A* **386** (2007), 414–438.

[3] M.A. Porter, P.J. Mucha, M. E. J. Newman, and C. M. Warmbrand, A network analysis of committees in the U.S. House of Representatives, *PNAS* **102** (2005), 7057–7062.

[4] F. E. Lee, Geographic Politics in the U.S. House of Representatives: Coalition building and distribution of benefits, *American Journal of Political Science* **47** (2003), 714–728.

[5] G.W. Cox and K.T. Poole, On Measuring partisanship in roll-call voting: The U.S. House of Representatives, 1877-1999, *American Journal of Political Science* **46** (2002), 477-489.

[6] J. M. Snyder, Jr. and Tim T. Grose, Estimating party influence in Congressional roll-call voting *American Journal of Political Science* **44** (2000), 193-211.

[7] D. Macrae, Jr., The relation between roll call votes and constituencies in the Massachusetts House of Representatives *The American Political Science Review* **46** (1952), 1046-1055.

[8] D.J. Watts, The new science of networks, *Annual Review of Sociology* **30** (2004), 243 – 270.

[9] C. Bron and J. Kerbosch, Finding all cliques of an undirected graph, *Comm of the ACM* **16** (1973), 575-577.

[10] S.P. Borgatti, M.G. Everett and L.C. Freeman, Ucinet for Windows: Software for Social Network Analysis, *Analytic Technologies*, Harvard, Massachusetts, (2002).

[11] S.C. Johnson, Hierarchical clustering schemes, *Psychometrika* **32** (1967), 241-253.

[12] M.E.J. Newman, Detecting community structure in networks, *Eur Phys J B* **38** (2004), 321–330.

[13] M.E.J. Newman, Modularity and community structure in networks, *PNAS* **103** (2006), 8577–8582.

[14] L. C. Freeman, Centrality in Social Networks: Conceptual clarification, *Social Networks* **1** (1979), 215-239.

[15] L. C. Freeman, S. P. Borgatti and D.R. White, Centrality in valued graphs: A measure of betweenness based on network flow, *Social Networks* **13** (1991), 141-154.

[16] P.V. Marsden, Egocentric and sociocentric measures of network centrality, *Social Networks* **24** (2002), 407—422.

[17] H. Ibarra and S.B. Andrews, Power, social influence, and sense making: Effects of network centrality and proximity on employee perceptions, *Administrative Science Quarterly* **38** (1993), 277—303.

[18] J. M. Kleinberg, Authorative sources in a hyperlinked environment, *Journal of the ACM*, (1999).

[19] D.J. Watts, A simple model of global cascades on random networks, *PNAS*, **99** (2002), 5766.

[20] R. Cowan and N. Jonard, Network structure and the diffusion of knowledge, Journal of Economic Dynamics and Control **28** (2004), 1557-1575.

[21] M. O. Jackson, Social and Economic Networks. Princeton, New Jersey, Princeton University Press, (2008).

[22] M. McPherson, L. Smith-Lovin, and J. Cook, Birds of a Feather: Homophily in social networks, *Annual Review of Sociology* **27** (2001), 415-444.

[23] N. Cressie, Statistics for Spatial Data, Terra Nova: *John Wiley and Sons,* (1992).

[24] P.A.P. Moran, Notes on continuous stochastic phenomena, *Biometrika* **37** (1950), 17–33.

[25] R. C. Geary, The contiguity ratio and statistical mapping, *The Incorporated Statistician* **5** (1954), 115–145.

[26] L. Anselin, LISA Local Indicators of Spatial Autocorrelation, *Geographical Analysis* **27** (1995), 93-115.

6 FIGURES

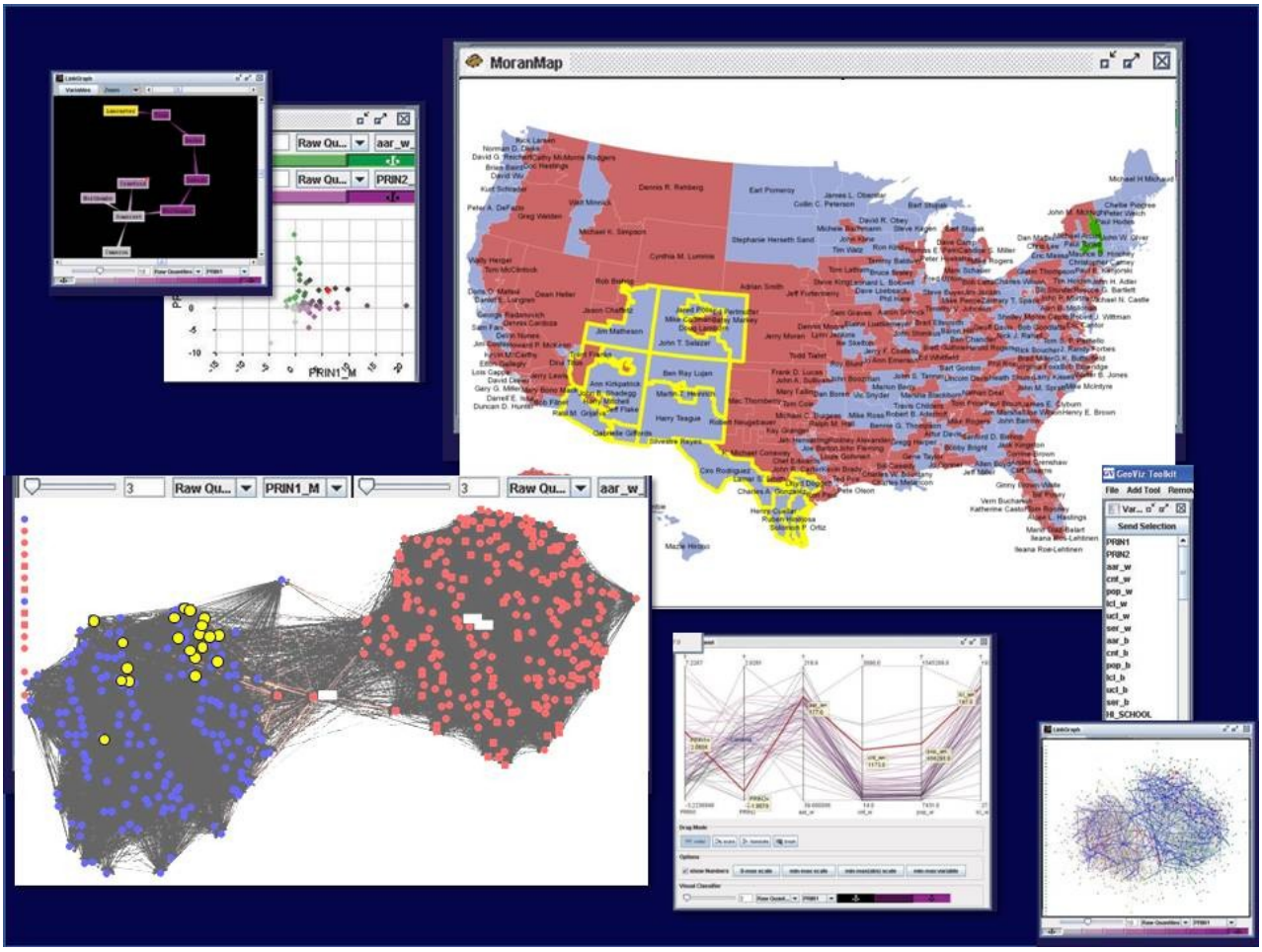


Figure 1 (top): The java-based GeoViz Social analysis system allows for the simultaneous and dynamic exploration of social networks through interactive highlighting (see yellow), statistical and geographic representations of the relationships between congressmen, and their corresponding constituencies.

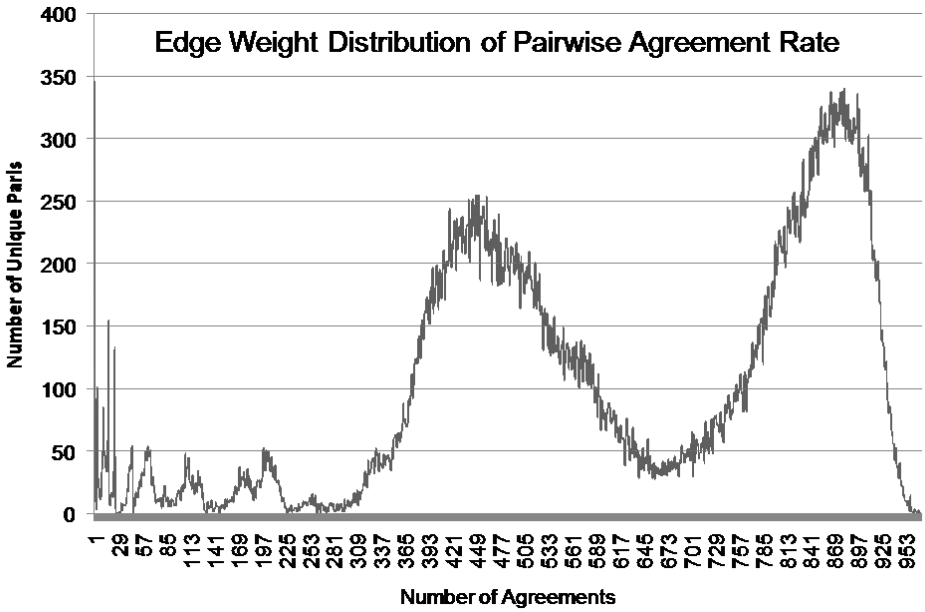


Figure 2 (left): The allocation of edge weight values among all node pairs depicts a bimodal distribution.

Poster Session Abstracts

Topology Discovery of Sparse Random Graphs With Few Participants

Animashree Anandkumar
EECS Dept.,
University of California,
Irvine, CA 92697
a.anandkumar@uci.edu

Avinatan Hassidim
Google Research Tel Aviv,
23 Menachem Begin street,
Tel Aviv, Israel
avinatanh@gmail.com

Jonathan Kelner
CSAIL & Dept. of Math.
Massachusetts Inst. of Tech.
Cambridge, MA 02139
kelner@mit.edu

ABSTRACT

We consider the task of topology discovery of sparse random graphs using end-to-end random measurements (e.g., delay) between a subset of nodes, referred to as the participants. The rest of the nodes are hidden, and do not provide any information for topology discovery. We consider topology discovery under two routing models: (a) the participants exchange messages along the shortest paths and obtain end-to-end measurements, and (b) additionally, the participants exchange messages along the second shortest path. For scenario (a), our proposed algorithm results in a sub-linear edit-distance guarantee using a sub-linear number of uniformly selected participants. For scenario (b), we obtain a much stronger result, and show that we can achieve consistent reconstruction when a sub-linear number of uniformly selected nodes participate. This implies that accurate discovery of sparse random graphs is tractable using an extremely small number of participants. We finally obtain a lower bound on the number of participants required by any algorithm to reconstruct the original random graph up to a given edit distance. We also demonstrate that while consistent discovery is tractable for sparse random graphs using a small number of participants, in general, there are graphs which cannot be discovered by any algorithm even with a significant number of participants, and with the availability of end-to-end information along all the paths between the participants.

Categories and Subject Descriptors

G.2.2 [**Mathematics of Computing**]: Discrete Mathematics—*Graph Theory.*

General Terms

Algorithms, Theory.

Keywords

Topology Discovery, Sparse Random Graphs, End-to-end Measurements, Hidden Nodes, Quartet Tests.

SIGMETRICS'11, June 7–11, 2011, San Jose, CA, USA.

1. INTRODUCTION

Inference of global characteristics of large networks using limited local information is an important and a challenging task. The discovery of the underlying network topology is one of the main goals of network inference, and its knowledge is crucial for many applications. For instance, in communication networks, many network monitoring applications rely on the knowledge of the routing topology, e.g., to evaluate the resilience of the network to failures [34, 40]; for network traffic prediction [25, 51] and monitoring [9], anomaly detection [7], or to infer the sources of viruses and rumors in the network [47]. In the context of social networks, the knowledge of topology is useful for inferring many characteristics such as identification of hierarchy and community structure [27], prediction of information flow [5,53], or to evaluate the possibility of information leakage from anonymized social networks [11].

Traditionally, inference of routing topology in communication networks has relied on tools such as traceroute and mtrace [3] to generate path information between a subset of nodes. However, these tools require cooperation of intermediate nodes or routers to generate messages using the Internal Control Message Protocol (ICMP). Increasingly, today many routers block traceroute requests due to privacy and security concerns [30, 54], there by making inference of topology using traceroute inaccurate. Moreover, traceroute requests are not scalable for large networks, and cannot discover layer-2 switches and MPLS (Multi-protocol Label Switching) paths, which are increasingly being deployed [43].

The alternative approach for topology discovery is the approach of *network tomography.* Here, topology inference is carried out from end-to-end packet probing measurements (e.g., delay) between a subset of nodes, without the need for cooperation between the intermediate (i.e., non-participating) nodes in the network. Due to its flexibility, such approaches are gaining increasing popularity (see Section 1.2 for details).

The approach of topology discovery using end-to-end measurements is also applicable in the context of social networks. In many social networks, some nodes may be unwilling to participate or cooperate with other nodes for discovering the network topology, and there may be many hidden nodes in "hard to reach" places of the network, e.g., populations of drug users, and so on. Moreover, in many networks, there may be a cost to probing nodes for information, e.g., when there is a cash reward offered for filling out surveys. For

Social Network-Based Interventions

Weihua An
Department of Sociology
Institute for Quantitative Social Science
Harvard University
March 9, 2011

Social network-based interventions (SNI) can target at least five different facets of social networks: environment and contexts of social networks, structures of social networks, processes (e.g., diffusion, learning, and searching) on social networks, meanings and trust attached to social ties, and strategies of networking. In this study, I will focus on how to utilize social network features, more specifically, how to choose central nodes or tight groups in a social network, to accelerate or consolidate the effects of any hypothetical intervention.

In practice, such SNIs have a lot of applications in areas such as public health, social marketing, commercial delivery, or even military operations. For examples, we may want to choose central nodes (i.e., students) in a classroom to act as opinion leaders to accelerate the diffusion of some positive information, attitudes or behaviors that are desirable to us researchers, or we may want to choose certain central cities on a transportation network as hubs to optimize the efficiency of a delivery system, and lastly, we many want to choose certain cities on a railroad network as military objectives to implement targeted attacks and maximize the efficacy of military offense.

In this study, I will first review existing methods in choosing central nodes and tight groups in a social network, pointing out their contributions and limitations, and then introduce the algorithms I have developed for that use. For example, previous methods in choosing central nodes tend to either generate nodes that are redundant with one another or ignore the direction of social ties, only applicable to undirected social networks. In this study, I develop an algorithm to address both issues and provide more options than previous ones to select central seed nodes, e.g., by indegree (for the purpose of influence), by outdegree (for the purpose of spreading information) or by both. Same importantly, the new algorithm, compared with previous greedy optimization algorithm which is not guaranteed to find an optimal set of central nodes, has been confirmed to find at least one optimal or very close to optimal set of central nodes. In short, my algorithm starts with a random set of nodes and then adds the next least

redundant node one by one until reaching the designed size of seed nodes. The above process can be repeated many times to achieve better optimization.

How to extract groups from a social network for group-based interventions is another challenge. Previous community (or group) detection approaches either ignore the directionality of social networks or tend to choose groups not from a sociological perspective but only from a data mining perspective and so often produce unmeaningful groups. In contrast, in this study I propose a theory-driven approach. First, I rewire a social network by keeping only symmetrical ties so that only strong relationships are retained in the network. Then I extract tight cliques (i.e., a maximal set of nodes in which all nodes have connections to each other) from the network. The last step is to pick cliques so that the total number of nodes being selected matches with the designed sample size and that any node being selected has at least one connection with others being selected. This will ensure no one will be treated as isolate in any group-based intervention. The last step is not trivial, as the last node to be selected must come from the neighborhood of the already selected nodes. In my view, this approach guarantees generating meaningful social groups and utilizing group pressure in a maximally positive way. Both this group selection algorithm and the above central nodes selection algorithm have been developed and incorporated in an R module “SNIP: Social Network-Based Interventions & Policies” and will be geared and released for public use soon.

Last to illustrate, I have applied the above algorithms to a health education program involving around 90 classes and 4500 students that aim to accelerate the diffusion of positive information, attitudes and behaviors regarding cigarette smoking among adolescents.

Parsimonious Algorithms for Decentralized Ranking in Social Networks

Kyomin Jung
KAIST
kyomin@kaist.edu

Boyoung Kim
KAIST
combicola@kaist.ac.kr

Milan Vojnović
Microsoft Research
milanv@microsoft.com

We consider decentralized algorithms where each node in a network aims at computing an aggregate quantity of all node states using only local information without any centralized agency. Recently, a large amount of interest on this type of algorithms has been arisen in various contexts such as social networks [1, 2], Internet [3, 4], and biological systems [5], because they are not only useful to explain phenomena observed in a network system, but also useful for the design of new computation protocols. Since many network systems inherently contain restrictions on memory and communications (i.e. parsimonious), designing a decentralized algorithm under such restrictions is an important problem. There have been many studies trying to account for these restrictions. For example, in the context of averaging algorithms, randomized gossip algorithms based on reversible Markov chains [6] have been considered as well as averaging algorithms based on non-reversible Markov chains [7]. For the averaging problem, also the effects of quantization of messages exchanged between nodes have been studied [8, 9].

In this paper, we study a *rank aggregation problem* where the goal is to rank a set of alternatives in decreasing order of users' preference in a decentralized manner. A specific example is voting over a set of alternatives, which frequently arises in social networks including surveys of consumer preferences. The goal is to identify a list of top k popular products in decreasing order of their popularity. Such cooperative decision making problem arises in a variety of applications such as in surveys of preference in social networks, decentralized database systems, and sensor networks.

The main contributions of our work are in (1) allowing for arbitrary number of alternatives $m \geq 2$, and (2) algorithms for ranking that are based on computing a generalized version of the *mode*, and (3) allowing for user preference across a set of alternatives. For computing the mode in network systems where each node prefers one out of two alternatives, the classical voter model [10–12] has been extensively studied. Algorithms for binary consensus were proposed to serve as an improvement of the voter model with respect to the error probability and the convergence speed to the correct consensus [13–15]. A quantized version of the gossip algorithm was suggested to identify the quantization interval containing the average value [8]. Using this algorithm, the majority voting problem can be solved with only four states per node. However, these works are restricted to the case of two alternatives.

The detailed setup that we consider is as follows: we consider a network system that consists of nodes $[n] = \{1, 2, \ldots, n\}$ where $n \geq 1$ and a finite set of alternatives $[m] = \{1, 2, \ldots, m\}$ where $m \geq 2$. The preference of each node $j \in [n]$ over alternatives is described by the vector of ranking scores $\vec{v_j} = (v_1, v_2, \ldots, v_m)$ where $v_i \geq 0$ and $\sum_{i=1}^{m} v_i = 1$. A vector of ranking scores $\vec{v}_j$ is such that the i-th coordinate of this vector represents preference of node j for alternative i. A *top-k ranking* is a tuple of alternatives $(a_1, a_2, \ldots, a_k)$, for $k \leq m$, such that $a_i \in [m]$ for every i, and $U(a_1) \geq U(a_2) \geq \ldots \geq U(a_m)$ where $U(a_i)$ is the sum of ranking scores for alternative a_i over all nodes. The ranking problem is to construct a decentralized algorithm which ensures that every node computes a top-k ranking correctly after finitely many number of iterations.

First, we propose an algorithm that computes the full ranking of alternatives for any connected network graph by generalizing the discretized averaging algorithm [8]. Our algorithm runs in a time

equivalent to the mixing time of the corresponding random walk in the network. Our algorithm uses $2^{m(m-1)}$ states per node. Although this algorithm runs correctly on any connected network graph, it is not parsimonious in the required memory per node. Next, we present parsimonious algorithms for the mode computation and the top-k ranking computation for the case of small k.

Our mode computation algorithm is described as follows. As the behavior of bloggers, at each time step, a randomly chosen node observes another node chosen uniformly at random and updates its preference state. The main idea of the update rule is the introduction of two extra states (weak and strong) for each alternative $j \in [m]$. Based on this idea, we prove that this algorithm converges correctly with probability of error that diminishes exponentially with the total number of nodes; this result is established using mean field arguments along with a concentration inequality for random processes.

Finally, we propose an efficient algorithm for computing a top-k ranking. This algorithm starts with assigning to each node a random *k-ranking state* $(b_1, b_2, \ldots, b_k)$, where $b_i \in [m]$ for every i, according to a probability that depends on the ranking score vector of the node. At this step, $m(m-1)\ldots(m-k+1) = O(m^k)$ many k-ranking states are needed. We prove that the problem of computing the mode among a set of k-ranking states is equivalent to the problem of computing the top-k ranking on the set of the original alternatives. Using our mode computation algorithm, we prove that the error probability of our top-k ranking algorithm decays exponentially with the total number of nodes. We examine convergence of our algorithms using simulations.

References

[1] J. Kleinberg, "The small-world phenomenon and distributed search," in *SIAM News*, vol. 37, 2004.

[2] D. Acemoglu, M. A. Dahleh, I. Lobel, and A. Ozdaglar, "Bayesian learning in social networks," in *Review of Economic Studies*, 2010.

[3] J. Kleinberg, "Complex networks and decentralized search algorithms," in *Proc. of the International Congress of Mathematicians*, 2006.

[4] F. Vega-Redondo, A. Galeotti, S. Goyal, M. Jackson, and L. Yariv, and A. Steger, "Network games," in *Review of Economic Studies*, vol. 77, 2010.

[5] F. Kuhn, K. Panagiotou, J. H. Spencer, and A. Steger, "Synchrony and asynchrony in neural networks," in *Proc. of ACM-SIAM SODA*, 2010.

[6] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," in *IEEE Trans. on Information Theory*, vol. 52, pp. 2508-2530, 2006.

[7] K. Jung, D. Shah, and J. Shin, "Distributed averaging via lifted Markov chains," in *IEEE Trans. on Information Theory*, vol. 56, pp. 634-647, 2010.

[8] F. Benezit, P. Thiran, and M. Vetterli, "Interval consensus: from quantized gossip to voting," in *Proc. of IEEE ICASSP*, 2009.

[9] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," in *IEEE Trans. on Automatic Control*, vol. 54, pp. 2506-2517, 2009.

[10] P. Donnelly and D. Welsh, "Finite particle systems and infection models," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 94, pp. 167-182, 1993.

[11] D. Aldous and J. A. Fill, "Reversible Markov chains and random walks on graphs," in *Monograph in Preparations*, 1999.

[12] J. Cruise and A. Ganesh, "Probabilistic consensus via polling and majority rules," in *Proc. of Allerton Conference*, 2010.

[13] E. Perron, D. Vasudevan, and M. Vojnović, "Using three items for binary consensus on complete graphs," in *Proc. of INFOCOM*, 2009.

[14] M. Draief and M. Vojnović, "Convergence speed of binary interval consensus," in *Proc. of INFOCOM*, 2010.

[15] E. Mossel, G. Schoenebeck, "Reaching consensus on social networks," in *International Journal of Intelligent Systems - Decision Making in Social Networks*, vol. 25, 2010.

Dual approaches to network science

Patrick O. Perry (patperry@seas.harvard.edu)
Patrick J. Wolfe (patrick@seas.harvard.edu)
Harvard University

There are two complementary approaches to extracting information from a social network. The first approach is structure-based: it looks for interesting structural features (nodes, sub-networks, diffusion properties, etc.), where "interesting" is defined according to some external criterion. The second approach is model-based: it attempts to gain insight into the dynamics driving a network's formation by treating it as a random instantiation from a parametrized stochastic process. It turns out that in many situations, the two approaches are duals of each other. Characterizing the structural features of a network corresponds to estimating the parameters of a random network model. In particular, degree centrality and modularity are closely related to two specific random network models.

Degree centrality

Degree centrality measures the importance of a node with respect to the rest of the network; it provides a way of ranking a network's nodes [3, 5]. Suppose we are given an undirected network with n nodes, encoded by an $n \times n$ symmetric binary matrix $\boldsymbol{X}$. Degree centrality d_i measures the importance of node i by

$$d_i = \sum_{j \neq i} X_{ij},$$

where X_{ij} is the ij element of $\boldsymbol{X}$. The degree centrality of node i is a count of the number of edges with i as an endpoint.

Degree centrality turns out to be related to a random network model parametrized by node specific importance parameters $\alpha_1, \ldots, \alpha_n$. The model, $\mathcal{M}^{(0)}$, considers the edges to appear randomly, where edge $i \sim j$ appears independently of all others with probability p_{ij} specified as

$$\mathcal{M}^{(0)} : \operatorname{logit} p_{ij} = \alpha_i + \alpha_j.$$

Under this model, if $\boldsymbol{X}$ is sparse enough, then the maximum likelihood estimate of the ith node-specific parameter is

$$\hat{\alpha}_i = \hat{\alpha}_0 + \log d_i + \mathcal{O}(1/n),$$

where $\hat{\alpha}_0$ is a constant depending on $\boldsymbol{X}$, and the remainder term is often negligible. Notably, $\hat{\alpha}_i$ is essentially a monotone transformation of d_i.

Modularity

Modularity and its scale-dependent generalization measure the quality of a structural grouping of the nodes in a network; maximizing modularity gives rise to a natural node partition [1, 2, 4]. As above, suppose we are given an undirected network with n nodes, encoded by an $n \times n$ symmetric binary matrix $\boldsymbol{X}$. A partition of the n nodes into K groups can be encoded by vector $\boldsymbol{g} = (g_1, \ldots, g_n)$, where g_i is the group of node i and $g_i \in \{1, \ldots, K\}$. Modularity $Q_\gamma(\boldsymbol{g})$ measures the quality of partition $\boldsymbol{g}$ by

$$Q_\gamma(\boldsymbol{g}) = \sum_{i<j} (X_{ij} - \gamma\, p_{ij}^{(0)})\, \delta(g_i, g_j),$$

where $\gamma \geq 1$ is a scale parameter, $p_{ij}^{(0)}$ is a "null" estimate of the probability of edge $i \sim j$ appearing in the absence of community structure, and δ is the Dirac delta function. The modularity of node

partition $\boldsymbol{g}$ is motivated as a sum of residuals between the observed withtin-community edges and the null expected within-community edges.

Modularity turns out to be related to a random network model parametrized by group membership parameters $g_1, \ldots, g_n$ and positive strength factor λ. The model, $\mathcal{M}$, considers the edges to appear randomly, where edge $i \sim j$ appears independently of all others with probability p_{ij} specified as

$$\mathcal{M} : \operatorname{logit} p_{ij} = \operatorname{logit} p_{ij}^{(0)} + \lambda\, \delta(g_i, g_j).$$

Under this model, if $\max_{ij}\{p_{ij}^{(0)}\}$ is small enough, the log-likelihood is

$$\ell(\boldsymbol{g}, \lambda) = \ell_0 + \lambda\, Q_\gamma(\boldsymbol{g}) + \mathcal{O}(\lambda^2),$$

where ℓ_0 is a constant depending on $\boldsymbol{X}$ and $\gamma = \gamma(\lambda) = [\exp(\lambda) - 1]/\lambda$, and the remainder term is often negligible. Notably, $\ell(\lambda, \boldsymbol{g})$ is essentially a monotone transformation of $Q_\gamma(\boldsymbol{g})$.

Implications

The explicit relationship between degree centrality and model $\mathcal{M}^{(0)}$ gives us a parametric interpretation of d_i. Likewise, the explicit relationship between modularity and model $\mathcal{M}$ gives us a parametric interpretation of γ and $Q_\gamma(\boldsymbol{g})$. With these interpretations, we can leverage standard statistical machinery to incorporate prior information about node importance and node partitions, we can construct confidence intervals and perform hypothesis tests related to these quantities, and we can employe model selection tools like AIC and BIC to choose the number of groups, K. Taking the dual perspective, we can gain insight into how degree centrality and modularity are related to the dynamics driving network formation. The two approaches to extracting information from networks—structure-based and model-based—turn out to be closely related, and both approaches can profit from this relationship.

References

[1] R. Lambiotte, J.-C. Delvenne, and M. Barahona. Laplacian dynamics and multiscale modulare structure in networks. arXiv:0812.1770v3 [physics.soc-ph], 2009.

[2] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

[3] J. Nieminen. On the centrality in a graph. *Scand. J. Psychol.*, 15(1):332–336, 1974.

[4] J. Reichardt and B. Stefan. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, 2006.

[5] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

Degree Distributional Metric Learning

Bert Huang, Blake Shaw, and Tony Jebara
Department of Computer Science, Columbia University, New York, NY 10027
{bert, blake, jebara}@cs.columbia.edu

Introduction. Real-world networks often consist of nodes with informative attributes as well as links. To properly model these networks, it is necessary to learn how attributes of the nodes relate to the connectivity structure. Metric learning is a natural framework for transforming the raw node features to match the structural properties of a graph. Traditional metric learning algorithms primarily model the similarity between nodes and not structural properties, such as degree distributions. Degree distributions play a central role in graph structure analysis [1]. The degree distribution for some nodes may be non-stationary and depend on their attributes, particularly if some attributes naturally relate to connectedness. For example, in the LinkedIn network, an individual whose job area is "Software Sales" is likely to have more connections than an individual whose area is "Software Programmer". We propose *degree distributional metric learning* (DDML), a method for simultaneously learning a metric and degree preference functions such that the combination captures the structure of the input graph and allows for more accurate link prediction from only node features.

Algorithm Description. The learning algorithm is given training data consisting of N pairs of node feature vectors and corresponding adjacency matrices $\{(\mathbf{X}^1, \mathbf{A}^1), \ldots, (\mathbf{X}^N, \mathbf{A}^N)\}$, where each row i of $\mathbf{X}^k \in \mathbb{R}^{n_k \times D}$ represents one of n_k D-dimensional real-valued node feature vectors denoted by $(\mathbf{x}_i^k)^\top$, and each $\mathbf{A}^k \in \mathbb{B}^{n_k \times n_k}$ is a directed adjacency matrix. DDML then outputs a similarity function $f : \{\mathbb{R}^D, \mathbb{R}^D\} \mapsto \mathbb{R}$ that takes two vectors as input and outputs a real value, and a degree preference function $g : \{\mathbb{R}^D, \mathbb{N}\} \mapsto \mathbb{R}$ takes a node descriptor vector and a candidate degree d and outputs a real valued preference score for that node having degree d.

Matrices $\mathbf{M} \in \mathbb{R}^{D \times D}$, and $\mathbf{T}, \mathbf{S} \in \mathbb{R}^{n \times D}$, where $n = \max_k n_k$, define a degree distributional metric. The similarity function is [1] $f(\mathbf{x}_i, \mathbf{x}_j; \mathbf{M}) = \mathbf{x}_i^\top \mathbf{M} \mathbf{x}_j$. Using the notation that $\mathbf{s}_c$ is the $1 \times D$ dimensional c'th row of $\mathbf{S}$, the degree preference function is $g(\mathbf{x}_i^k, b; \mathbf{S}) = \sum_{c=1}^{n_k - b} \mathbf{s}_c \mathbf{x}_i^k$. A graph is predicted by maximizing $F(\mathbf{A}|\mathbf{X}^k, \mathbf{M}, \mathbf{S}, \mathbf{T}) = \sum_{ij|A_{ij}=1} f(\mathbf{x}_i^k, \mathbf{x}_j^k; \mathbf{M}) + \sum_i g\left(\mathbf{x}_i^k, \sum_j A_{ij}^k; \mathbf{S}\right) + \sum_j g\left(\mathbf{x}_j^k, \sum_i A_{ij}^k; \mathbf{T}\right)$. This optimization is computable by a reduction to a maximum weight b-matching [3]. Using normalized Hamming distance $\Delta(\mathbf{A}^k, \tilde{\mathbf{A}}) = \sum_{ij|A_{ij}^k \neq \tilde{A}_{ij}} 1/(n_k^2 - n_k)$, i.e., the proportion of misclassified edges, as a loss function and the Frobenius ℓ_2-norm of the parameter matrices as a regularizer, learning is done by solving

$$\min_{\mathbf{M},\mathbf{S},\mathbf{T},\xi \geq 0} \frac{1}{2}\left(||\mathbf{M}||_{\text{Fro}} + ||\mathbf{S}||_{\text{Fro}} + ||\mathbf{T}||_{\text{Fro}}\right) + C\xi, \text{ s.t.}$$

$$\frac{1}{N}\sum_{k=1}^{N}\left[F(\mathbf{A}^k|\mathbf{X}^k, \mathbf{M}, \mathbf{S}, \mathbf{T}) - F(\tilde{\mathbf{A}}^k|\mathbf{X}^k, \mathbf{M}, \mathbf{S}, \mathbf{T})\right] \geq \frac{1}{N}\sum_{k=1}^{N}\Delta(\mathbf{A}^k, \tilde{\mathbf{A}}) - \xi, \quad \forall\{\tilde{\mathbf{A}}^1, \ldots, \tilde{\mathbf{A}}^n\}. \quad (1)$$

The optimization is a quadratic program with exponentially many linear constraints, and is of the same form as a *structural support vector machine* (SVM) [4]. Thus, the established cutting-plane approach (and efficiency guarantees) can be applied. The solution to (1) is found by maintaining a working set of constraints, solving for the optimal $\mathbf{M}$, $\mathbf{S}$, and $\mathbf{T}$, then adding the worst-violated constraint by the current solution and repeating. The worst violated constraint is found by a *separation oracle*, $\tilde{\mathbf{A}}^k = \text{argmax}_{\mathbf{A}} F(\mathbf{A}|\mathbf{X}, \mathbf{M}, \mathbf{S}, \mathbf{T}) + \Delta(\mathbf{A}^k, \mathbf{A})$. This is computed by adding the decomposed loss to the primary edge weights of the b-matching input.

Experiments. We consider comparisons against two baseline models of varying richness. The simplest model classifies node-pairs using a support vector machine (SVM). The SVM receives training data as pairs of inputs and outputs (binary labels) $\{[\mathbf{x}_i^k(1)\mathbf{x}_j^k(1), \ldots, \mathbf{x}_i^k(D)\mathbf{x}_j^k(D)], (\mathbf{A}^k)_{ij}\}$, and then estimates a weight vector $\mathbf{w}$. The second model, $\mathbf{M}$-learning, learns a linear transform matrix $\mathbf{M}$ without any degree information, predicting the presence an edge if $\mathbf{x}_i \mathbf{M} \mathbf{x}_j$ is a positive quantity. We learn $\mathbf{M}$ by the same optimization as DDML except with the $\mathbf{S}$ and $\mathbf{T}$ matrices fixed at zero. The SVM, $\mathbf{M}$-learning and DDML approaches bring increasing model richness. The SVM approach is equivalent to learning an $\mathbf{M}$ matrix that is only nonzero along the diagonal. Similarly, $\mathbf{M}$-learning is equivalent to DDML with no degree distribution information.

[1] If $\mathbf{M}$ is positive semi-definite (PSD), it can be used to describe a metric. Omitting the PSD requirement allows the similarity function to be asymmetric, which allows representation of directed graphs. Nevertheless, we always refer to the algorithm as degree distributional metric learning.

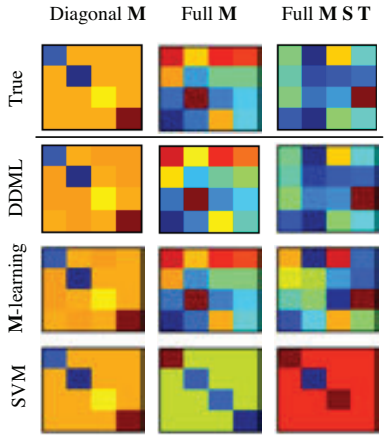


Figure 1: True and Learned **M** matrices from Synthetic Tests. For each sampling scheme, the models that do not learn the active sampling parameters inadequately model the feature interactions.

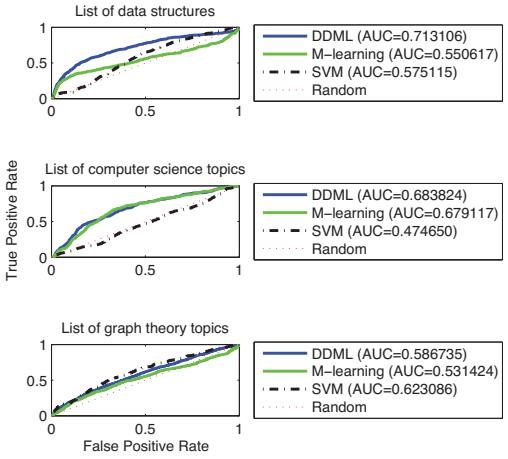


Figure 2: ROC Curves for Wikipedia Link Prediction. These plots compare the true and false positive rates of the rankings returned by the three learning algorithms for predicting the edges of held out graphs.

Synthetic Graphs. We generate data and graphs from three sampling schemes. For each scheme, we train on five graphs and test on five new graphs. To generate graphs, we randomly sample 50 data vectors from $\mathbb{R}^4$ uniformly from $[0, 1]^4$ and predict different **M**, **S** and **T** matrices. First, we use a diagonal **M** matrix and zero degree preference. Second, we use a full **M** matrix and zero degree preference. Third, we generate a random **M** matrix and random **S** and **T** matrices. All methods perform well when data is generated from their corresponding models, but the baselines fail on graphs from the richer generative processes. E.g., in the third scheme, only DDML predicts near-perfectly. See Fig. 1.

Wikipedia Lists. We conducted an experiment to predict the link structure between Wikipedia articles in predefined categories using bag-of-words features for each article. For each category, we collected the count of word-occurrences in articles listed on the main category page and directed links between the articles within each category. We squash the word counts with the square root function and reduce dimensionality to 20 by applying non-negative matrix factorization [2]. We train the algorithms on categories "linear algebra topics", and "mathematical functions", and test on "computer science topics", "data structures", and "graph theory topics". DDML predictions produce an average F_1-score of 0.1255, **M**-learning scores 0.0930, and SVM scores 0.0534 and a fully-connected graph scores 0.0561.

We also compared the ranking of edges obtained by the three models. Since the DDML model is richer than a simple ranking of edges, we greedily select edges in order according to the gain in current overall weight, which takes into account the edge weight itself as well as the reward for the change in degree induced by adding each edge. For **M**-learning and SVM, the ranking is the ordering of prediction values. The receiver order statistics (ROC) curves for each of the held-out test graphs are in Fig. 2.

Discussion. Metric learning is a natural framework for modeling graph data containing both connectivity information and node attributes. We have demonstrated that it is insufficient to learn only a metric that defines node similarity merely pairwise. To properly model how nodes connect, it is necessary to estimate both a metric and a set of degree preference functions which allow the model to better match the structural properties of real networks.

References

[1] Barabási, A. Linked: The new science of networks. *J. Artificial Societies and Social Simulation*, 6(2), 2003.

[2] Berry, M., Browne, M., Langville, A., Pauca, V., and Plemmons, R. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155 – 173, 2007.

[3] Huang, B. and Jebara, T. Exact graph structure estimation with degree priors. In *ICMLA*, pp. 111–118, 2009.

[4] Joachims, T., Finley, T., and Yu, C. Cutting-plane training of structural svms. *Mach. Learning*, 77(1):27–59, 2009.

Visualizing Social Networks with Structure Preserving Embedding

Blake Shaw and Tony Jebara
Department of Computer Science, Columbia University, New York, NY 10027
{blake, jebara}@cs.columbia.edu

1 Introduction

We propose an adaptation to Structure Preserving Embedding (SPE) based on stochastic gradient decent that allows for visualization of large social network datasets. SPE finds a low-dimensional representation of nodes in a network which is *structure-preserving*, meaning a connectivity algorithm such as k-nearest neighbors will recover the original connectivity pattern of the network exactly from only the coordinates of the nodes in the low-dimensional embedding. There are many possible goals for network visualization algorithms, such as minimizing edge crossings, bringing neighbors close, pushing away unconnected nodes, highlighting clusters, and preserving graph distances. We propose that accurate visualizations of social networks should preserve the underlying topological structure of the network. In previous work, we have presented Structure Preserving Embedding (SPE) [3], an algorithm based on semidefinite programming and singular value decompositions designed to find such embeddings. In this abstract, we present a low-rank approximation to the original algorithm, implemented using a fast custom solver based on projected stochastic gradient descent, which allows the technique to scale to larger networks.

2 Algorithm

Given a network of n nodes represented as a graph with adjacency matrix $\mathbf{A} \in \mathbb{B}^{n \times n}$, SPE finds an embedding $\mathbf{L} \in \mathbb{R}^{d \times n}$ such that d is small and running a connectivity algorithm such as k-nearest neighbors on $\mathbf{L}$ returns $\mathbf{A}$. As first proposed, SPE learns a matrix $\mathbf{K}$ via a semidefinite program (SDP) and then decomposes $\mathbf{K} = \mathbf{L}^\top \mathbf{L}$ by performing singular value decomposition. In contrast, this article proposes optimizing $\mathbf{L}$ directly. Although for $d < N$, this problem is now non-convex, because of the stochastic nature of the optimizer we have found the algorithm does not suffer from local minima in practice.

SPE for greedy nearest-neighbor constraints solves the following SDP:

$$\max_{\mathbf{K} \in \mathcal{K}} \operatorname{tr}(\mathbf{K}\mathbf{A})$$
$$D_{ij} > (1 - A_{ij}) \max_m (A_{im} D_{im}) \ \forall_{i,j}$$

where $D_{ij} = K_{ii} + K_{jj} - 2K_{ij}$ and $\mathcal{K} = \{\mathbf{K} \succeq 0,\ \operatorname{tr}(\mathbf{K}) \leq 1, \sum_{ij} K_{ij} = 0\}$. The constraints require the embedding of each node to be more distant from its non-neighbors than its neighbors. Let $S = \{\mathbf{C}_1, \mathbf{C}_2, ...\mathbf{C}_m\}$ be the set of all triplet constraints, where each $\mathbf{C}_l$ is a constraint matrix corresponding to a triplet (i, j, k) such that $A_{ij} = 1$ and $A_{ik} = 0$. This set of all triplets clearly subsumes the distance constraints above, and allows each individual constraint to be written as $\operatorname{tr}(\mathbf{C}_l \mathbf{K}) > 0$ where $\operatorname{tr}(\mathbf{C}_l \mathbf{K}) = K_{jj} - 2K_{ij} + 2K_{ik} - K_{kk}$. Temporarily dropping the centering and scaling constraints, we can now formulate the SDP above as maximizing the following objective function over $\mathbf{L}$:

$$f(\mathbf{L}) = \lambda \operatorname{tr}(\mathbf{L}^\top \mathbf{L}\mathbf{A}) - \sum_{l \in S} \max(\operatorname{tr}(\mathbf{C}_l \mathbf{L}^\top \mathbf{L}), 0).$$

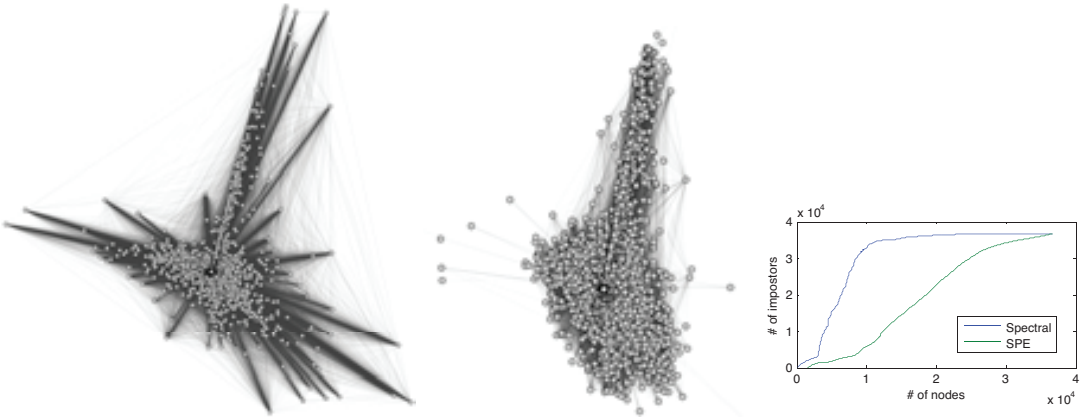
Note that we have introduced a Lagrange multiplier λ as an additional parameter which trades-off between the loss term and regularization term. We will maximize $f(\mathbf{L})$ via projected stochastic subgradient decent. Define the subgradient in terms of a single randomly chosen triplet:

$$\Delta(f(\mathbf{L}), \mathbf{C}_l) = \begin{cases} 2\mathbf{L}(\lambda \mathbf{A} - \mathbf{C}_l) \text{ if } \operatorname{tr}(\mathbf{C}_l \mathbf{L}^\top \mathbf{L}) > 0 \\ 0 \ \text{ otherwise} \end{cases}$$

and for each randomly chosen triplet constraint C_l, if $\operatorname{tr}(\mathbf{C}_l \mathbf{L}^\top \mathbf{L}) > 0$ then update $\mathbf{L}$ according to:

$$\mathbf{L}_{t+1} = \mathbf{L}_t + \eta \Delta(f(\mathbf{L}_t), \mathbf{C}_l)$$

where the step-size $\eta = \frac{1}{\sqrt{t}}$. After each step, we can use projection to enforce that $\operatorname{tr}(\mathbf{L}^\top \mathbf{L}) \leq 1$ and $\sum_{ij} (\mathbf{L}^\top \mathbf{L})_{ij} = 0$, by subtracting the mean from $\mathbf{L}$ and dividing each entry of $\mathbf{L}$ by its Frobenius norm. $\mathbf{L}$ is initialized either randomly or from the solution of spectral embedding or Laplacian eigenmaps [1]. The



(a) Spectral embedding (b) SPE-SGD (c) Impostors

Figure 1: The Enron email network embedded into 2D by spectral embedding (a), and SPE-SGD (b). The plot on the right shows how many nodes have fewer than x impostors. We see that embedding this network into 2D yields many impostors; however on average nodes in the SPE embedding have many fewer impostors than nodes in the spectral embedding.

algorithm terminates when $|\mathbf{L}_{t+1} \quad \mathbf{L}_t| < \epsilon$, where ϵ is an input parameter.

In practice, instead of optimizing over a single randomly chosen triplet at each iteration, we find it useful to randomly select a node at each iteration, and use the gradient computed from all *impostor* triplets, since it is only for these triplets that a gradient step is taken. As shown in Figure 2 an impostor is a node which violates the neighborhood of another node. For each impostor triplet $\{i, j, k\}$, i is the randomly chosen target node, j is the furthest connected neighbor of i and k is a node unconnected to i but currently closer than j.

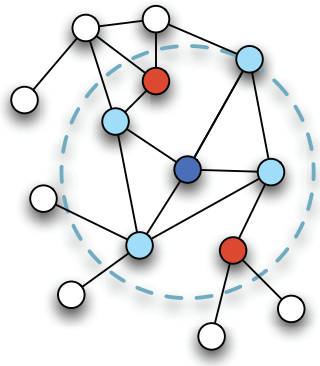


Figure 2: The red nodes are identified as impostors to the neighborhood of the center node (dark blue), because the impostors (red) are closer than the furthest of the connected nodes (light blue).

3 Experiments

In Figure 1 we see two embeddings of the Enron email network [2]. Each of the 36692 nodes in the network represents a person, and there exist edges between each pair of people who have communicated via email. Because of the high degree of many of the nodes in the network, it is likely impossible to find a 2D embedding which preserves topology exactly – meaning all nodes have zero impostors. The network may require a higher dimensional embedding. However we see that the 2D visualization produced by SPE has far fewer impostors than that produced by spectral embedding, and thus provides a more accurate visualization.

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2002.

[2] J. Leskovec, J Kleinberg, and C Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proc. of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 2005.

[3] B. Shaw and T. Jebara. Structure preserving embedding. In *Proc. of the 26th International Conference on Machine Learning*, 2009.

Ranking: Compare, Don't Score

Ammar Ammar
ammar@mit.edu
LIDS - MIT

Devavrat Shah
devavrat@mit.edu
LIDS - MIT

Abstract

The need to rank items based on user input arises in many practical applications such as elections, betting, and recommendation systems. Consider, for example, the popular movie rental website, Netflix, which is faced with the challenge of having to recommend movies to users based on partial historical information about their preferences. A popular approach to this challenge is to ask users to provide explicit numerical ratings of the movies they have watched. The ratings are then used to obtain the desired ranking. The main appeal of the rating-based methods is that they are fairly easy to implement. However, the rating scale and the individual ratings are often arbitrary and may not be consistent from one user to another. Furthermore, each user rates only few movies, which could lead to the "loss" of valuable information. A more natural alternative to numerical ratings relies on asking users to compare pairs of movies. These comparisons provide an "absolute" indicator of the user's preference, but it is often hard to combine comparisons from different users to obtain a consistent ranking.

In this work, we provide a general and tractable framework for utilizing comparison data for the purpose of ranking. In this framework, comparisons are treated as partial samples from a distribution over permutations. Using the Principle of Maximum Entropy, we devise a concise parameterization of one such distribution using only $O(n^2)$ parameters, where n is the number of items in question. We also propose a distributed algorithm for estimating the parameters, and producing rankings over the items in question. Finally, we provide results from an experiment where our aglorithm was useds as advisory component in the review process for the The ACM International Symposium on Mobile Ad Hoc Networking and Computing.

1. Problem Satement and Main Results

We consider a set $\mathcal{N} = \{1, ..., n\}$ of n items. We represent the preferences of each user by a permutation σ of the elements of $\mathcal{N}$. We denote the position of item i in the user's preference by $\sigma(i)$, and say that the user prefers item i to item j iff $\sigma(i) > \sigma(j)$. In our model, user preferences come from a distribution over permutations $\mu : S_n \to [0, 1]$, where S_n is the set of all permutations of n elements.

For the simplicity of presentation, suppose that the available data for each pair of distinct items consists of the fraction of users who prefer item i to item j, denoted by w_{ij}. Suppose also that each w_{ij} is the marginal of some underlying distribution (i.e. $w_{ij} = \mathsf{P}[\sigma(i) > \sigma(j)]$. Then the set of distributions consistent with our data, can be defined using the following constraints:

$$\sum_{\sigma \in S_n} \mu_\sigma \mathbb{I}_{\{\sigma(i) > \sigma(j)\}} = w_{ij}, \ \forall i, j \tag{1a}$$

$$\mu_\sigma \geq 0, \ \forall \sigma \in S_n \tag{1b}$$

where $\mathbb{I}_{\{E\}}$ denotes the indicator variable for event E. It is easy to see that there are multiple distributions that satisfy these constraints, and one has to come up with a criterion to choose among these distributions. Ideally, the selected distribution should utilize the information provided by the data, without imposing any additional structure on unseen data. One criterion that achieves this objective is that of Maximum Entropy, where we choose a distribution that fits the data while maximizing the entropy. Using the method of Lagrange Multipliers, we obtain a distribution of the form:

$$\mu_\sigma \propto \exp\Big(\sum_{i,j\in\mathcal{N}} \lambda_{ij}\mathbb{I}_{\{\sigma(i)>\sigma(j)\}}\Big) \tag{2}$$

One appealing property of this distribution is that it can be fully characterizes using only $O(n^2)$ paremeters that can be estimated by solving the following dual problem:

$$\max_{\lambda} \sum_{i,j} \lambda_{ij} w_{ij} - \sum_{\sigma\in S_n} \exp\Big(\sum_{i,j} \lambda_{ij}\mathbb{I}_{\{\sigma(i)>\sigma(j)\}}\Big) \tag{3}$$

The second term in the objective above, known as the partition function, involves summing over an exponential ($n!$) space, and is usually hard to compute exactly. To address this issue, we propose a distributed algorithm that uses MCMC techniques to compute the partition function and other marginal. In our algorithm, the parameters λ are maintained and updated separately, and the marginals corresponding to each pair of items are computed approximately when needed. Furthermore, the estimates produced by the algorithm are asymptotically consistent, and converge to the true model, under the assumption that $0 < w_{ij} < 1$ for all distinct items i, j.

Once the distribution μ is obtained, our algorithm can be used to estimate different marginals for the purpose of ranking. For example, we could obtain a ranking by assigning each item a score that reflects its average position according to our distribution. For example, the score for item i could be:

$$\mathsf{E}_\mu[e^{\theta(n-\sigma(i))}] = \sum_{r=1}^{n} e^{\theta(n-r)} P[\sigma(i) = r] \tag{4}$$

where the parameter $\theta > 0$ could be used to adjust the importance of the top part of the ranking.

2. Evaluation

We used our algorithm and ranking scheme, outlined in the previous section, as an advisory component to the technical committee for The ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2011). In addition to the scores traditionally provided by the reviewers, we asked reviewers to provide pairwise comparisons of their papers. We then used our algorithm to rank the papers, and the result were compared with the final decision made by the committee. Despite the limited amount of available data, the top 20 papers recommended by our scheme contained 9 of the 20 accepted papers, and the remaining 11 received high rankings. This is surprising, given the fact that the data used was fairly "thin", due to the high paper-to-reviewer ration. Moreover, the final decision was made by a committee over the course of a two day meeting held at MIT, and our algorithm had no access to any information from meeting.

References

[1] Cynthia Dwork and Ravi Kumar and Moni Naor and D. Sivakumar. *Rank Aggregation Revisited*. In Proceedings of WWW10, 2001.

[2] Yiling Chen and Evdokia Nikolova and Lance Fortnow. *Betting on permutations*. In ACM Conference on Electronic Commerce (2007).

[3] Jonathan Huang and Carlos Guestrin and Leonidas Guibas. *Inference for Distributions over the Permutation Group*. Machine Learning Department, Carnegie Mellon University (2008).

Community detection with fuzzy community structure

Qinna Wang[1], and Eric Fleury[1]

[1]LIP ENS-LYON
D-NET INRIA
Université de lyon
France,
qinna.wang@ens-lyon.fr, eric.fleury@inria.fr

Abstract. In real networks, communities may overlap. That is the problem of the cover which allows the nodes to be shared among communities. Currently, the most accepted and widely used measurement to evaluate the quality of the community structure is the modularity proposed by Newman and Girvan. Although it fails to evaluate the quality of the covers, it is suitable to uncover the fuzzy community structure which allows the nodes to have memberships with more than one group. Reichardt et al. [1] has proposed an energy landscape survey method, which uncovered the fuzzy community structure with a co-appearance matrix by collecting the spin configuration with the minima local energy. We use a similar method to uncover the fuzzy community structure with a co-appearance matrix. Here, the co-appearance matrix whose element shows the probability of nodes in the same community is calculated by running Louvain algorithm [2] several times. With the co-appearance matrix, the membership of nodes is revealed by the high co-appearance in off-diagonal. Comparing to the method proposed by Reichardt et al., the Louvain algorithm is more efficient than the simple Monte-Carlo heat-bath algorithm, and setting random orders of nodes is much simpler than tuning the temperature.

We also propose a new extension of the modularity to evaluate the quality of the covers. Combined with the Hamiltonian in forms of the cohesion and the adhesion, the quality of the cover is derived from the union of the communities with overlapping nodes, which is,

$$Q_{ov} = \frac{1}{2M} \sum_{i \neq j} \left(A_{ij} - \frac{k_i k_j}{2M}\right) \frac{|d_i \cap d_j|}{|d_i \cup d_j|}. \quad (1)$$

where d_i, d_j denote the memberships of i and j in the graph, $d_i, d_j \subseteq \{\mathcal{C}_1, \ldots, \mathcal{C}_q\}$.

Since the overlapping nodes make the classification of nodes into the partition undecidable, a strict overlapping community definition is proposed in [3]. That is, if removing the node group l from the community $\mathcal{C}_\alpha$ to the community $\mathcal{C}_\beta$ will not degenerate the quality of the total graph, we consider the node group l to be the overlapping group between $\mathcal{C}_\alpha$ and $\mathcal{C}_\beta$. Furthermore, the quality of the cover should not be lower than the partition. We can find the cover by modularity optimization. For the efficiency of the partition detection algorithm, it is a good choice to find the cover by adjusting the fuzzing node groups based on the partition by improving the quality of the community structure. In this way, we can find the cover $\mathcal{S}$ with a good quality: $Q_{ov}(\mathcal{P}) \leq Q_{ov}(\mathcal{S})$.

Furthermore, benefited from the hierarchical structure of Louvain algorithm, the hierarchal structure with the fuzzy community structure can be mined, too. Although it seems that the hierarchical structure can be provided by tuning the resolution parameter γ [4],

more studies have shown the relation between the stability and the parameter γ. Therefore, it is meaningful to study the relationship between the fuzzing nodes and γ. Both of them are related with the stability of communities.

Having applied our method on benchmark graphs, the results have demonstrated the availability of our method in fuzzy community detection, the overlapping community detection, the hierarchical structure detection and the community structure with the different resolution parameters γ. The results show that the fuzzing node groups which cause the instability of communities can reveal how small clusters are nested in large clusters for the modularity resolution limitation, with which the size of the community is in the order of $\sqrt{m}$, where m is the total number of links of the graph. The hierarchical organization with fuzzing nodes is suitable to study the stability of communities at different levels. With different resolution parameters γ, we observe that the fuzzing nodes join the small clusters into large clusters by minimizing the parameter γ. Studies on a geographical co-citation network whose points denote the laboratories in different regions and the edges denote the number of articles cooperated between laboratories, are very interesting. Results have shown that the laboratories with few articles may degenerate the robust of the communities. A good example is the community composed of the laboratories in UK. Its instability is related with the laboratories located in few population regions. And the laboratories in Belgium having good collaborations with its neighbor countries are also shown in the overlapping community structure. For the feasibility of our method, we hope it is helpful in community detection.

References

[1] Stefan Reichardt, Joerg Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. 93,:218701, 2004.

[2] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:8, October 2008.

[3] Stefan Reichardt, Joerg Bornholdt. Statistical mechanics of community detection. 74:016110, 2006.

[4] Peter Ronhovde and Zohar Nussinov. Local resolution-limit-free potts model for community detection. *Physical Review E*, 81,:046114, March 2008.

A new Leaders-followers algorithm for detecting overlapping communities in social networks

Rushed Kanawati
LIPN CNRS UMR 7030, University Paris Nord, 93430 Villetaneuse, FRANCE
rushed.kanawati@lipn.univ-paris13.fr

1 Introduction

Complex networks exhibit a mesoscopic level of organization, called *communities* [NG04]. A community is a connected sub-graph whose nodes are much linked with one each other that with nodes out-side the sub-graph. Nodes in a community, are generally supposed to share common properties or play similar roles within the network. This suggests that we can gain much insight into the complex networked systems by discovering and examining their underlaying communities. A great number of algorithms have been proposed for detecting the community structure in complex networks. Examples are [BGLL08; NG04; SZ10; BGMI05; ZW07; DWWW08; LM09; Mur10]. Different algorithms apply different techniques. However a major trend of these algorithms are *modularity* guided approaches [NG04]. Amazingly, maximizing the modularity seems not to be a guarantee to detect real communities in a network. For example, considering the now well known Zachary's Karate club dataset [Zac77] the modularity of the partition composed of the two real communities forming the club is only 0.37. The *Louvain* approach, which is one of the top community detection algorithms in the state of the art [BGLL08], computes a different partition with a higher modularity of 0.41. However, compared to the real partition the obtained purity is rather low (0.64). This suggests that modularity might not be the right criteria to optimize.

We propose here a new modularity independent algorithm for community detection that is inspired from human community formation. A community is led by a set of core leaders followed by other community members. A single node can belong to different communities at once allowing to compute overlapping communities. A leader may also follow another leaders of another communities. The proposed algorithm is structured into two mains steps : identifying nodes in the network that are playing the role of communities' leaders, then assigning other nodes to leaders in order to construct the communities. Our algorithm computes automatically the number of communities to identify. Moreover it identifies overlapping communities rather than disjoint ones as most of existing algorithms do.The idea of constructing communities around leader nodes has been recently applied in two independent and different algorithms described in [KCZ10] and [SZ10]. Both cited algorithms are designed to compute a network partition rather than overlapping communities. In addition the [KCZ10] requires the number of communities to identify as an input. Next we sketch briefly the outlines of our proposal and show first results obtained by applying the approach to some small benchmark networks.

2 Leaders-centered community detection algorithm

The basic idea underlaying the proposed algorithm is that a community is composed of two types of nodes: *Leaders* and *Followers*. Roughly speaking, leaders form a subset of nodes (eventually one node) whose removal form the network implies community collapse. Algorithm 1 sketches the general outlines of the proposed approach. The algorithm functions as follows. First we extract from the set of the graph nodes a list $\mathcal{L}$ of nodes that are likely to be leaders of communities. This is the role of the $isLeader()$ function (line 3). A leader is a node that has greater *centrality* that its neighbors, whatever the applied centrality is.

Algorithm 1 Leader-based community detection algorithm

Require: $G =< V, E >$ a connected graph
1: $\mathcal{L} \leftarrow \emptyset$ {set of leaders}
2: **for** $v \in V$ **do**
3: **if** $isLeader(v)$ **then**
4: $\mathcal{L} \leftarrow \mathcal{L} \cup \{v\}$
5: **end if**
6: **end for**
7: $\mathcal{C} \leftarrow computeComumunitiesLeader(\mathcal{L})$
8: **for** $v \in V$ **do**
9: **for** $c \in \mathcal{C}$ **do**
10: $M[v, c] \leftarrow membership(v, c)$ {see equation 1}
11: **end for**
12: $P[v]$ = **sortAndRank**$(M[v])$
13: **end for**
14: **for** $v \in V$ **do**
15: $P^*[v] \leftarrow \mathbf{rankAggregate}_{\mathbf{x} \in \{\mathbf{v}\} \cap \mathbf{\Gamma_G(v)}} \mathbf{P}[\mathbf{x}]$
16: */* assigning v to communities */*
17: **for** $c \in P^*[v]$ **do**
18: **if** $|M[v, c] - M[v, P^*[0]]| \leq \epsilon$ **then**
19: $COM(c) \leftarrow COM(c) \cup \{v\}$
20: **end if**
21: **end for**
22: **end for**
23: **return** $\mathcal{C}$

In the current implementation, a leader is defined as a node whose degree centrality is greater or equal to $\sigma \in [0, 1]$ percent of its neighbors. This allows to recover leaders connected to other leaders. Nodes in $\mathcal{L}$ are then grouped into sets; each composing a set of leaders of a community. This is the role of the function $computeComumunitiesLeader()$ (line 7). The length of the list of sets $\mathcal{C}$ is the number of communities in the network. Two leader nodes are grouped in the same community if the ratio of common neighbors to the total number of neighbors is above a given threshold $\delta \in [0, 1]$. Next, each node (leaders and followers) computes its membership degree to each of the identified communities. This is done by applying the $membership()$ function (line 10). Again different implementations can be proposed for this function. In the current version, the membership degree of node v to a community c is given by the inverse of the minimal shortest path that links v to one of the leaders of c.

$$membership(v, c) = \frac{1}{(min_{x \in COM(c)} SPath(v, x)) + 1} \quad (1)$$

Notice that for a community c, the membership of all its leaders is equal to 1. The membership vector of each node v is sorted in decreasing order in order to compute ranked list of membership preference vector P_v (line 12). Then, for each node we compute a permutation of P_v in a way the aggregates preference vectors of all its direct neighbors (line 15). Different social choice algorithms can then be applied to compute P_v^* [CELM07]. In the current implementation of the algorithm we apply the classical Borda approach [Bor81]. Lastly, a node v is assigned to top ranked communities in P_v^* for which membership degree is at most within $\epsilon \in [0, 1]$ from the membership degree of the first ranked community. ϵ in another system parameter. $COM(v \in V)$ returns the community in $\mathcal{C}$ led by v if it exists, $\emptyset$ otherwise (line 19).

3 Experimental results

We apply our approach to two well known *small* benchmark networks for which ground truth community decompositions are available: *Zachary karate club* [Zac77] and the sawmill strike movement dataset [1]. The first network is composed of 34 nodes divided into two communities while the second is composed of 36 nodes linked with 63 edges and decomposed into three communities. While both datasets are very small compared to target networks of our approach, they have the advantage of offering a ground truth decomposition into communities and have been used by different community identification approaches. We compare performances of our approach with basic community detection algorithms: The Newman-Girvan algorithm [NG04] and the *Louvain* algorithm [BGLL08]. Input parameter of our approach are mainly the three thresholds used in the three steps described earlier. In this experimentation we have fixed the ϵ parameter to 0 hence assigning a node to different communities only if the node has an equal membership degree. This choice is mainly motivated by the fact we compare our approach with algorithms detecting disjoint communities. The σ threshold used for identifying leader nodes has been varied from 0.7 to 1. Lastly the δ threshold used in grouping leaders into groups is fixed to 0.5. Table 1 gives obtained results on both datasets using classical clustering quality metrics: purity and the ARI index [YR01]. These preliminary results shows clearly that the modularity metric does not correspond to the best decomposition into communities as measured by both purity and adjusted rand index. For instance, the *Louvain* method obtain always the best modularity (even better than the modularity of the ground truth decomposition) however it is ranked last according to purity. Best results are obtained by our approach for high values of σ. This suggests that heuristics we propose is able to detect precisely real leaders in these social networks.

[1] available on http://vlado.fmf.uni-lj.si/pub/networks/pajek/ensa/strik

4 Conclusion

In this work we have proposed a new approach for community detection in complex networks based on identifying a set of community leaders then assigning nodes to these leaders. This approach presents the advantage of applying node neighborhood bounded computation steps in order to conduct the three of steps of leader detecting, community number identification and node assignments to communities. This allows to consider developing a full distributed version of the algorithm that can handle very large scale networks. Results obtained on small benchmark social network argue for the capacity of the approach to detect real communities.

References

[BGLL08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks. *CoRR*, abs/0803.0476, 2008.

[BGMI05] Jeffrey Baumes, Mark K. Goldberg, and Malik Magdon-Ismail. Efficient identification of overlapping communities. In Paul B. Kantor, Gheorghe Muresan, Fred Roberts, Daniel Dajun Zeng, Fei-Yue Wang, Hsinchun Chen, and Ralph C. Merkle, editors, *ISI*, volume 3495 of *Lecture Notes in Computer Science*, pages 27–36. Springer, 2005.

[Bor81] J.C. Borda. Mémoire sur les élections au scrutin. *Comptes rendus de l'Académie des sciences, traduit par Alfred de Grazia comme Mathematical Derivation of a election system , Isis, vol 44, pp 42-51*, 1781.

[CELM07] Yann Chevaleyre, Ulle Endriss, Jérôme Lang, and Nicolas Maudet. A short introduction to computational social choice. In Jan van Leeuwen, Giuseppe F. Italiano, Wiebe van der Hoek, Christoph Meinel, Harald Sack, and Frantisek Plasil, editors, *SOFSEM (1)*, volume 4362 of *Lecture Notes in Computer Science*, pages 51–69. Springer, 2007.

[DWWW08] Nan Du, Bai Wang, Bin Wu, and Yi Wang. Overlapping community detection in bipartite networks. In *Web Intelligence*, pages 176–179. IEEE, 2008.

[KCZ10] Reihaneh Rabbany Khorasgani, Jiyang Chen, and Osmar R. Zaiane. Top leaders community detection approach in information networks. In *4th SNA-KDD Workshop on Social Network Mining and Analysis*, Washington D.C., July 2010.

[LM09] Xin Liu and Tsuyoshi Murata. Community detection in large-scale bipartite networks. In *Web Intelligence*, pages 50–57. IEEE, 2009.

[Mur10] Tsuyoshi Murata. Detecting communities from tripartite networks. In Michael Rappa, Paul Jones, Juliana Freire, and Soumen Chakrabarti, editors, *WWW*, pages 1159–1160. ACM, 2010.

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physics review E*, 69:026113:1–022613:15, 2004.

[SZ10] D. Shah and T. Zaman. Community detection in networks: The leader-follower algorithm. In *Workshop on Networks Across Disciplines in Theory and Applications, NIPS*, November 2010.

[YR01] Ka Yee Yeung and Walter L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[Zac77] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452473, 1977.

[ZW07] R.S. Zhang and X.S. Wang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A*, 374:483490, 2007.

Table 1: Comparison of performances of different community detection algorithms

	Zacahary dataset				Strike dataset			
Approach	# Communities	Purity	ARI	Modularity	# Communities	Purity	ARI	Modularity
Newman	2	0.97	0.87	0.36	3	0.90	0.82	0.35
Louvain	4	0.64	0.83	0.41	5	0.68	0.80	0.60
Our approach $\sigma = 0.7$	3	0.74	0.59	0.23	4	0.68	0.5	0.20
Our approach $\sigma = 0.8$	4	0.95	0.93	0.30	3	1.0	0.60	0.45
Our approach $\sigma = 0.9$	2	1.0	1.0	0.37	3	1.0	1.0	0.55
Our approach $\sigma = 1.0$	2	1.0	1.0	0.37	3	1.0	1.0	0.55

Estimation of Dynamic Social Network Structure via Online Convex Programming

Maxim Raginsky, Corinne Horn, and Rebecca Willett

I. Introduction

Consider a dynamic social network composed of p agents, where the designation "dynamic" means that the network topology or, more broadly, the influence of each agent on the other agents, may evolve with time. We collect sequential observations of the agents' actions, and would like to use them to infer something in real time about the structure of the network. We assume that each agent has only two actions available to him. This assumption is valid in many contexts, e.g., voting, communication, or meeting patterns.

From the modeling point of view, both the topology of the network and the strengths of the agents' influence can be encoded in a time-varying sparse binary pairwise Markov random field (also known in statistical physics as the Ising model). Models of this type have been popular in statistical analysis of social networks [1], [2]. Thus, our goal is to infer the parameters of such a model given the observed binary actions of the agents. The offline (batch) version of this problem was recently treated by Banerjee et al. [3], Ravikumar et al. [4], Kolar and Xing [5], and Höfling and Tibshirani [6]. As noted in [3] and [4], sparsity regularization can play a critical role in accurately estimating network structure from a relatively limited amount of data; this effect has been noted both empirically and theoretically in an analysis of sample complexity of different methods.

By contrast, we work in the *online* setting, i.e., when the observations are available sequentially, one at a time. Moreover, even if all the observations are available at once, online algorithms often offer a computationally feasible and robust alternative to batch inference [7]. A notable feature of our approach is that we do *not* assume that the dynamical evolution of the agents' actions is, in fact, described by a time-varying Ising model; nor do we assume that sequential observations are conditionally independent. Rather, we treat the Ising model as a descriptive tool, which allows for easy interpretation of the observed network behavior in terms of a time-varying weighted graph, where the presence of a nonzero weight on an edge connecting a particular pair of agents indicates mutual influence, while the sign of the weight indicates the nature of the correlation between the agents' actions (positive or negative). Our theoretical results show that we can use the sequential observations of the agents' actions to construct a sequence of network structure estimates that is nearly as good as if all the observations had been available at once.

Moreover, if the observations do happen to be generated according to a time-varying Ising model, then we can combine the online estimates to form a "final" estimate and bound its generalization error in terms of both (a) the choice of regularization and (b) the penalty due to sequential, rather than batch, processing of observations. Recently, there has been increased interest in the role of regularization, especially sparsity regularization, in the context of online convex programming. In particular, Xiao [8], Langford et al. [9] and Duchi et al. [10] have demonstrated empirically that treating a regularization term separately in an online algorithm can help ensure sparse estimates at each time step. However, this line of research does not show how the choice of regularizer impacts generalization performance, despite its obvious importance empirically. The work described in this abstract helps close that gap.

The authors are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. E-mail: {m.raginsky, ceh23, willett}@duke.edu.

II. Ising models for network structure

Let $\mathcal{V} = \{1, \ldots, p\}$ denote the set of agents making up the network. We assume that the number of agents p is known, and that no new agents enter the network throughout the observation period. Let us first consider a single time instant. For each agent $\alpha \in \mathcal{V}$, let $x_\alpha \in \{-1, +1\}$ denote the action of that agent. Then the Ising model for the network configuration $x = (x_\alpha : \alpha \in \mathcal{V})$ is given by the probability distribution

$$\mathbb{P}_\theta(x) = \frac{1}{Z(\theta)} \exp\left\{ \sum_{\alpha,\beta \in \mathcal{V}} \theta_{\alpha\beta} x_\alpha x_\beta \right\}. \qquad (2.1)$$

The $p \times p$ matrix $\theta = (\theta_{\alpha\beta})_{\alpha,\beta\in\mathcal{V}}$ is symmetric ($\theta_{\alpha\beta} = \theta_{\beta\alpha}$); if $\theta_{\alpha\beta} = 0$ for some α and β, then we say that agent α has no influence on agent β and vice versa; on the other hand, if $\theta_{\alpha\beta} \neq 0$, then there is a correlation between the actions x_α and x_β, which is positive or negative depending on the sign of $\theta_{\alpha\beta}$. The normalization constant $Z(\theta)$ is known as the *partition function*.

The basic problem is as follows: Suppose we observe T "snapshots" of the network in the form $x^t = (x^t_\alpha : \alpha \in \mathcal{V}) \in \{-1, +1\}^{\mathcal{V}}$, $t = 1, \ldots, T$. We would like to come up with a corresponding sequence $\widehat{\theta}^t \in \mathbb{R}^{p\times p}$, $t = 1, \ldots, T$, such that the product distribution $\mathbb{P}_{\widehat{\theta}^1} \otimes \mathbb{P}_{\widehat{\theta}^2} \otimes \ldots \otimes \mathbb{P}_{\widehat{\theta}^T}$ provides a good "fit" to the observed data $x^{1:T} = (x^1, \ldots, x^T)$. Moreover, we would like the estimates $\widehat{\theta}^t$ to be *sparse*, i.e., only a small number of $\widehat{\theta}^t_{\alpha\beta}$ should be nonzero. This sparsity requirement meshes well with observational studies of certain decision-making phenomena in social networks, e.g., in the voting patterns of legislators.

The problem of inferring the network parameters θ of the Ising model (2.1) is, in general, intractable due to the presence of the partition function $Z(\theta)$ [11]. To circumvent the need for computing (or even approximating) the partition functions [4], [5], several authors, starting with Ravikumar et al. [4], have exploited the observation that, for each agent $\alpha \in \mathcal{V}$, the conditional distribution x_α given the actions of the remaining agents (denoted here by $x_{\backslash\alpha}$) is

$$\mathbb{P}_\theta(x_\alpha|x_{\backslash\alpha}) = \frac{\exp\left(2x_\alpha \sum_{\beta\in\mathcal{V}\backslash\alpha} \theta_{\alpha\beta}x_\beta\right)}{\exp\left(2x_\alpha \sum_{\beta\in\mathcal{V}\backslash\alpha} \theta_{\alpha\beta}x_\beta\right)+1}$$

where $\mathcal{V}\backslash\alpha$ denotes the set of all vertices excluding α. The negative log likelihood $-\log\mathbb{P}_\theta(x_\alpha|x_{\backslash\alpha})$ is a convex function of θ; thus, the vector of weights θ can be found by minimizing, separately for each vertex α, the ℓ_1-regularized negative log likelihood. The final graph estimate is then assembled from the individual vertex estimates. Alternatively, as discussed by Höfling and Tibshirani [6], the individual vertex terms $-\log\mathbb{P}_\theta(x_\alpha|x_{\backslash\alpha})$ can be combined into a *pseudolikelihood*

$$\ell(\theta;x) \triangleq -\sum_{\alpha\in\mathcal{V}} \log\mathbb{P}_\theta(x_\alpha|x_{\backslash\alpha}).$$

This avoids the need for reconciling the individual vertex estimates. Although there is no proof of consistency in [6], the empirical results are promising, and the method of [6] easily lends itself to an online implementation.

III. Our approach and summary of results

We recap our problem: given sequentially arriving co-occurrence observations $x^{1:T} = (x^1, x^2, \ldots, x^T)$, construct a sequence of sparse network graph estimates $\widehat{\theta}^{1:T}$, where $\widehat{\theta}^t$ may depend only on the currently available observations $x^{1:t-1} = (x^1, \ldots, x^{t-1})$. Since we do not assume anything about the dynamics of the agents' behavior, we take the *regret minimization* point of view: Given a regularization parameter $\tau > 0$, we consider the regularized pseudo-likelihood $\ell_\tau(\theta;x) = \ell(\theta;x) + \tau\|\theta\|_1$, and our goal is to ensure that the *regret*

$$R_T \triangleq \sum_{t=1}^T \ell_\tau(\widehat{\theta}^t;x^t) - \inf_{\theta\in\Theta}\sum_{t=1}^T \ell_\tau(\theta;x^t)$$

is as small as possible. Here, Θ is a convex feasible set of graphical model weights (e.g., an ℓ_1 ball). The big picture here is that the proposed online estimation scheme should compete favorably against the best ℓ_1-regularized pseudolikelihood estimator that has access to the entire data record $x^{1:T}$. For the problem of tracking a dynamically evolving network, we can consider a class $\Theta^{1:T}$ of time-varying reference models and define the *tracking regret*

$$R_T(\Theta^{1:T}) \triangleq \sum_{t=1}^T \ell_\tau(\widehat{\theta}^t;x^t) - \inf_{\theta^{1:T}\in\Theta^{1:T}}\sum_{t=1}^T \ell_\tau(\theta^t;x^t).$$

The goal in both cases is to ensure that the regret behaves sublinearly as a function of the record size T. In the stochastic case, sublinear regret will lead to consistency in the usual sense [7] — if the best ℓ_1-regularized batch estimator is consistent, then the sequence of outputs of any online algorithm with sublinear regret can be used to construct a final estimator which is nearly as good.

In a nutshell, our theoretical results are as follows: If we compare our causally constructed sequence of estimates $\widehat{\theta}^t$ to the best single offline estimate θ^*, then we obtain a regret bound of the form $R_T = O(\sqrt{T})$, where the constants implicit in the $O(\cdot)$ notation depend on the geometry of the reference model class Θ and on the regularization constant τ. For tracking regret, we obtain a bound of the form

$$R_T(\Theta^{1:T}) = O(V_T\sqrt{T}),$$

where

$$V_T \triangleq \sup_{\theta^{1:T}\in\Theta^{1:T}} \sum_{t=1}^T \|\theta^t - \theta^{t-1}\|$$

is a natural measure of the complexity of the time-varying reference class $\Theta^{1:T}$. In other words, we can successfully track a dynamically changing network structure, provided the changes are sufficiently infrequent and/or smooth.

We also use the above results on the regret to establish generalization error bounds, which explicitly account for the role of sparsity regularization. Intuitively, if sparsity regularization gives strong performance guarantees in a batch setting, then applying the same regularization in the online setting should yield similar gains. We prove that this intuition is in fact true.

Finally, we will demonstrate the efficacy of the proposed approach on a simulated data set generated from a time-varying binary Markov random field, as well as on the US Senate voting records.

References

[1] O. Frank and D. Strauss, "Markov graphs," *J. Amer. Statist. Soc.*, vol. 81, no. 395, pp. 832–842, 1986.

[2] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, "A survey of statistical network models," *Foundations and Trends in Machine Learning*, vol. 2, no. 2, pp. 1–117, 2009.

[3] O. Banerjee, L. El Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Machine Learn. Res.*, vol. 9, pp. 485–516, 2008.

[4] P. Ravikumar, M. J. Wainwright, and J. Lafferty, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *Ann. Statist.*, vol. 38, no. 2, pp. 1287–1319, 2010.

[5] M. Kolar and E. Xing, "Sparsistent estimation of time-varying discrete Markov random fields," *Ann. Statist.*, 2009, Submitted.

[6] H. Höfling and R. Tibshirani, "Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods," *J. Machine Learn. Res.*, vol. 10, pp. 883–906, 2009.

[7] N. Cesa-Bianchi, A. Conconi, and C. Gentile, "On the generalization ability of online learning algorithms," *IEEE Trans. Inform. Theory*, vol. 50, no. 9, pp. 2050–2057, September 2004.

[8] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *J. Machine Learn. Res.*, vol. 11, pp. 2543–2596, 2010.

[9] J. Langford, L. Li, and T. Zhang, "Sparse online learning via truncated gradient," *J. Machine Learn. Res.*, vol. 10, pp. 777–801, 2009.

[10] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, "Composite objective mirror descent," in *Conf. on Learning Theory (COLT)*, 2010.

[11] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, December 2008.

Dynamic network centralities and saltatoric information transmission: lessons of biological networks

Peter Csermely,[1,*] Eszter Hazai,[2] Huba J. M. Kiss,[1] István A. Kovács,[1,3] Ágoston Mihalik,[1] Robin Palotai,[1] Gábor I. Simkó,[1,4] Kristóf Z. Szalay,[1] Máté Szalay-Bekő,[1] and Shijun Wang[5]

In the last decade several analogies have been uncovered between the topology and dynamics of complex social and biological networks [1]. Our studies on community-based, perturbation-based and spatial game-based centralities showed that inter-modular nodes and links play a key role in information transmission, and led us to suggest a novel information transmission mechanism of complex networks.

Moduland is a novel method family to detect pervasively overlapping communities ([2], www.linkgroup.hu/modules.php). First, local influence zones of each links (or nodes) are defined. Next, a community landscape is constructed, where the horizontal plane corresponds to a 2D visualization of the network, while the vertical scale is the sum of the influence zones containing the given link (or node). The value of the vertical scale is called as community centrality, since it characterizes the influence reaching the given link or node from the whole network. The overlapping modules are the 'hills' of the community landscape. The method also constructs a hierarchical coarse-grained representation of the network, where the nodes correspond to the modules of the original network, and the link-weights denote the overlaps of the modules. Our studies on changes of protein-protein interaction networks during abrupt changes of the environment (stress) [3] showed that the overlap of modules decreases and modules became partially disintegrated as an initial response to stress. The stress-induced decrease of inter-modular connections is beneficial, since it A.) allows a better focusing on vital functions, and thus spares resources; B.) localizes damage (e.g. of free radicals) to the affected modules; C.) reduces the propagation of noise; D.) allows a larger 'degree of freedom' of the individual modules to explore different adaptation strategies; and E.) helps the 'mediation of inter-modular conflicts' during a period of violent intra-modular changes Modular overlaps emerge as keys of adaptive processes in cells – and in all complex systems including social networks. Changes in community centrality identified key players of the response to the cellular challenge.

Turbine ([4] www.linkgroup.hu/Turbine.php) is a widely applicable, Matlab-compatible algorithm, to assess the propagation of perturbations in any cellular networks. In these studies intermodular nodes emerged as highly efficient transmitters of perturbations.

Based on our earlier studies on spatial games (*where agents playing repeated rounds of social dilemma-type games, like the prisoner's dilemma game, can play only with their neighbors*), we constructed NetworGame ([5] www.linkgroup.hu/NetworGame.php), which is a versatile

[1] Department of Medical Chemistry, Semmelweis University, 37-47 Tűzoltó Street, Budapest, H-1094, Hungary. [2]Virtua Drug Co., Budapest, Hungary. [3]Department of Physics, Loránd Eötvös University, Pázmány P. s. 1/A, H-1117 Budapest, Hungary and Research Institute for Solid State Physics and Optics, H-1525 Budapest, P. O. Box 49, Hungary. [4]Vanderbilt University, Nashville TN, USA. [5]Clinical Center, National Institutes of Health, Bethesda MD, USA.

[*]Presenter and corresponding author. E-mail, csermely@eok.sote.hu; E-mail addresses of other authors: EH: eszter.hazai@virtuadrug.com; HJMK: kisshuba@googlemail.com; IAK: kovacs.pisti@gmail.com; AM: anaston@gmail.com; RP: palotai.robin@gmail.com; GIS: gsimko@gmail.com; KZS: kris@sch.bme.hu; MS: szalay.beko.mate@gmail.com; SW: wangshi@cc.nih.gov.

program package to model any types of two-agent games (with 2 to 5 strategies) in any real world, or model networks using any types of strategy update rules, update dynamics and starting strategies. The NetworGame program interprets game centrality as the ability of a networked agent (or a link of two agents) with a single initial defective strategy to change an overall initial starting cooperation to defection (and vice versa: a cooperative strategy of a linked node-pair/triangle changing overall defection to cooperation). Spatial games can also be rationalized in networks of non-conscious agents, such as amino acids, or proteins [6]. Our game centrality measures correctly identified the major decision makers of social cooperation in benchmark networks, such as the Zachary karate club network or Michael's strike network, and pinpointed key 'actors' determining the cooperation of biological networks.

Recently we summarized the features of particularly dynamic central elements, and called them as 'creative elements' [7]. These elements bridge Ronald S. Burt's 'structural holes', and provide a key subset of Mark Granovetter's 'weak links'. Active centers and binding sites of proteins often occupy such a position in protein structure networks. As the complexity of the system increases, the mobility of creative elements expands, and covers more and more the entire network [7].

Based on our earlier studies demonstrating the partial disassembly of networks as a response to stress [3], recently we proposed that information transmission of 'cumulus-type' networks (*which have a limited overlap between their modules and a more compact, rigid module structure*) can be described by an 'energy transfer' mechanism. In 'stratus-type' networks (*having a significant overlap between their modules*) the information transfer utilizes multiple trajectories. These signaling trajectories converge at modular boundaries. Bridging nodes may have a decisive role in the regulation of signal transmission from one network module to another [8]. Such inter-modular nodes, called as cross-talks in cellular information transfer networks, are key players of biological information transmission [9].

Authors would like to thank members of the LINK-group (www.linkgroup.hu) for helpful suggestions. Work in the authors' laboratory was supported by research grants from the Hungarian National Science Foundation (OTKA-K69105 and OTKA-K83314), from the EU (FP6-016003).

References:

1. P. Csermely, *Weak links: The universal key to the stability of networks and complex systems*, Springer Verlag (2009) www.weaklink.sote.hu/weakbook.html
2. I. A. Kovács, R. Palotai, M. S. Szalay, P. Csermely, Community landscapes: a novel, integrative approach for the determination of overlapping network modules. *PLoS ONE* **7**, e12528 (2010).
3. R. Palotai, M. S. Szalay, P. Csermely, Chaperones as integrators of cellular networks: changes of cellular integrity in stress and diseases. *IUBMB Life* **60**, 10–18 (2008).
4. M. A. Antal, C. Böde, P. Csermely, Perturbation waves in proteins and protein networks: Applications of percolation and game theories in signaling and drug design. *Curr. Prot. Pept. Sci.* **10**, 161–172 (2009).
5. S. Wang, M. S. Szalay, C. Zhang, P. Csermely, Learning and innovative elements of strategy update rules expand cooperative network topologies. *PLoS ONE* **3**, e1917 (2008).
6. P. Csermely, R. Palotai, R. Nussinov, Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35**, 539–546 (2010).
7. P. Csermely, Creative elements: network-based predictions of active centres in proteins, cellular and social networks. *Trends Biochem. Sci.* **33**, 569–576 (2008).
8. P. Csermely, K. S. Sandhu, E. Hazai, Z. Hoksza, H. J. M. Kiss, F. Miozzo, D. V. Veres, F. Piazza, R. Nussinov, Disordered proteins and network disorder in network representations of protein structure, dynamics and function. Hypotheses and a comprehensive review. *Curr. Prot. Pept. Sci.* **12**, in press, http://arxiv.org/abs/1101.5865 (2011).
9. T. Korcsmáros, I. J. Farkas, M. S. Szalay, P. Rovó, D. Fazekas, Z. Spiró, C. Böde, K. Lenti, T. Vellai, P. Csermely, Uniformly curated signaling pathways reveal tissue-specific cross-talks, novel pathway components, and drug target candidates. *Bioinformatics* **26**, 2042–2050 (2010).

Converging an Overlay Network to a Gradient Topology

Håkan Terelius, Guodong Shi, Jim Dowling, Amir Payberah,
Ather Gattami and Karl Henrik Johansson*†

Abstract

The paper investigates the topology convergence problem for the gossip-based gradient overlay network. In an overlay network where each node has a local utility value, a gradient overlay network is characterized by the properties that each node has a set of neighbors with the same utility and a set of neighbors containing higher utilities, such that paths of increasing utilities emerge in the network topology. We show how a gossip-based overlay network, built using a preference function that samples random nodes using a peer sampling service, converges to a complete gradient structure. A sufficient and necessary condition is proposed on the sampling probability function for the considered network converging to a gradient structure with probability 1. We illustrate through simulations how the gradient overlay network can be used to build a more efficient live-streaming peer-to-peer system than one built using random sampling.

Keywords: Overlay networks; topology convergence; gossiping; gradient topology

*Håkan Terelius, Guodong Shi, Amir Payberah, Ather Gattami and Karl Henrik Johansson are with Royal Institute of Technology, Stockholm, Sweden {guodongs,hakante,amir,gattami,kallej}@kth.se

†J. Dowling is with Swedish Institute of Computer Science (SICS) jdowling@sics.se

A Model of Strategic Behavior in Networks of Influence

Mohammad T. Irfan and Luis E. Ortiz

Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
{mtirfan,leortiz}@cs.sunysb.edu

We propose *influence games*, a new class of graphical games, as a model of the behavior of large but finite networked populations. Grounded in non-cooperative game theory, we introduce a new approach to the study of influence in networks that captures the strategic aspects of complex interactions in the network. We study computational problems on influence games, including the identification of the most influential nodes. We characterize the computational complexity of various problems in influence games, propose several heuristics for the hard cases, and design approximation algorithms, with provable guarantees, for the influential nodes problem based on a connection we establish to the *minimum hitting set problem* [4].

To date, the study of influence in networks has concentrated mostly on analyzing the diffusion (or "contagion") processes induced by the influences in the network [1, 6, 7]. The notion of "influential nodes" considered in this paper is different, and is aimed at complementing the traditional line of work with a new game-theoretic perspective. Inspired by threshold models in social science [2], we define *influence games* as a class of graphical games [5] where each player corresponds to a node in a directed graph encoding"influence factors." The set of *actions* (or *pure strategies*) of each player is $\{1, -1\}$, for "adopt" or "not-adopt" a behavior, respectively. Each player also has an *influence function* f_i mapping each joint-action of the player's parents in the game graph to a real number, and a real-valued *tolerance threshold* b_i. Each player i's *payoff function* is defined such that, given a joint-action $\mathbf{x}_{\mathrm{Pa}(i)}$ of player i's parents $\mathrm{Pa}(i)$ in the graph, player i's best-response is 1 (respectively, -1) if $f_i(\mathbf{x}_{\mathrm{Pa}(i)})$ exceeds (is below) b_i; and *indifferent* if $f_i(\mathbf{x}_{\mathrm{Pa}(i)}) = b_i$. In the special class of *linear influence games* (LIGs), each f_i is a weighted *sum* of $\mathbf{x}_{\mathrm{Pa}(i)}$.

We define a set S of players in a game as most influential, *with respect to a specified pure strategy Nash equilibrium (PSNE)* $\mathbf{x}^*$, if the players in S to choosing actions according to $\mathbf{x}^*$ enforces all others to also choose actions according to $\mathbf{x}^*$. Said differently, the players in S are collectively so influential that they are able to restrict the choice of actions of every other player in a stable solution to a *unique* one. We further extend this definition by allowing for a preference function over all possible sets of the most influential nodes (e.g., a minimum-cardinality set). Departing from the contagion model and rather concentrating on the PSNE of the influence game, we capture significant, basic, and core *strategic* aspects of complex interaction in networks that naturally appear in many real-world problems (e.g., determining the most influential Senators in Congress).

We study two fundamental algorithmic questions in this setting—computing PSNE of influence games and finding the most influential set of nodes. We show that various versions of these problems (e.g., existence of a PSNE, uniqueness of PSNE, counting the number of PSNE even in star networks, etc.) are intractable, unless P = NP. Nevertheless, on the positive side, we show how to compute a PSNE of special types of influence games, such as the ones with non-negative influence factors and the ones having tree structures, in polynomial time. Furthermore, given the set of all PSNE H, we give a $(1 + \log |H|)$-factor approximation algorithm for the most influential nodes selection problem. We also illustrate the whole computational scheme empirically, using random influence games and influence games learned from the US Congress voting records using machine learning techniques [3].

Bibliography

[1] E. Even-Dar and A. Shapira. A note on maximizing the spread of influence in social networks. In *Workshop on Internet and Network Economics (WINE)*, 2007.

[2] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978. ISSN 00029602. URL `http://www.jstor.org/stable/2778111`.

[3] J. Honorio and L. Ortiz. Learning graphical games from behavioral data. Submitted for review, 2011.

[4] R. Karp. Reducibility among combinatorial problems. In R. Miller and J. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[5] M. Kearns, M. Littman, and S. Singh. Graphical models for game theory. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 253–260, 2001.

[6] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In N. Nisan, T. Roughgarden, Éva Tardos, and V. V. Vazirani, editors, *Algorithmic Game Theory*, chapter 24, pages 613–632. Cambridge University Press, 2007.

[7] S. Morris. Contagion. *The Review of Economic Studies*, 67(1):57–78, 2000. ISSN 00346527. URL `http://www.jstor.org/stable/2567028`.

Optimal Marketing and Pricing over Social Networks

Vahab S. Mirrokni
Google Research, New York
mirrokni@google.com

ABSTRACT

We discuss the use of social networks in implementing viral marketing strategies. While infuence maximization has been studied in this context, we study revenue maximization, arguably, a more natural objective. In our model, a buyer's decision to buy an item is influenced by the set of other buyers that own the item and the price at which the item is offered. We focus on algorithmic question of finding revenue maximizing marketing strategies and study the problem in two main cases with and without price discrimination.

1. INTRODUCTION

Social networks pervade our lives and significantly influence the decisions we make. Our decisions to buy a cell phone, go to college, or smoke a cigarette are fundamentally affected by the decisions of our friends. Traditionally, social scientists have studied the effects of social networks on decision-making in a very abstract sense or in explicit small-scale settings. However, the proliferation of social-networks on the Internet has allowed companies to collect information about social-network users and their social relationships, yielding explicit information regarding connections in massive social networks. As a result, it has become increasingly relevant to understand how the social network structure affects the choices of the society it describes, and how this knowledge can be leveraged to monetize these networks in an Internet setting.

In this talk, our main focus will be on how to leverage network structure to promote a product and/or maximize revenue. Consider a company interested in promoting a product. As one agent adopts the product, this impacts other potential adopters. Such an effect is called an *externality*. Externalities that induce further adoption of the product are called *positive externalities*. These externalities significantly influence the diffusion of the product in the social network and the revenue that can be extracted during this diffusion process.

A far-sighted seller can take advantage of the existence of positive externalities to increase its revenue. For instance, in order to influence many buyers to buy the good, the seller could initially offer some popular buyers the good for free. Indeed such selling techniques are already employed in practice. TiVo, a company which makes digital video recorders, initially gave away its digital video recorder for free to a select few video enthusiasts [6]. Such promotions may be an effective way to create a buzz about the product.

The basic idea of giving away the item for free can be generalized in a couple of ways: First, rather than offering the item for free, sellers could offer discounts. There is a trade-off: larger discounts decrease the revenue earned from the transaction while increasing the likelihood of a sale and the influence on future buyers. *How large should the discounts be?* Second, the sequence in which sales happen has an impact on the effect of externalities. Influence is generally not symmetric. Often popular, well-connected users wield more influence. Clearly, we would like sales that have the potential to cause further sales to occur earlier. *In what sequence should the selling happen?* The goal of this paper is to explore marketing strategies that optimize a seller's revenue.

We investigate the marketing and pricing problems over social networks in three different parts: In the first part, we allow price discrimination and design pricing pricing strategies that can target special social network users with specific offers [4]. In the second part, we discuss optimal pricing strategies in settings without price discrimination, and by using publicly known posted pricing [1, 2].

We investigate marketing strategies that maximize revenue from the sale of digital goods. In this setting, a buyer's decision to buy an item depends on other buyers owning the item *and* the price offered to the buyer; the value of the buyer for the good is defined by a set function which models the influence from other set of buyers on this buyer. We assume that though the seller does not know the value functions, but instead has distributional information about them. In general, smaller prices increase the probability of sale.

Marketing with Price Discrimination. In this part, we discuss optimal marketing strategies and discuss results from the following paper:

J. Hartline, V. S. Mirrokni, M. Sundararajan. *Optimal Marketing Strategies over Social Networks.* The World Wide Web Conference (WWW), 2008.

V. S. Mirrokni, S. Roch, M. Sundararajan. *Optimal Posted-Price Marketing with Influence Propagation over Social Networks.* Manuscript.

In the first paper, the seller considers buyers in some sequence and offers each buyer a price. When the buyer accepts the offer, the seller earns the price of the item as the revenue. As a result, a marketing strategy has two elements: the sequence in which we offer the item to buyers, and the prices that we offer. In general it is advantageous to get influential buyers to buy the item early in the sequence; it

even makes sense to offer such buyers smaller prices to get them to buy the item. This paper first gives a polynomial-time algorithm for a symmetric setting, and then shows that the optimal marketing strategy is NP-Hard in the general settings. In order to design approximation algorithms[1] for the problem, this paper identifies a simple marketing strategy, called the *influence-and-exploit* strategy: In the initial *influence step*, motivated by the the form of the optimal strategy in the symmetric case, the seller starts by giving the item away for free to a specifically chosen set of buyers. In the *exploit step*, the seller visits the remaining buyers in a random sequence and attempts to maximize the revenue that can be extracted from each buyer by offering it the (myopic) optimal price. About these strategies, the paper first shows that they achieve a reasonable approximation of the optimal marketing strategy, which, by a hardness result is not polynomial-time computable, and it shows that if the buyer-specific revenue functions are submodular, then the expected revenue as a function of the set of buyers who get the item for free is also submodular. Therefore, in order to identify the set of buyers to influence, non-monotone submodular maximization [3] is employed.

The second paper mentioned above studies a set of influence and pricing setting strategies in which a set of buyers get the item for free, and then the seller sets a publicly available price (or a sequence of public prices) for all the remaining buyers. The problem is to find the right subset of buyers to influence and then set the right price, to maximize the total revenue. In certain Bayesian Influence Setting, the paper gives an algorithm with approixmaiton factor of 40% combining by dynamic programming and submodular maximization [5].

Optimal Pricing without Price Discrimination. In this part, we study the optimal iterative pricing strategies in the presence of positive network externalities. In contrast to the paper discussed in the previous part [4], we study the pricing problem without price discrimination where the goal is to maximize revenue by posting a sequence of publicly available prices. We study the following papers:

H. Akhlaghpour, M. Ghodsi, N. Haghpanah, H. Mahini, V. S. Mirrokni, and P. Nikzad. Iterative pricing with positive network externalities. WINE 2010, and

N. Ahmadi, S. Ehsani, M. Ghodsi, N. Haghpanah, N. Immorlica, H. Mahini, and V. S. Mirrokni. Optimal equilibrium pricing over social networks. WINE 2010.

The first paper studies the Bayesian setting in which there are some prior knowledge of the probability distribution on the valuations of buyers. In particular, the paper studies two iterative pricing models in which a seller iteratively posts a new price for a digital good (visible to all buyers), and any interested buyer can buy the item at the posted price. In one model, re-pricing of the items are only allowed at a limited rate. In particular, the item can be repriced only after a long period of time at which no new buyer is interested in buying the item. For this case, the paper gives an FPTAS for the optimal pricing strategy using a dynamic programming approach. Furthermore, using this FPTAS, the paper reports interesting observation about optimal pricing strategies over preferential attachment networks. In the second model, the paper shows that the revenue maximization problem is inapproximable even for simple deterministic valuation functions. In light of this hardness result, the paper presents constant and logarithmic approximation algorithms for a special case of this problem where the individual distributions are identical [1].

Finally, we discuss the second paper mentioned above which studies optimal posted pricing in the presence of strategic buyers. In this paper, the pricing problem is modeled as a strategic game amongst buyers and the seller, and the paper studies existence and revenue properties of the resulting equilibrium. In particular, for a special case where the type of all buyers is the same, it is proved that pure Nash equilibria exist and an efficient algorithm is given to compute the pricing strategy that will result in an equilibrium that maximizes the revenue of the seller among all such equilibria [2].

[1] An algorithm is a c-approximation if its revenue is at least c times the revenue of the optimal marketing strategy.

2. REFERENCES

[1] H. Akhlaghpour, M. Ghodsi, N. Haghpanah, H. Mahini, V. S. Mirrokni, and P. Nikzad. Iterative pricing with positive network externalities. In *WINE*, 2010.

[2] N. Anari, S. Ehsani, M. Ghodsi, N. Haghpanah, N. Immorlica, H. Mahini, and V. S. Mirrokni. Optimal equilibrium pricing over social networks. In *WINE 2010*.

[3] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrak. Maximizing non-monotone submodular functions. In *FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 461–471, Washington, DC, USA, 2007. IEEE Computer Society.

[4] J. Hartline, V. S. Mirrokni, and M. Sundararajan. Optimal marketing strategies over social networks. In *WWW*, pages 189–198, 2008.

[5] V. S. Mirrokni, S. Roch, and M. Sundararajan. Optimal posted-price with influence propagation over social networks. In *Manuscript*, 2009.

[6] Rob Walker. http://www.slate.com/id/1006264/.

Mood, Sleep and Face-to-Face Interactions in a Co-located Family Community

Sai Moturu[1], Inas Khayal[1, 2], Nadav Aharony[1], Alex (Sandy) Pentland[1]
[1]MIT Media Lab, [2]Masdar Institute of Science and Technology
{sai, nadav, sandy}@media.mit.edu, ikhayal@mit.edu

Today's smartphones have great potential as social sensors due to their numerous sensing capabilities. In this work, we present some insights into the correlations between social interactions, sleep and mood from an experiment where smartphones were used as social sensors to track face-to-face interactions in a co-located community of families.

The data used for this analysis was collected from 54 participants (MIT graduate students and spouses). Participants were provided Android smartphones with a proprietary software sensing platform that allowed us to track face-to-face interactions through bluetooth proximity sensing in addition to other behavioral, contextual and communication patterns. Regular surveys were used to obtain additional contextual and behavioral information about participants.

Using the data on face-to-face interactions over a one-month period, we compute an overall measure of "how social" an individual was. Other information used for this analysis includes the amount of sleep every night for each individual and the predominant mood for the following day.

Using the mood information, we separated the groups into two: 1) those who exhibited primarily good mood (relaxed, calm, happy, content) i.e., on at least 70% of the occasions 2) those who exhibited poor mood (stressed, anxious, frustrated, angry) i.e., on at least 30% of the occasions. We observed that people who fell into the latter group were significantly less social in this community than those who fell into the former. While such a connection might have been made in the past, we believe that this is the first time such an observation has been made using actual face-to-face interactions.

One of the key differences in this experiment when compared with our previous phone sensing experiments is the involvement of families. Hence, we compared behavior patterns between spouses. We observe that 1) the wives show poor mood more often than the husbands ($p<0.0001$), 2) the wives tend to sleep a bit better than their husbands do ($p<0.001$), and 3) there is no significant difference in how social the wives and husbands are in this community.

We also observed that the amount of sleep obtained and the following day's mood were not independent ($p<0.0001$). When the predominant mood observed was poor, participants had slept less than 7 hours the previous night in nearly 50% of the cases. However, when the mood observed was good, this number dropped to 30%. The following day's mood was significantly poorer for those who slept less than 7 hours when compared to those who slept more than 7 hours. This result holds even when the husbands and wives are separated ($p<0.001$).

Using smartphones as sensors provides us with a great opportunity to study interactions and behaviors in a social network. The ability to quantify face-to-face interactions provides us with a better way to study these relationships when compared to the traditional approach, primarily dependent on surveys. The results presented here only include partial analysis of the available data. We intend to have much more concrete analyses that provide novel insights into these factors by the time we present at WIDS.

A Game Theoretic Perspective on Network Topologies

Shaun Lichter[1] and Christopher Griffin[2]

[1]Department of Industrial and Manufacturing Engineering, The Pennsylvania State University, University Park, PA 16802, lichter@psu.edu

[2]Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802, griffinch@ieee.org

March 6, 2011

Special structure in networks has been considered in several recent papers [1, 2, 3, 4] that have cut across various subjects including social networks, information networks, and biological networks as well as physical networks such as power grids and road networks. The network science literature was largely inspired by the observation of macroscopic structural properties (e.g., small world, power law degree distribution) of networks that occurred in several distinct network types (e.g., social, information, biological networks), which have diverse microscopic properties. The network science literature has largely been devoted to finding the mechanisms by which networks form and/or evolve in order to generate the structural properties that are observed. The momentum in this direction has largely been driven by the statistical physics community [1, 2, 3], who argue that the phenomena of complex networks (e.g., power laws) may be explained by laws that reach across all complex networks because they are phenomena that are inherent to the complexity of the networks.

Alternatively, there has been recent interest in the structural properties of networks that have been designed via optimization [5, 6, 7]. This perspective is motivated largely by the fact that many networks (in the abstract sense) are models for physical networks that are designed by humans to function with particular objectives (or even designed by nature to serve an evolutionary purpose). These networks are distinctly different from social networks, which are abstract models that describe interactions between actors (e.g., people talking, writing scientific papers or dating). Such networks (e.g., power grids, communication networks) are not designed through central coordination, but arise as a result of the objectives of multiple independent actors. As a result there are structural properties that often exist in these networks that are not explained by models that do not account for these functioning characteristics [8].

In this work we show through game theoretic principles, that networks may form with particular structural properties. Specifically, we show that as the result of a game, a graph with an arbitrary degree sequence may be formed as a stable network.

In this work we model the emergence of collaborations among players (e.g., firms) as a strategic network formation game [9], by allowing selfish agents to choose with which other agents they would like to form a link. Each agent has the option to deny a link to another agent, so the formation of a link requires the cooperation of both players. A value function assigns a value to each particular graph and this value is distributed to agents by an allocation function (or allocation rule). This distribution of value drives a player's preference for particular graph structures. This work is unique in that it takes a perspective similar to the network formation literature (i.e., modeling network formation as a result of player strategies), but with a motivation from the network science literature (i.e., finding the models that create particular topologies). The motivation in this work is different from the network formation literature in that, rather than modeling particular applications and finding the laws that govern stable networks, we focus on finding the network formation mechanisms (e.g., strategies, objectives, dynamics) that result in the formation of networks with particular structural characteristics that correspond to those observed in real networks.

Jackson and Wolinksy use *pairwise stability* to model stable networks without the use of non-cooperative Nash equilibrium [10]. Pairwise stability implies that in a stable network, for each link that exists, (1) both players must benefit from it and (2) if a link can provide benefit to both players, then it in fact must exist. Jackson notes that *pairwise stability* is not a perfect modeling mechanism, but allows us to consider models where each player may veto a link.

We show how to construct an allocation rule to ensure that a graph with an arbitrary degree sequence is stable. We show that players with convex objectives with minimums at points d_i will result in a stable graph of degree sequence $\mathbf{d} = \{d_1, \ldots, d_n\}$. This theorem implies there is a game that may result in a stable graph with a power law degree distribution or in fact any degree distribution. Unfortunately, this stable graph is not uniquely stable as graphs with a different degree sequence may also be stable for this game.

In addition to the problem of understanding the formation of graphs with arbitrary degree sequences as the result of game theoretic interaction, we also study the problem of generating graphs with these degree sequences. There has been recent interest in generating graphs with an arbitrary degree sequence [11, 12, 13, 14]. The approach we investigate is unique in that it is the result of an optimization problem. This approach compliments and completes the model of graph formation as the result of a game. Further, in the event that a degree sequence is not graphical (i.e. there is not a graph with that degree sequence [15]), the math program constructed in our work returns the nearest graph with [graphical] degree sequence. Here, nearness is measured using either the Earth Mover's metric or the standard ℓ_1 metric. The resulting mathematical programs can then be used to calculate the price of anarchy for the game theoretic model of stable graphs in the foregoing collaboration game.

References

[1] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, p. 509, 1999.

[2] M. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[3] S. Dorogovtsev and J. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, pp. 1079–1187, 2002.

[4] R. Albert and A. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, pp. 47–97, 2002.

[5] D. Alderson, "Catching the'Network Science'Bug: Insight and Opportunity for the Operations Researcher," 2008.

[6] A. Nagurney, "Comment on Catching the Network Science Bug by David L. Alderson," in *OPERATIONS RESEARCH (ONLINE FORUM COMMENTARY)*, vol. 56, no. 5, 2008, p. 5.

[7] D. Alderson, J. Doyle, R. Govindan, and W. Willinger, "Toward an optimization-driven framework for designing and generating realistic internet topologies," *ACM SIGCOMM Computer Communication Review*, vol. 33, no. 1, p. 46, 2003.

[8] J. Doyle, D. Alderson, L. Li, S. Low, M. Roughan, S. Shalunov, R. Tanaka, and W. Willinger, "The robust yet fragile nature of the Internet," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, p. 14497, 2005.

[9] B. Dutta and S. Mutuswami, "Stable networks," *Journal of Economic Theory*, vol. 76, no. 2, pp. 322–344, 1997.

[10] M. Jackson and A. Wolinsky, "A strategic model of social and economic networks," *Journal of economic theory*, vol. 71, no. 1, pp. 44–74, 1996.

[11] T. Britton, M. Deijfen, and A. Martin-L
"of, "Generating simple random graphs with prescribed degree distribution," *Journal of Statistical Physics*, vol. 124, no. 6, pp. 1377–1397, 2006.

[12] R. Milo, N. Kashtan, S. Itzkovitz, M. Newman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," *Arxiv preprint cond-mat/0312028*, 2003.

[13] C. Del Genio, H. Kim, Z. Toroczkai, and K. Bassler, "Efficient and exact sampling of simple graphs with given arbitrary degree sequence," *PloS one*, vol. 5, no. 4, p. e10012, 2010.

[14] M. Mihail and N. Vishnoi, "On generating graphs with prescribed vertex degrees for complex network modeling," *ARACNE 2002*, pp. 1–11, 2002.

[15] P. Erdős and T. Gallai, "Graphs with prescribed degree of vertices," *Mat. Lapok (NS). v11*, pp. 264–274.

Dynamical Control of DeGroot Learning to Shape Belief Structures in Social Networks

Sridhar Mandyam
sridhar.mandyam@ecometrix.in

Usha Sridhar
usha.sridhar@ecometrix.in

Ecometrix Research
Bangalore 560078
India

Extended Abstract

The notion of learning in social networks derives from the proposition that individual agents update the beliefs they hold on some global truth by aggregating beliefs of connected neighbors. Among the many alternative Bayesian and non-Bayesian approaches [1, 2, 4, 5] to model the dynamics of the learning process, an averaging procedure attributed to DeGroot [6] has drawn much attention in the recent years due its relative tractability. The DeGroot learning scheme [3, 6] uses an intuitively appealing homogeneous Markov chain analog to capture the belief updating process as recursion in which new beliefs are produced as (convex) weighted averages of older beliefs [7, 3]. The achievement of stationarity in the chain under conditions of strong connectedness has been interpreted as a convergence of beliefs to a consensus on the global truth among the agents in the community. The convex weighting process in DeGroot learning leaves final beliefs enclosed within the range of initial beliefs with only the 'connectedness' or 'social influence' of the network driving the evolution of beliefs.

In this paper, we propose endogenous and exogenous mechanisms to apply some level of control on this update process. The rationale for control derives from the observations that a) opportunities for better utilization of 'privately held' social influence information at individual level with 'publically held' belief information may not have been exhausted in the basic DeGroot recursion; b) the emergence of new classes of social media offer opportunities for infusing individual and group level persuasive biases which can help shape the belief structures; and c) such control mechanisms may offer the means to lift and shape even consensus beliefs.

The control methods proposed here are new contributions, and extend the applicability of DeGroot social learning theory to new areas where special classes of biases may be introduced to change the belief evolution patterns even as they are learnt through averaging processes. They also allow new ways to study the emergence of consensus, not merely as a result of network structure, but as a means to control, and even perhaps destroy it.

We consider a social network of m agents who hold beliefs $\boldsymbol{b}$ ($m \times 1$) about some global truth. Given a row-stochastic matrix $\boldsymbol{T}$, where T_{ij} denotes the social influence or attention agent i pays to agent j we consider *two* approaches to apply control on the basic DeGroot recursion: $\boldsymbol{b_t} = \boldsymbol{T b_{t-1}}$, where the subscripts refer to a learning or update cycle: one, which ***endogenously*** perturbs the matrix $\boldsymbol{T}$ to lift beliefs, and another, which ***exogenously*** controls the DeGroot recursion by embedding it in an exogenous linear control system.

1. Endogenous Dynamic Control

In this approach, we develop a new iterative control algorithm, which we refer to as BLIFT, in which each agent i perturbs its own private social influence weights (the row $\boldsymbol{T_i}$ of matrix T) and belief

information within the same cycle of the DeGroot recursion to obtain a lift of its own belief. We show how the perturbation is calculated to *guarantee* at least an increase in the value of belief in every learning cycle, for every agent.

We show the perturbation transforms the resultant Markov chain analog into a non-homogenous one and provide proofs of convergence and consensus for the following proposition.

Endogeneous Control with BLIFT: Proposition 1
Given a $\boldsymbol{T}$ *matrix that represents an analog of an aperiodic and irreducible Markov chain, and the sequence of perturbations to* $\boldsymbol{T}$ *as defined in Algorithm BLIFT, the resultant non-homogeneous chain converges to a consensus that is bounded below by the belief which DeGroot update will converge to with the same* $\boldsymbol{T}$.

Sketch of Proofs

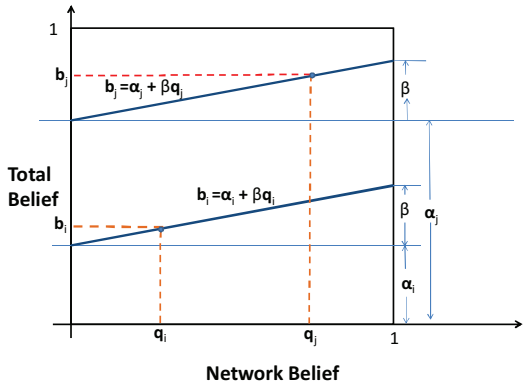
We first show that BLIFT perturbation produces a non-homogeneous Markov chain equivalent in which each influence matrix $\boldsymbol{T_t}$, when multiplied by previous beliefs to produce a new belief vector is guaranteed to be at least greater than its DeGroot update. We then show how this property produces a contraction of the variance in the sequence of belief vectors over time, thus guaranteeing convergence, as well as consensus under strong connectivity conditions

We illustrate the application of BLIFT with examples of small social networks.

2. Exogenous Dynamic Control

We develop theory to explore linear exogenous control of the form $\boldsymbol{b}_t = \boldsymbol{\alpha}_{t-1} + \beta_{t-1}(\boldsymbol{q}_{t-1})$ where $\boldsymbol{q}_{t-1} = \boldsymbol{T}\boldsymbol{b}_{t-1}$ and $\mathbf{T}$ is a *constant* influence matrix, $\boldsymbol{\alpha}_{(t-1)}$ is an (*m x 1*) vector of parameter values, one for every agent, set for cycle *(t-1)*. $\beta_{(t-1)}$ is a scalar value set (same value for all agents) in cycle (*t-1*). This formulation embeds the DeGroot recursion within a linear system. We refer to $\boldsymbol{q}$ as DeGroot or network-learned belief, and $\boldsymbol{b}$ as total belief.

Geometrically, this control formulation permits the treatment of $\boldsymbol{\alpha}$ and β as individual and group 'persuasive bias' parameters that may set to achieve a goal to pattern beliefs in the social network in a desired manner, and is depicted below pictorially:



We develop proofs for convergence of the new recursion above, and show that:

i) Linear dynamic control of network learning of initial beliefs produces a weighted network learning of the control parameter $\boldsymbol{\alpha_t}$, which is also learnt by the agents using the influence matrix $\boldsymbol{T}$, just as they learn beliefs.
ii) As $t \to N$, a suitably large positive integer, the control mechanism converges; and control parameters determine asymptotic convergence behavior;
iii) As $t \to N$, the control parameters can swamp out network learning of beliefs, and the agents will be left with a learning of supplied $\boldsymbol{\alpha_t}$.
iv) No specific network structure assumptions were required to be made to achieve 'weak' convergence or network-learning of beliefs or even control parameter $\boldsymbol{\alpha_t}$

We illustrate the theory with examples, and show how the choice of $\boldsymbol{\alpha_t}$ and β_t may be used for achieving different forms of purposeful control of belief learning in a social network.

We present concluding remarks on possibilities of dynamic control social learning by extending the theory of non-Bayesian DeGroot learning with new contributions in this paper. We discuss potential applications in social media for purposeful communications.

This paper extends preliminary work by the present authors on this subject dynamic control of belief learning, some of which is expected to appear soon [8, 9].

References

1. Acemoglu, Daron, Dahleh, Munther, Lobel, Ilan and Ozdaglar, Asuman.(2008), Bayesian Learning in Social Networks. Mimeo., M.I.T
2. DeMarzo, Peter M., Vayanos, Dimitri and Zwiebel, Je_rey. "Persuasion Bias, Social Influence, and Uni-Dimensional Opinions," Quarterly Journal of Economics 118: 909-968, 2003.
3. Eugene Seneta. *Non-negative matrices and Markov chains*. Springer-Verlag, 1981.
4. Friedkin, N.E. and Eugene C. Johnsen. "Social Influence Networks and Opinion Change- Advances in Group Processes," Vol. 16, 1-29, 1999.
5. Golub, B. and M.O. Jackson. "Naive Learning and Influence in Social Networks: Convergence and Wise Crowds," American Economic Journal Microeconomics, 2007.
6. Jackson, M.O. Social and Economic Networks, Princeton University Press, 2008
7. Roger L. Berger. "A Necessary and Sufficient Condition for Reaching a Consensus using DeGroot's Method," Journal of American Statistical Association, Vol.76, N0. 374, 415-418, 1981.
8. Usha Sridhar and Sridhar Mandyam. "Lifting Beliefs over DeGroot Learning in Social Networks", under review
9. Usha Sridhar and Sridhar Mandyam. "Exogenous Control of Belief Learning in Social Networks" to be presented at The 2011 Symposium on Social Media, New Media and Collaboration, Conference on Collaboration Technologies and Systems (CTS 2011).

Convergence to consensus in multiagent systems and the lengths of inter-communication intervals

Jan Lorenz
Center for Social Science Methodology
Carl von Ossietzky Universität Oldenburg,
Ammerländer Heerstraße 114–118, 26129 Oldenburg, Germany
post@janlo.de
http://www.janlo.de

January 22, 2011

Abstract

A theorem on (partial) convergence to consensus of multiagent systems is presented. It is proven with tools studying the convergence properties of products of row stochastic matrices with positive diagonals which are infinite to the left. Thus, it can be seen as a switching linear system in discrete time. It is further shown that the result is strictly more general than results of Moreau (IEEE Transactions on Automatic Control, vol. 50, no. 2, 2005), although Moreau's results are formulated for generally nonlinear updating maps. This is shown by demonstrating the existence of an appropriate switching linear system which mimics the nonlinear updating maps. Further on, an example system is given for which convergence to consensus can be shown by using the theorem. In this system the lengths of inter-communication intervals in the switching communication topology grow without bound. This makes other theorems not applicable.

Full preprint arXiv:1101.2926

Randomized Optimal Consensus for Multi-agent Systems with Time-varying Communication Graphs*

Guodong Shi and Karl Henrik Johansson†

Abstract

In this paper, we formulate and solve randomized optimal consensus problem for a multi-agent system with time-varying interconnection topology. The multi-agent system with a simple randomized iterating rule achieves an almost sure consensus meanwhile solving the following optimization problem

$$\min_x \sum_{i=1}^N f_i(x),$$

in which the information of objective function f_i corresponding to agent i can only be observed by node i itself.

At each time step, each agent independently chooses either taking an average among its time-varying neighbor set, or projecting onto the optimal solution set of its own objective function randomly:

$$x_i(k+1) = \begin{cases} \sum\limits_{j\in N_i(k)} a_{ij}(k)x_j(k), & \text{with probability } p \\ P_{X_i}(x_i(k)), & \text{with probability } 1-p \end{cases}$$

Both directed and bidirectional communications are studied. Connectivity conditions are proposed to guarantee an optimal consensus almost surely. The convergence analysis is carried out using convex analysis.

The results illustrate that a group of autonomous agents can reach an optimal opinion with probability 1 by each node simply making a randomized trade-off between following its neighbors or sticking to its own opinion at each time step.

Keywords: Multi-agent systems, Optimal consensus, Set convergence, Distributed optimization, Randomized methods

*This work has been supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Research Council and KTH SRA TNG.

†G. Shi and K. Johansson are with ACCESS Linnaeus Centre, School of Electrical Engineering, Royal Institute of Technology, Stockholm 10044, Sweden. Email: guodongs@kth.se, kallej@ee.kth.se

Extended Abstract: The Robustness of Consensus under Non-Bayesian Learning in Social Networks

Manuel Mueller-Frank*
Nuffield College and Department of Economics, University of Oxford

March, 2011

Social networks play an important role as communication platforms. Individuals interact with their social peers on an ongoing basis and use the information gained through the interaction when forming opinions or making decisions. The properties of opinion processes under repeated interaction in social networks have received substantial attention recently. Both Bayesian and non-Bayesian opinion formation processes have been considered. The usual setting is the following; agents are organized in a social network and receive initial private information. Interaction occurs in countable rounds and takes the form of all agents simultaneously announcing an opinion, a real number, to their neighbors in the network. Bayesian agents make Bayesian inference regarding the private information of all agents based on the announcements they observe, while the announcement of a non-Bayesian agent in a given period is simply the weighted average of the last period announcements of his neighbors and herself. One property of the opinion formation process that has been established for both Bayesian and non–Bayesian learning is asymptotic consensus of all agents in a strongly connected network. See DeMarzo, Vayanos and Zwiebel(2003) and Golub and Jackson(2010) for the non-Bayesian, and Mueller-Frank(2011) for the Bayesian case.

In terms of positive relevance, Bayesian models have the weakness that Bayesian inferences require potentially highly complex calculations from the agents. A potential weakness of the non-Bayesian models is the reliance on a particular functional form, weighted average, of the updating functions. Agents are assumed to apply weighted averages when forming their opinions, and weights are fixed in every period.[1]

This paper is concerned with the robustness of asymptotic consensus when considering a larger class of updating rules. I coincide with the existing non-Bayesian approach in assuming that the

*E-mail: manuel.mueller-frank@economics.ox.ac.uk

[1] DeMarzo, Vayanos and Zwiebel allow for changes in the weight agents assign to their own opinion while keeping the relative weights assigned to the opinions of neighbors constant.

announcement of an agent in a given period is a function of the last period announcements of his neighbors and herself.

Three properties of the updating function are shown to be of relevance when considering asymptotic consensus; continuity, a property which I denote as constricting, and finite types. A function is constricting if the following property holds for all announcement vectors; if in a given period a neighbor of agent i announces an opinion unequal to the announcement of agent i, then the announcement of i in the next period is strictly smaller than the highest announcement in his neighborhood and strictly larger than the smallest announcement in his neighborhood.[2] Opinion formation satisfies finite types if the number of different updating functions an agent uses over time is finite.

I show that if the social network is strongly connected and the updating functions of all agents are constricting, continuous and satisfy finite types, then the opinions of all agents in the network converge. This generalizes the existing asymptotic consensus result of DeMarzo, Vayanos and Zwiebel(2003), and Golub and Jackson(2010) in two ways. First, I allow agents to switch their updating function infinitely often over time and second, I consider a larger class of updating functions.

This larger class of updating functions captures interesting types of opinion formation behavior that cannot be captured by the fixed weighted average functions used in the literature. For one, approximate cognitive dissonant behavior, where the weight an agent assigns to his neighbor decreases with the difference in opinion, is captured by constricting and continuous updating functions. Also, under fixed weighted average updating functions the opinions of all neighbors matter and moreover, which neighbor announces which opinion. However, some individuals might only focus on the extreme opinions in their neighborhood, independent of the identity of the agent expressing the extreme opinion, and disregard all others. Here one can think of partisan behavior that assigns very high (low) weight to the largest opinion, very low (high) weight to the smallest opinion, and no weight to all other opinions. This type of partisan behavior can be used to model political partisans where opinions are ordered in a left-right spectrum.

As my second result I show that asymptotic consensus under the above conditions is also robust against introduction of random noise. If the opinion of each agent in a given period equals the value of his updating function plus an idiosyncratic random error term, and the probability of the error term equaling to zero converges to one geometrically, then asymptotic consensus holds with probability one.

[2] $f_i : \mathbb{R}^v \to \mathbb{R}$ is **constricting** if for all $a \in \mathbb{R}^v$, $a_l \neq a_i$ for some $l \in N_i$ implies $\min_{j \in N_i \cup \{i\}} a_j < f_i(a) < \max_{j \in N_i \cup \{i\}} a_j$.

COMPUTING SYMMETRIC FUNCTIONS ON MEMORY BOUNDED SOCIAL NETWORKS

ELCHANAN MOSSEL [1], ANUPAM PRAKASH [2] AND GREGORY VALIANT [3]

1. ABSTRACT

1.1. **Context and previous work.** Social networks are computationally constrained by bounds on the memory of agents and limits on the amount of interaction between neighbors. Computational tasks that would be trivial without these restrictions become interesting in the social network setting. We investigate the computational power of social networks in the bounded memory dynamics model formalized in [7]. The model is parameterized by the amount of memory m available to each agent and the rate of communication ν between neighbors, where interaction between neighbors is modeled by a Poisson process of rate ν.

Computation over a social network is modeled as the task of evaluating a function $f : [k]^n \to [k]$ where the inputs to f are drawn from $[k]$ and are distributed among the n agents in the network. Except for their immediate neighborhood, agents have no information about the size of the network, its structure or the identities of other agents. It is natural to restrict the family of functions f being computed, we restrict the function f to be symmetric capturing the anonymity of agents in the network.

The dynamics proceeds by updating the local state of the agents for every pairwise interaction. We say that *the network computes the function* f if for all possible inputs, every agent in the network eventually converges to the correct value of f. The question that we address is for which functions f there is a protocol for computing f for all connected networks. Furthermore for cases where a protocol exists we are interested in the convergence time in terms of the size of the graph n.

The computational problems of achieving consensus [6] [5], computing the majority over two colors [3] and graph coloring [4] on social networks have been studied in different contexts including experimental sociology and computer science. The computational model is motivated by these works, and abstracts some of the features common to the experimental studies. Protocols for reaching consensus with no additional memory and for computing the majority over two colors with 1 extra bit of memory per agent were presented in [7]. Closely related results bounding the convergence time of distributed consensus protocols have appeared in the recent literature on distributed computing [2, 1].

1.2. **Motivation:** The values assigned to the nodes in the network are referred to as colors, in accordance with the existing literature. The value of a binary symmetric function $f : \{0,1\}^n \to \{0,1\}$ depends on the number of nodes in the network with color 0. Computation of a binary symmetric function is motivated by the problem of estimating the number of nodes in the network that satisfy a given property, as can be seen by assigning color 0 to nodes having the property. For the case of k different colors, we consider the problem of computing comparison statistics like the

[1] UC Berkeley and Weizmann Institute of Science, *E-mail:* mossel@stat.berkeley.edu.
[2] Computer science division, UC Berkeley, *E-mail:* anupamp@eecs.berkeley.edu.
[3] Computer science division, UC Berkeley, *E-mail:* gvaliant@eecs.berkeley.edu.

plurality color or the least common color. Computation of symmetric functions is thus motivated by the problem of estimating statistics over social networks.

1.3. **Results:** Extending the results of [7] we show that given c bits of memory per agent it is possible to compute: (i) The c least significant bits in the binary representation of r, where r is the number of nodes having color 0. (ii) The threshold function $f : [n] \to \{0, 1\}$ which is 1 if $\frac{r}{n-r} \geq \frac{a}{b}$ and 0 otherwise for integers $a, b \leq 2^c$.

An information theoretic argument shows that there are symmetric functions that require $O(\log n)$ bits of memory per node to compute. Furthermore we show that this is tight and that given the truth table of a symmetric function it can be computed on a social network using $O(\log n)$ bits of memory per node.

Theorem 1. *Every symmetric function can be computed by a network with $O(\log n)$ bits of memory per node and this result is tight.*

Interestingly we show that every digit in the binary representation of r can be computed with $O(\log \log n)$ memory per node.

Theorem 2. *Every digit in the binary representation of r can be computed by a network with $\log \log n + 2$ bits of memory per node.*

While we currently do not have an explicit example of a symmetric function that requires more than $O(1)$ memory per node, we conjecture that the threshold function which is 1 if $\frac{r}{n-r} \geq \tau$ for an irrational number τ requires $O(\log n)$ memory per node.

For the case of k different colors we consider statistics that can be determined by comparison between the number of nodes having a given color, such as the plurality color, the least common color or the median color. While it easy to construct protocols for computing the plurality where the memory requirement per node is $O(k^2)$ or $O(k)$ using the majority protocol of [7], we provide a protocol that computes the plurality with a much smaller number of bits. We show that $O(\log k)$ bits of memory per node suffice to compute plurality and this is optimal,

Theorem 3. *The plurality function over k colors can be computed by a network with $O(\log k)$ bits of memory per node.*

The proof extends to the computation of statistics that can by evaluated using comparison trees of depth $O(\log k)$. All of the protocols that we describe are independent of the geometry of the graph and the expected time for convergence is poly(n).

References

[1] Moez Draief and Milan Vojnovic. Convergence Speed of Binary Interval Consensus. *Microsoft Technical Report, MSR-TR-2009-86*, 2009.

[2] Dinkar Vasudevan Etienne Perron and Milan Vojnovic. Using Three States for Binary Consensus on Complete Graphs. *Microsoft Technical Report, MSR-TR-2008-114*, 2008.

[3] M. Kearns, S. Judd, J. Tan, and J. Wortman. Behavioral experiments on biased voting in networks. *Proceedings of the National Academy of Sciences*, 106(5):1347, 2009.

[4] M. Kearns, S. Suri, and N. Montfort. An experimental study of the coloring problem on human subject networks. *Science*, 313(5788):824, 2006.

[5] M. Kearns and J. Tan. Biased voting and the democratic primary problem. *Internet and Network Economics*, pages 639–652, 2008.

[6] B. Latané and T. L'Herrou. Spatial Clustering in the Conformity Game: Dynamic Social Impact in Electronic Groups* 1. *Journal of Personality and Social Psychology*, 70(6):1218–1230, 1996.

[7] E. Mossel and G. Schoenebeck. Reaching Consensus on Social Networks. *Innovations in Computer Science, ICS*, 2009.

Social consensus through the influence of committed minorities

J. Xie* and B. K. Szymanski†

Dept. of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180

S. Sreenivasan‡

Dept. of Computer Science, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180 and

Dept. of Physics, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180

G. Korniss§

Dept. of Physics, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180

W. Zhang¶ and C. Lim**

Dept. of Mathematics, Rensselaer Polytechnic Institute, 110 8th Street, Troy, NY 12180

We study how the prevailing majority opinion in a social network can be rapidly reversed by a *committed* minority who consistently proselytize an opinion contrary to the majority, and are uninfluencable. There are many precedents for such events in human history, and it could be argued that several social upheavals have been facilitated through the action of such inflexible minorities, the suffragette movement, and the American civil rights movement being two such examples.

Previous work on opinion spreading and the diffusion of innovations has relied on models such as the Bass model [1] and the Threshold model [2]. While these models are suitable for studying influence spreading in the case where investment in a new idea, behavior or opinion comes at a cost, they do not account for the case of competing opinions where switching one's state has little overhead.

In order to address the latter case, we focus on an opinion dynamics model which we call the *binary agreement* model [3]. In this model, each node can either possess one of the two competing opinions, or both the opinions simultaneously - the presence of this "intermediate" state is what distinguishes it from the voter model. In addition to being useful in modeling opinion evolution where the cost of changing one's opinion is low, the binary agreement model is

*Electronic address: xiej2@cs.rpi.edu

†Electronic address: szymansk@cs.rpi.edu

‡Electronic address: sreens@rpi.edu

§Electronic address: korniss@rpi.edu

¶Electronic address: zhangw10@rpi.edu

**Electronic address: limc@rpi.edu

amenable to situations where changes in state are not deliberate or calculated, but unconscious [4]. Furthermore, by its very definition, the model may be applicable to situations where agents while trying to influence others, simultaneously also have a desire to reach global consensus [5].

The evolution of opinions in this model takes place through the following rules: at each time step, a randomly chosen speaker from the social network voices a random opinion from his list of opinions to a randomly chosen neighbor, designated the listener. If the listener has the spoken opinion in his list, both speaker and listener retain only that opinion, else the listener adds the spoken opinion to his list. The initial condition that we focus on is one where a fraction of $p < 0.5$ nodes possess opinion A and the rest of the nodes possess opinion B. In addition, we assume that the initial minority opinion holders are committed i.e. they can influence other nodes to alter their opinions, but are un-influencable themselves. Given these initial conditions, the only absorbing fixed point of the system is the consensus state where all influenceable nodes adopt opinion A - the opinion of the committed nodes. The question that we specifically ask is: *how does the consensus time vary with the size of the committed fraction?* More generally, our work addresses the conditions under which an inflexible set of minority opinion holders can win over the rest of the population.

Specifically, we show that when the committed fraction grows beyond a critical value $p_c \approx 10\%$, there is a dramatic decrease in the time, T_c, taken for the entire population to adopt the committed opinion. In particular, for complete graphs we show using a quasistationary approximation that when $p < p_c$, $T_c \sim \exp(\beta(p)N)$ (where $\beta(p) > 0$), while for $p > p_c$, $T_c \sim \ln N$. We conclude with simulation results for Erdős-Rényi random graphs, and scale-free graphs which show qualitatively similar behavior.

[1] F. M. Bass, Management Science **15**, 215 (1969).

[2] M. Granovetter, Am. J. Sociol. **83**, 1420 (1978).

[3] X. Castelló, A. Baronchelli, and V. Loreto, Eur. Phys. J. B **71**, 557 (2009).

[4] N. A. Christakis and J. H. Fowler, New Engl. J. Med. **357**, 370 (2007).

[5] M. Kearns, S. Judd, J. Tan, and J. Wortman, PNAS **106**, 1347 (2009).

ON THE EQUILIBRIUM OF BINARY CHOICE MODELS WITH POSITIVE EXTERNALITIES

J. TIPAN VERELLA, STEPHEN D. PATEK

Abstract. Models of binary decision with positive externalities are of practical significance since they enhance the understanding of social and economic processes. Analysis of such model often relies on assuming a continuum of agents or particles whose types are drawn from a distribution. We show that the problem of characterizing the equilibrium of a threshold model of collective behavior is a exactly solvable on a complete graph. We provide the distribution of equilibrium percentage of particles/agents switching to the new state and the distribution of the time to reach this equilibrium. We can recover the folk fixed point result by taking a continuum limit.

We consider a particle system with n particles labeled $i \in \{1, \ldots, n\}$. A particle i can be in either of two states, $s_i(t) \in \{0, 1\}$ at discrete time t, indicating that the particle is active, $s_i(t) = 1$, or not. All particles start out being inactive. To each particle i we associate an idiosyncratic constant Π_i, which are *iid* according to a distribution function F. We also attribute to a particle i a function $U_i(t) = B(\Pi_i, X(t))$, the utility or energy of the particle at time t; where $B(\Pi, X)$ is an non-decreasing function of both Π and X, and $X(t)$ is the percentage of particles that are active at time t, i.e. $X(t) = \frac{1}{n}\sum_{i=1}^{n} s_i(t)$. A particle becomes active whenever $U_i(t) > 0$. Note that since $B(\Pi, X)$ is non-decreasing in both Π and X, $X(t)$ is a monotone non-decreasing sequence in t. The underlying graph structure for this model is simply the completely connected graph on n vertices. We are interested in characterizing the equilibrium percentage of active particles as a function of the distribution function F, the function $B(.,.)$ and the number of particles n.

The object of this work is to supplement our understanding of this kind of model by providing a description of the above particle systems when $n < \infty$. We will refer to the model described above as a model of binary choice with positive externalities (BCPE) after [Sch73], although the form that it takes in this paper is closer to the formulation provided in [Cab90], in which the model is applied to the study of diffusion of innovations. Models of diffusion of innovations, [Cab90, FB95, GDGP$^+$02, MPV09, SJGH10, Kem10], can be framed as models of binary choice with positive externalities, the choice being whether or not to adopt the innovation. The BCPE model has been related in the literature to the random-field Ising model [SDP04, SGN08, GNPS09] or referred to as a model of cascades [Wat02, HMG10] in situations where the underlying graph structure is more intricate than a complete graph. Unlike those models of cascades on networks, in our simpler setting, we are able to pursue an exact description of the equilibrium.

In order to analyse the BCPE model, we write the following recursive definition for the percentage of active particles at time t:

$$\begin{cases} X_n(0) = & 0 \\ X_n(t+1) = & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \mathbf{1}\left\{ B(\Pi_i, X_n(t)) > 0 \right\} \end{cases} \tag{1}$$

For a given choice of F and $B(.,.)$, we are interested in X_n:

$$X_n := \lim_{t\to\infty} X_n(t)., \tag{2}$$

which can be interpreted as the equilibrium in which the most recent activation fails to encourage another particle to activate.

We note that X_n is a random variable with support on $\left\{0, \ldots, \dfrac{j}{n}, \ldots, 1\right\}$ for $j \in \{0, \ldots, n\}$, that the limit exists by the monotonicity of $X_n(t)$ in t, and that $X_n(t) = X_n(n)$ for all $t \geq n$.

Our main result follows.

Theorem 0.1. *Given a particle system of size n, define a configuration of the system $\sigma = (\sigma_0, \ldots, \sigma_n)$ with $\sigma_k \in \mathbb{Z}^+$ for all $k \in \{0, \ldots, n\}$ and $\sum_k \sigma_k = n$ to be an array indicating that σ_0 particles are such that they will activate on their own, σ_n particles will never become active, and σ_k particles will become active when k particles are active, but not before. Also let $\{p_k\}_{k=0}^n$ be a sequence of probabilities defined as follows:*

$$p_0 = \mathbb{P}\left\{B(\Pi, 0) > 0\right\} \tag{3}$$

$$\vdots$$

$$p_k = \mathbb{P}\left\{B\left(\Pi, \frac{k-1}{n}\right) < 0, B\left(\Pi, \frac{k}{n}\right) > 0\right\}, \quad k \in 1, \ldots, n-1 \tag{4}$$

$$\vdots$$

$$p_n = \mathbb{P}\left\{B\left(\Pi, \frac{n-1}{n}\right) < 0\right\}. \tag{5}$$

Furthermore, let $T = \min\left\{k > 0 \colon \sum_{i=0}^{k} \sigma_i < k\right\}$. Then the probability mass function of X_n is given by:

$$\mathbb{P}\left\{X_n = \frac{S}{n}\right\} = \sum_{\left\{\ \sigma_0 + \ldots + \sigma_T = S\ \right\}} \binom{n}{\sigma_0, \ldots, \sigma_n} \prod_{k=0}^{n} p_k^{\sigma_k}. \tag{6}$$

The continuum version of the above model where introduced in economics [Sch73] and in sociology [Gra78]. Models of binary decision with externalities are of practical significance since they enable a theoretical understanding of social and economic processes like diffusion of innovations [Gri57, Rog62, CKM66]. In the socio-economic context, one thinks of the particles as agents and arrives at the equilibrium percentage of active agents using the following heuristic. The marginal agent i^* is the last agent to find it beneficial to become active. This must mean that everyone before them had positive utility, while everyone after had negative utility. Hence the marginal agent has utility zero, i.e. Π_{i^*} solves $B(\Pi_{i^*}, X(t)) = 0$, and all agents with $\Pi_i \geq \Pi_{i^*}$ become active. Therefore the proportion of agents X^* to become active is the proportion of agents that are active as the marginal agent becomes active:

$$X^* = \int_{\Pi_{i^*}}^{\infty} dF(u). \tag{7}$$

We show that the continuum characterization can be obtained as $X^* = \lim_{n\to\infty} X_n$.

References

[Cab90] L. M. B. Cabral. On the adoption of innovations with network externalities. *Mathematical Social Sciences*, 19:299–308, 1990.

[CKM66] J.S. Coleman, E. Katz, and H. Menzel. *Medical Innovation: A Diffusion Study*. Bobbs-Merrill Co., 1966.

[FB95] H. Fukś and N. Boccara. Cellular automata models for diffusion of innovations. In *Proceedings of the Sixth Meeting on Instabilities and Nonequilibrium Structures*, 1995.

[GDGP+02] X. Guardiola, A. Díaz-Guilera, C. J. Pérez, A. Arenas, and M. Llas. Modeling diffusion of innovations in a social network. *Physical Review E*, 66(2), 2002.

[GNPS09] M. B. Gordon, J.-P. Nadal, D. Phan, and V. Semeshenko. Discrete choices under social influence: Generic properties. *Mathematical Models and Methods in Applied Sciences*, 19:1441–1481, 2009.

[Gra78] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6):1420–1443, 1978.

[Gri57] Z. Griliches. Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4):501–522, 1957.

[HMG10] A. Hackett, S. Melnik, and J. P. Gleeson. Cascades on a class of clustered random networks. *CoRR*, December 2010.

[Kem10] D. Kempe. *Structure and Dynamics of Information in Networks*. 2010.

[MPV09] A. C. R. Martins, C. B. Pereira, and R. Vicente. An opinion dynamics model for the diffusion of innovations. *Physica A: Statistical Mechanics and its Applications*, 388:3225–3232, 2009.

[Rog62] E. M. Rogers. *Diffusion of Innovations*. Free Press of Glencoe, Macmillan Company, 1962.

[Sch73] T. C. Schelling. Hockey helmets, concealed weapons, and daylight saving: A study in binary choices with externalities. *Journal of Conflict Resolution*, 17(3):381–428, July 1973. Discussion Paper no. 9.

[SDP04] J. P. Sethna, K. A. Dahmen, and O. Perkovic. Random-Field Ising Models of Hysteresis. *ArXiv Condensed Matter e-prints*, June 2004.

[SGN08] V. Semeshenko, M. B. Gordon, and J.-P. Nadal. Collective states in social systems with interacting learning agents. *Physica A*, 387:4903–4916, 2008.

[SJGH10] S. Sen, Y. Jin, R. A. Guérin, and K. Hosanagar. Modeling the dynamics of network technology adoption and the role of converters. *IEEE/ACM Transactions on Networking*, 18:1793 – 1805, 2010.

[Wat02] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99:5766–5771, 2002.

Mining social networks for customer churn prediction

Wouter Verbeke[*,a], Karel Dejaeger[a], Thomas Verbraken[a], David Martens[b], Bart Baesens[a,c]

[a] *Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium*
[b] *Faculty of Applied Economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium*
[c] *School of Management, University of Southampton, Highfield Southampton, SO17 1BJ, United Kingdom*

Abstract

Customer churn prediction models aim to detect customers with a high propensity to attrite. This study investigates the applicability of relational learning techniques to predict customer churn using social network information. A range of existing, extended, and novel relational classifiers and collective inference procedures have been (re-) implemented and applied on two large-scale real life data sets obtained from international telco operators, containing both networked (call detail record data) and non-networked (usage statistics, socio-demographic, marketing related) information about millions of customers. The results of the experiments indicate the existence of a limited but relevant impact of network effects on customer churn behavior. Combining a relational and a local classifier therefore improves the predictive power of a customer churn prediction model compared to a stand-alone local or relational classifier. Collective inference procedures however are shown to have a negative impact on the classification performance.

1. Methodology and experimental setup

1.1. Social network mining for customer churn prediction

Huge amounts of networked data on a broad range of network processes and information flows between interlinked entities are available, such as for instance call logs linking telephone accounts (Dasgupta et al., 2008), money transfers connecting bank accounts, or hyperlinks relating web pages (Neville and Jensen, 2007). These massive data logs potentially hide information that is extremely valuable to companies and organizations, but as well is extremely difficult to discover due to the size and the fragmentation of the data.

Networked data present both complications and opportunities for predictive data mining. The data are patently not independent and identically distributed, which introduces bias to learning and inference procedures (Jensen and Neville, 2002; Macskassy and Provost, 2007). Relational learning aims to exploit the information contained within the network structure of data instances, and to incorporate this information within a network classification or regression model (Džeroski and Lavrač, 2001; Getoor and Taskar, 2007). The aim of this study is to apply and develop relational learners to predict customer churn using social network information derived from call detail record (CDR) data, containing a vast amount of communication logs between customers of a telecom operator.

Macskassy and Provost (2007) introduced a node-centric, modular framework with network learners consisting of a local model, a relational model, and a collective inference procedure.

[*]Corresponding author. Tel. +32 16 32 68 87; Fax +32 16 32 66 24
Email addresses: wouter.verbeke@econ.kuleuven.be (Wouter Verbeke), karel.dejaeger@econ.kuleuven.be (Karel Dejaeger), thomas.verbraken@econ.kuleuven.be (Thomas Verbraken), david.martens@econ.kuleuven.be (David Martens), bart.baesens@econ.kuleuven.be (Bart Baesens)

ID	# nodes	# edges	CDR time span	# mean degree
1	1.365.451	2.446.672	3 months	3.58
2	8.337.285	16.494.177	6 months	3.96

Table 1: Summary of data set characteristics: ID, number of customers or nodes, total number of links between the nodes, call data time span, and the average degree calculated according to Nanavati et al. (2008).

This framework will be followed and applied in this study in order to compare stand-alone versions of network learners with combinations of a local classifier and a network model, both with and without collective inference procedures. To this end, two large real life data sets have been obtained from international telco operators. The characteristics of these data sets are summarized in Table 1. Data set 1 does not include local attributes for the nodes, whereas data set 2 contains 58 or 119 local attributes for the customers depending on the contract type.

1.2. Relational classification techniques

Due to the large scale of the networks, a number of existing relational learners have been re-implemented using sparse and parallel computation techniques. However, the time complexity of certain relational learning techniques (such as, e.g., the network-only Bayes Classifier, (Chakrabarti et al., 1998)) appeared prohibitive for application to large scale networks (cfr. infra, future research).

Furthermore, adapted versions of computationally expensive collective inference procedures, i.e. Gibbs sampling (Geman and Geman, 1984) and Iterative Classification (Lu and Getoor, 2003), have been developed with a reduced complexity, by making inferences concurrently for the entire network in each iteration, instead of node by node within each iteration as currently the case.

Relational learners typically restrict the impact of the network on a node to the first order neighborhood, i.e. the nodes in the network that are directly connected to a particular node (e.g., Macskassy and Provost (2007); Neville and Jensen (2007)). However, in many applications first order Markov behavior is violated, and higher order nodes in the network have an impact which should be accounted for. Therefore, a module has been developed which can be applied in combination with any network learning technique to incorporate the impact of higher order neighborhood nodes. This problem is closely related to the All Pairs Shortest Path problem (APSP) (Seidel, 1992) and the distance or minimum-plus matrix product. An adjusted version of the minimum-plus matrix product has been developed to incorporate higher order nodes for each node in the network represented by the adjacency or weight matrix, thus allowing to incorporate non-Markovian behavior in network learners. The first order network weight matrix is similar to the first order distance matrix in APSP, but not identical, since weights are the inverse of distances as a *closeness measure*, and therefore existing APSP solutions require adjustment.

2. Preliminary experimental results, discussion and future research

Figure 1 plots the call graph around a specific customer up to degree eight, with churners represented by red dots, and non-churners represented by blue dots, measured over a time frame of three months. The length of the links indicates the *closeness* between two connected customers (except for connections between customers from the different *branches* starting from the central customers, these are not proportional to closeness), i.e. the more contact between customers, the closer they are in the network. Multiple sequences of connected churners (with a maximum length of 12 sequential churners) can be observed in this network, indicating the existence of propagation of churn and social network effects throughout the customer network. It are exactly these social network effects that a relational classifier aims to model in order to predict customer churn.

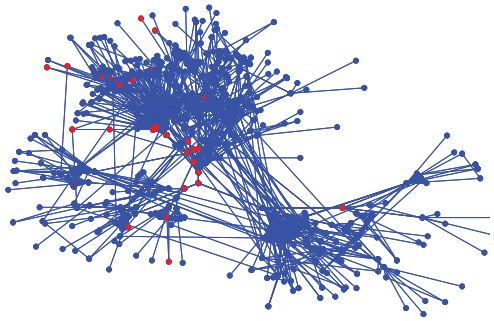


Figure 1: The call graph around a customer to degree 8, with churners represented by red dots, and non-churners represented by blue dots.

The results of a number of experiments, in which a range of existing, adjusted and extended network learners have been applied to the data sets summarized in Table 1, confirm that network effects do exist in a customer churn prediction setting (although only for a small fraction of the customers), and therefore should be taken into account by customer churn prediction models in order to improve classification power. Collective inference procedures however do not appear to improve classification power, and even have a negative impact. Finally, preliminary experiments indicate that increasing the order of the network neighborhood of the adjacency or weight matrix in order to model non-Markovian network behavior appears to have a minor positive effect, which is nonetheless existent and improves classification accuracy.

Currently, various experiments are ongoing, such as combining local and network classifiers, increasing the order of the network neighborhood, and the application of newly developed network classification algorithms. As a topic of future research, specific targeted sampling techniques, based on node and community diversity, will be developed in combination with active learning techniques in order to deal with scalability issues, as well as to improve learning. Generally, sampling in networked data is believed by the authors to be an important issue for future research in network learning. A second avenue of future research deals with including time-dependencies within network learning.

Acknowledgements

We extend our gratitude to the Flemish Research Council for financial support (FWO Odysseus grant B.0915.09), and the National Bank of Belgium (NBB/10/006).

References

Chakrabarti, S., Dom, B., Indyk, P., 1998. Enhanced hypertext categorization using hyperlinks. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. pp. 307–319.

Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A. A., Joshi, A., 2008. Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on Extending Database Technology: Advances in database technology, EDBT '08. pp. 697–711.

Džeroski, S., Lavrač, N., 2001. Relational Data Mining. Kluwer, Berlin, Germany.

Geman, S., Geman, D., 1984. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence 6, 721–741.

Getoor, L., Taskar, B., 2007. Statistical Relational Learning. MIT Press, Cambridge, MA, USA.

Jensen, D., Neville, J., 2002. Linkage and autocorrelation cause feature selection bias in relational learning. In: Proceedings of the 19th International Conference on Machine Learning. pp. 259–266.

Lu, Q., Getoor, L., 2003. Link-based classification. In: Proceedings of the 20th International Conference on Machine Learning (ICML). pp. 496–503.

Macskassy, S., Provost, F., 2007. Classification in networked data. Journal of Machine Learning Research 8, 935–983.

Nanavati, A. A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S., Joshi, A., 2008. Analyzing the structure and evolution of massive telecom graphs. IEEE Transactions on Knowledge and Data Engineering 20 (5), 703–718.

Neville, J., Jensen, D., 2007. Relational dependency networks. Journal of Machine Learning Research 8, 653–692.

Seidel, R., 1992. On the all-pairs-shortest-path problem. In: Proceedings of the 24rd Annual ACM Symposium on Theory of Computing, Victoria, Canada. pp. 745–749.

Making sense of the Chat Dialogs: the network, the communities and the text

Fahad Shah

University of Central Florida
sfahad@cs.ucf.edu

1 Introduction

There has been growing body of work and recent interest in the social networks like Twitter and Facebook alongside their phenomenal growth and impact on our lives. However, we feel that there exists a hidden component of information within these networks that can help to get better understanding of the 'real' social network of a person. For instance, a person may have 100's number of friends but might be actually 'friends' with only a short number of them. What is there that can help us make that distinction? We believe that text analysis is an important portion of the social network analysis and a lot of it is hidden and can't just be captured by 'links' alone. Also, we have availability of Masssive multi-player online social games like Second Life, that allow for much greater interaction than that is possible on facebook like sites in terms of expressivity, the user-experience and services. Similar to Facbook in having friends, Second Life had the kind of spatio-temporal 'situated' presence that the user can opportune to interact with the other people in the surrounding. This posits the question that we asked earlier about inferring the 'real' social network of the people, which is more challenging in Second Life in that there is no publically accessible way to get the friends list of a person (as opposed to Facebook). This forms the basis of our research: making sense of the social interactions among people in Second Life. In this abstract our focus is on chat dialog analysis within Second Life. When we embarked on this research project, we collected across 4 weeks of data more than 180 thousand dialogs across eight regions in Second Life (selected on varying category of the region and the user traffic) on the public chat channel, of which we separated out the 2 weeks of data for which we were able to collect enough chat dialogs across all regions (see [1] for data collection details).

We devised two methods for social network extraction from dialogs. This is a more challenging task, for we are operating in an environment where the users can enter or leave the chat at any time and the data consists of multiple overlapping conversations with variagated topics. An obvious first challenge was to identify the user the conversation is directed to for establishing the link between two users The first of the two algorithms, uses a time overlap, where we establish a link between a user and all other users within a twenty minute (default logout timeout after inactivity in SL) window, for the first and last occurrence from the user. This resulted in a raw baseline that we can compare our more refined algorithm that relies on the shallow semantic information (essentially rule based approach), which we call Shallow semantic temporal overlap or

SSTO (for details see [2]). The SSTO algorithm relies on occurrence of a username, a salutation (hi etc.), a question word (who/what etc.), second person pronouns and previous conversation information for the current user in determining the to/from labeling for the chat dialog. Once we have the to/from labeling for the chat dialogs, we can then construct the social network of the users based on these exchanges. In ([2]), we show that we can further use the community detection on the resulting social network to uncover the latent group structure present within this network using state-of-the-art community detection algorithms. This added information can then be used in refining the to/from labeling for the dialogs - in the event that we are unsure about the labeling we can look into a previous user from the same community the current user is in, for the last spoken utterance and establish a link between them for SSTO. Similarly for the temporal overlap, we can restrict ourselves to constructing links only between the users that are from the same community , thus filtering out the spurious links. To establish a quantitative measure, we compared the results from the original, community enhanced and hand-labeled dialogs for 1 hour of data. The results (see [2] for details) in table 1 for the frobenius norm, show that addition of community information adds little to the SSTO algorithm, while greatly improving upon the temporal overlap algorithm. This shows that the addition of community information is a useful measure that can refine the link detection algorithms.

Table 1. Frobenius norm: comparison against hand-annotated subset.

	SSTO	SSTO+LC	SSTO+SC	TO	TO+DT	TO+HT
Help Island Public	35.60	41.19	46.22	224.87	162.00	130.08
Help People Island	62.23	60.50	66.34	20.29	20.29	54.88
Mauve	48.45	45.11	51.91	58.44	58.44	49.89
Morris	24.67	18.92	20.76	43.12	37.54	38.98
Kuula	32.12	30.75	32.66	83.22	73.15	77.82
Pondi Beach	20.63	21.77	21.56	75.07	62.62	71.02
Moose Beach	17.08	18.30	21.07	67.05	53.64	50.97
Rezz Me	36.70	39.74	45.78	38.72	39.01	41.10
Total error	277.48	**276.28**	306.30	610.78	507.21	514.74

In Future, we plan to enhance our understanding of the chat dialogs by utilizing the content, that we thus far have ignored save for the shallow heuristics (we chose a rule based approach in first place because we wanted our technique to be unsupervised and able to work on large data-set in reasonable time period).

References

[1] Shah, F., Usher, C., Sukthankar, G.: Modeling group dynamics in virtual worlds. In: Proceedings of the Fourth International Conference on Weblogs and Social Media. (2010)

[2] Shah, F., Sukthankar, G.: Constructing social networks from unstructured group dialog in virtualworlds. In: Proceedings of the International conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP). (2011)

EXPLOITING SOCIAL NETWORK ANALYSIS FOR CAREER PATHS RECOMMENDATIONS.

María F. González-Gutiérrez[1], Oleg Morajko[1], Marc Pou Miquel[2], David Monreal[2],
[1]InnoQuant, Barcelona, Spain, {mafe, oleg}@innoquant.com,
[2]InfoJobs, Barcelona, Spain, {marc.pou, david.monreal}@infojobs.net.

Today, unemployment is one of the most pressing socio-economic problems in the European Community. In particular, Spain ranks first in unemployment rate among young people under 25 years. Moreover, according to data from Eurostat [1] most people have found their last job through recommendations from family and friends, despite being active job seekers on the Internet.

The online job search market has rapidly evolved and has attracted millions of candidates expecting improvement in the process of looking for jobs, and recruiters expecting to find them. There are large employment databases where both candidates and employers put their information, but the knowledge from that data has not yet been fully exploited.

InfoJobs.net is the leading job site in Spain with about 47% of market share in 2009 and is ranked third in Europe. The site manages more than 60.000 employers and 4.5 millions of candidates. Last year the site registered 800.000 job offers and more than 51 millions of subscriptions that lead to 250.000 job deals. It means that each employer contributes an average of 13.3 jobs and that each offer receives an average of 63 subscriptions per year.

The wealth of available information enables studying job market trends and user behavior on the education and employment decisions during their professional careers. This collective knowledge can be exploited to help users discover career paths that suit them best and promise to meet their expectations.

We develop a large-scale graph-based prediction model of labor market and a recommendation system to provide answers to the questions of "*What options you have in your career path?*", "*Do you need to enhance your skills to move ahead?*", "*How you compare with people like you?*", "*What is the social influence of an education center?*" and generate suitable recommendations for both job seekers and employers. As of this writing, our model has been built from about 2 millions of real-world curriculum vitaes, including information on job experiences, skills and education data, and is growing due to new subscribers since the beginning of the year.

In our approach, we build a social network model by representing individual actors as nodes and their relationships as links within *career-path graph*. A job seeker is related to education centers by its academic career and to positions, companies and industry sectors through its professional trajectory. Employers generate offers directed to the candidates and education centers provide them training necessary to obtain education and qualifications. The relationships between the actors are characterized by duration, management level, salary and other relevant attributes.

The model data is captured and updated by automated harvesting of CV database and pre-processed by data normalization and de-duplication tools that help correcting data incoherencies and complete missing data. These steps result crucial when working on real-world datasets collected over years from different users and systems and suffer different kinds of data incoherency problems.

To provide recommendations we use a hybrid approach that combines collaborative filtering [2, 3, 4, 5] techniques with social network analysis [6, 7, 8, 9] into *graph-based collaborative filtering* [10]. This approach operates on career-path graph, and also includes attributes of job seeker nodes, offer nodes, and potentially other relevant entities from a database to make hybrid recommendations. We discover and group similar career paths by analyzing user links in the graph-path perspective and finding cohesive subgroups.

Our recommender offers job seekers personalized suggestions on training curses, missing skills, and professional experiences to guide them into successful career paths. We discover career path patterns of similar candidates and apply graph-based influence scoring [11, 12, 13] functions to rank the suggestions. For example, we measure social and economical impact of each education center on the society in function of jobs, positions, management level and salaries achieved by their graduates (Figure 1). In that sense, highly influential centers (such as MBA centers) are more likely to be recommended to candidates with career paths aiming at executive jobs and high salary ranges.

In this paper, we apply social network analysis to make better recommendations. We find that graph-based approach to be a powerful tool that provides great potential in making recommendations. Moreover, we demonstrate that out approach operates in high-traffic environment, produces results in real-time, and scales well to large datasets.

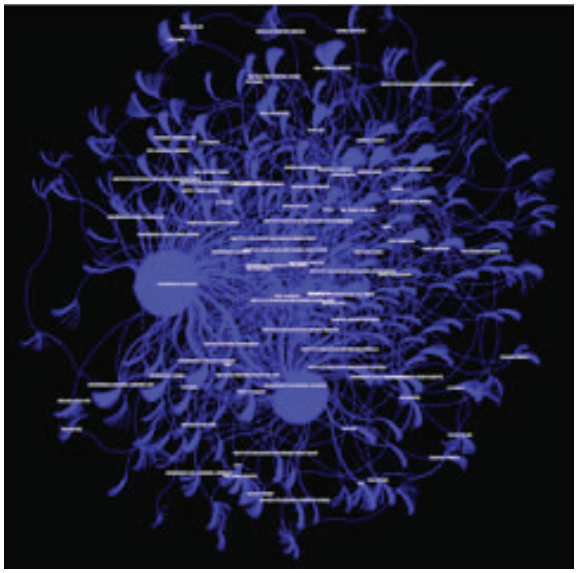


Figure 1: Example education centers network in Valencian Community in Spain. The graph shows relationships between individual candidates and education centers as declared in their curriculum vitaes. The network is composed of two big communities that correspond to public universities (Universidad de Valencia and Universidad Politécnica de Valencia) and a number of small education centers. The candidates that studied in a pair of centers enforce the relationship between the centers.

References:

[1] European economic statistics, 2010 Edition, Eurostat Statistical Books, CEE, Vol. 1, Belgium, 2011.

[2] J. Bennett and S. Lanning. The Netflix Prize. In Proc. of KDD Cup Workshop at SIGKDD-07, 13th ACM Int. Conf. on Knowledge Discovery and Data Mining, pages 3–6, San Jose, California, USA, 2007.

[3] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1): pages 5–53, 2004.

[4] D. Tikk, I. Pilászy, B. Németh, D. Tikk. Scalable Collaborative Filtering Approaches for Large Recommender Systems. Journal of Machine Learning Research, Vol. 10, Pages 623-656, 2009.

[5] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In Proc. of WWW-01, 10th Int. Conf. on World Wide Web, pages 285–295, Hong Kong, 2001.

[6] S. Wasserman, K. Faust. Social Networks Analysis: Methods and Applications. Cambridge University Press, Cambridge, 1994

[7] D.J. Watts. Six Degrees: The Science of a Connected Age. W. W. Norton & Company, 2004.

[8] Albert, R. and A.-L. Barabasi. Statistical mechanics of complex networks. Reviews of Modern Physics, 74, pages 47-97, 2002.

[9] Barabasi, A.-L. and R. Albert. Emergence of scaling in random networks. Science, 286, pages 509-512, 1999.

[10] Huang, Z., W. Chung and H. Chen. "A graph model for e-commerce recommender systems," Journal of the American Society for Information Science and Technology (JASIST), 55(3), pages 259-274, 2004.

[11] Kleinberg, J. Authoritative sources in a hyperlinked environment. In the Proceedings of ACMSIAM Symposium on Discrete Algorithms, 1998.

[12] Freeman, L. C. Centrality in social networks: Conceptual clarification. Social Networks 1(3), pages 215–239, 1979.

[13] R. Ghosh, K. Lerman, Leaders and Negotiators. An Influence-based Metric for Rank. Proceedings of 3rd International Conference on Weblogs and Social Media, 2009.

The Market Economy of Trips

Dimitris Papanikolaou, MIT Media Laboratory

dimp@media.mit.edu

The Market Economy of Trips (MET) investigates the potential of using market incentive mechanisms and a visual information system to create sustainable, self-organizing, one-way vehicle sharing systems. That is systems requiring minimum central intervention.

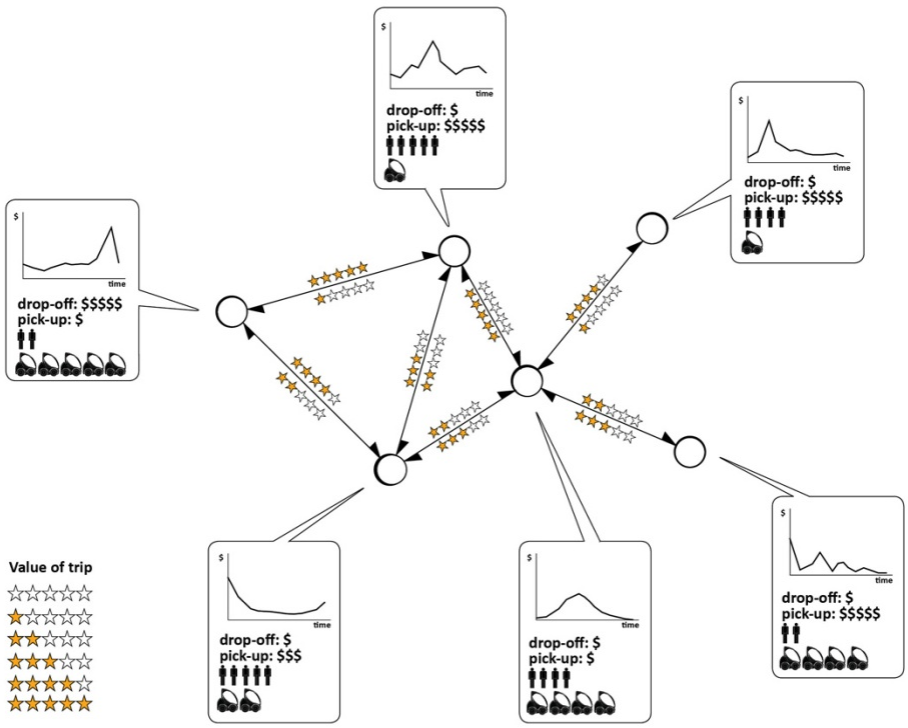
One-way vehicle sharing systems are distributed urban mobility networks of vehicles and parking stations that allow users to conveniently pick up a vehicle from any station and drop it off to any other station. Popular examples are bike sharing programs however this trend is rapidly entering automobile markets.

Despite their great convenience vehicle sharing systems have drawbacks too. Due to asymmetric demand patterns, eventually all vehicles are ending at the stations with no demand. This inventory imbalance, not only decreases throughput, but it furthermore increases trip time as drivers search for parking spaces. Existing policies redistribute manually vehicles, which is a complex, inefficient, and highly unsustainable solution. Not only it is operationally complex, but furthermore it is expensive: either the fleet needs to be too large or the redistributions need to be too frequent. In addition continuous redistributions keep vehicles away from the system reducing further service capacity. As a consequence, many vehicle sharing systems end up wasting more resources for sustaining their performance than the value of the service they provide.

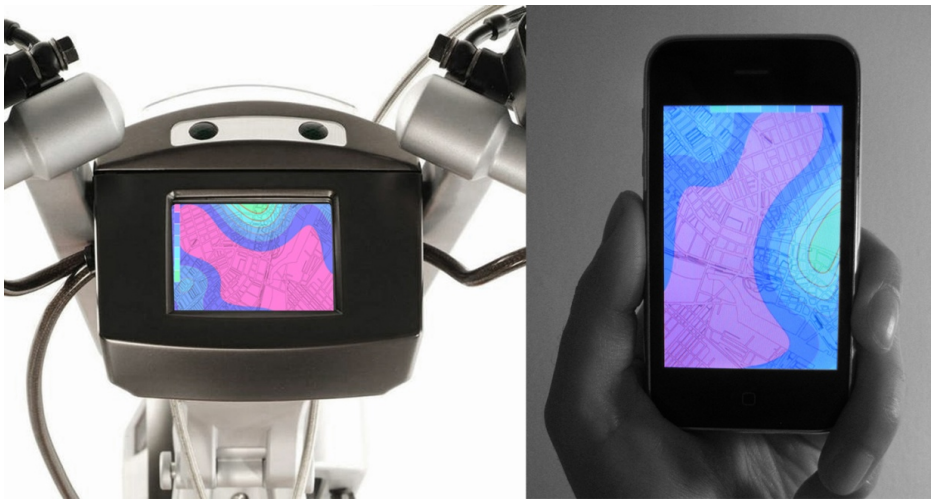
In this research I am exploring a new strategy to create autonomous self-organizing vehicle sharing systems that uses price incentives to smooth demand imbalances, and PriceScape, an interactive mapping tool and graphical user interface to intuitively communicate location-based price information to the users. Similarly to a market economy, prices adjust dynamically to parking needs incentivizing users to drive vehicles to stations that mostly need them while discouraging arrivals to stations that don't need them. PriceScape uses dynamic heat map display and isometric price curves to describe areas with similar payoffs. Like the analogy of navigating through a price landscape, climbing from valleys to hills is expensive, while descending from hills to valleys is rewarding.

This work explains decision-making in dynamically priced mobility systems, investigates the conditions under which a stable equilibrium state may exist, and if so, whether local price calculation and visual perception of the price landscape is sufficient to bring it. Is there a pricing policy that can make the system self-sustaining such that the funds from its overpaying users are enough to reward its underpaying users? How efficient can a dynamically priced vehicle sharing system be?

To address these questions I am developing both a game experiment to empirically evaluate how users' visual perception of payoffs affects decision making, and a computational framework using System Dynamics and Urban Economics, to explore the theoretical limits of efficiency of MET under different demand patterns, pricing policies and population's income distribution.



The Market Economy of Trips: trip value depends on pick-up and drop-off stations



PriceScape: interactive heat-map GUI displaying areas of high and low payoffs

Social distance drives the convergence of preferences in an online music sharing network

Lily Tran*, Manuel Cebrian**, Coco Krumme*, Sandy Pentland*
*MIT Media Lab
**UC San Diego
lilytran@mit.edu, mcebrian@cs.ucsd.edu, kak@mit.edu, sandy@media.mit.edu

We study a network of individuals belonging to Herzio (http:// herzio.com), a social music website in which information is shared dynamically via friendship links. On the site, participants can listen to songs, create friendships with others, and recommend songs to peers. In addition, it is possible to identify communities of nodes not necessarily linked by friendship.

Herzio is a relatively young network (less than two years old), which focuses on spreading the work of independent musicians; therefore, few exogenous shocks (i.e. hit singles shared by most of the network) exist. After removing individuals with abnormally high degrees of connectivity (defined as individuals with a degree greater than one hundred), the community included 3086 listeners. In order to study the behavior of only active individuals, we then reduced the data set to only include individuals who were active for a minimum of six months. The community studied and discussed represents 235 individuals.

Past work has highlighted the role of social influence in generating unpredictability and inequality in markets for cultural products [Salganik et al, 2006], as well as shown that social influence plays an informational role in certain systems [Krumme et al, 2011]. However, in these experiments, there was no network structure available to participants.

Other research has focused on the spread of behaviors, such as smoking, through social networks [Christakis and Fowler, 2008]. In the current work, we ask to what degree does network proximity drive the convergence of shared information.

In the Herzio network, individuals can both invite others to be "friends" (friendship is reciprocal and must be accepted by the invitee), as well as listen to a selection of songs available to participants throughout the network. We study individuals linked by direct friendship, as well as those who are "friends of friends", and measure the dynamic listening patterns of pairs of individuals over time. We find that closeness of the social tie predicts the rate of convergence of song preferences. This effect occurs independently of normative influence via direct song recommendations.

Moreover, not all friends are equal. For any individual, convergence occurs more quickly with the interests of some friends over others. Over the entire network, this "influence spectrum" is either normally distributed, or distributed normally with some friends exerting no influence on the preferences of the given individual. That is, the social network does not correspond to the preference network in a one-to-one fashion: individuals follow some friends for some things, and ignore others altogether.

Methods and Results

We quantify preference across the entire network as the proportion of songs shared by any pair of individuals to the total unique songs listened to by the pair. Over time, preference increases: that is, normalizing initiation time and filtering for those who remain active over six months, the preferences of individuals converge more quickly initially and then more slowly.

For each month similarity between two listeners was calculated as follows:

$$\text{Preference} = \frac{A \cup B}{A \cap B}$$

Where A and B represent the set of songs listened to by two listeners.

We define pairs of "friends" as those with a direct friendship link, pairs of "length 2" as those pairs who share a friend but are not themselves friends, and so forth. There exist differential rates of convergence, depending on distance of social connection, and we use these differences

to compare different aspects of social effects in an informational network. We find that the convergence of listening preference depends to a large extent on social distance. We test for significant differences in the rate of growth using a t-test, and find differences.

The more closely one is linked to one's social peers, the more musical tastes converge over time. This convergence occurs without direct (normative) influence and controlling for events or publicity exogenous to the network, and supports that observed with other elective behaviors in social networks [Christakis & Fowler]. Indeed it is a more important effect than that of recommendations or of shared community (Figure 1).

These results lend insight into the mechanism of social influence in networks of individuals who share information (here, listening patterns). Future work might consider the way in which a network restructures dynamically in response to changes in preferences.

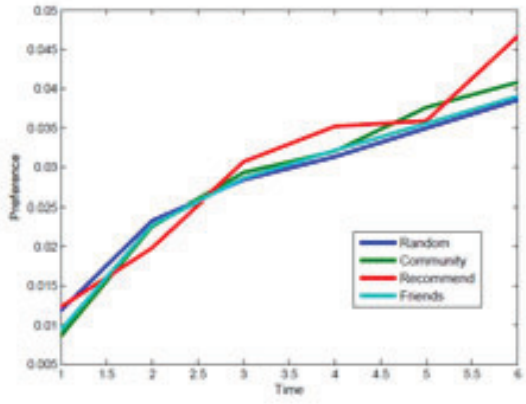


Figure 1: rates of convergence for shared listens of random links, members of the same community, friends who recommend to one another, and friends who don't recommend to one another. The Random label indicates two users who are not friends. The Community label indicates two users in the same community but are not friends. Here, the community assignment was determined with the Walktrap algorithm with random walks of length 2 [Pons & Latapy, 2005]. The Recommend label refers to users who are friends and recommend songs to each other. Lastly, the Friends label refers to friends who do not recommend songs to each other.

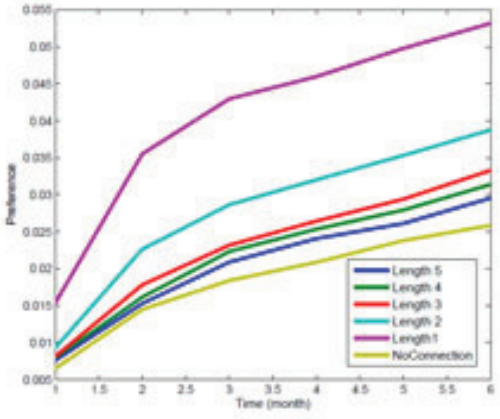


Figure 2: rates of convergence of shared listens for pairs of individuals 1, 2, 3, 4, and 5 hops away from one another in the network, as well as individuals with who are not connected (defined as completely unconnected individuals and individuals 6 or more hops away from each other).

References

Salganik et al, 2006. Experimental study of inequality and unpredictability in an artificial cultural market.
Krumme et al, 2011. A model of social influence in markets for cultural products
Christakis&Fowler, 2008. The collective dynamics of smoking in a large scale network
Pons&Latapy, 2005. Computing communities in large networks using random walks

Modeling the Coevolution of Network Structure and Node State in a Student Dorm

Wen Dong, Anmol Madan, Alex (Sandy) Pentland
(wdong, anmol, pentland)@media.mit.edu

In recent years, mobile phones have enabled researchers to record the minute-level behavior and interactions over hundreds people and many months. They have hence enabled researchers to observe diffusions and test social dynamics models with real-world face-to-face network data (Madan, Farrahi, Gatica-Perez, & Pentland, 2010).

We have developed a stochastic process model to describe the coevolution of network structure (such as who talks to whom) and node state (such as preferring Democratic vs. Republican), in a framework similar to physics models of social dynamics (Castellano, Fortunato, & Loreto, 2009). We use MCMC and variational (Wainwright & Jordan) methods to fit the parameters of our model, and proceed to answer questions such as, who is the most influential person, how widely distributed the individuals' influence is, and how the individuals' states are asymptotically correlated. Our model extends the influence model that was developed in LIDS (Asavathiratham, Roy, Lesieutre, & Verghese, 2001).

Specifically, we study a network with C nodes, each node $c \in \{1, \dots, C\}$ taking state from $\{1, \dots, m_c\}$, and each node c signaling node c' to change state with rate $h_{c\to c'}$ (i.e., node c has an influence $h_{c\to c'}$ over node c'): Node c in state $s_t^{(c)} = i$ at time t can signal node c' to change state to $s_{t+1}^{(c')} = j$ at time $t+1$ with rate $h_{c\to c'}\times a_{i\to j}^{c\to c'}$, where $\sum_j a_{i\to j}^{c\to c'} = 1$. (i.e., node c can distribute its influence $h_{c\to c'}$ into how much it wants node c' change to each of the $m_{c'}$ states). Hence the signaling rate of the whole network is $\sum_c \sum_{c'} h_{c\to c'}$, and the probability of a specific event is $h_{c\to c'}\times a_{s_t^{(c)}\to j}^{c\to c'} / \sum_c \sum_{c'} h_{c\to c'}$ when any signaling event happens at time *t*. The structure of the network is different when its nodes are in different states.

So far, we have been able to confirm that the change of political opinions and body weight is significantly related to the exposure to similar individuals ($p<0.001$ in both cases) but not to influence from self-declared friends or discussants.

Bibliography

Wainwright, M. J., & Jordan, M. I. *Graphical Models, Exponential Families, and Variational Inference.*

Asavathiratham, C., Roy, S., Lesieutre, B., & Verghese, G. (2001). The Influence Model. *IEEE Control Systems Magazine* .

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics* .

Madan, A., Farrahi, K., Gatica-Perez, D., & Pentland, A. (2010). *Pervasive Sensing to Model Political Opinions in Face-to-Face Networks.* Retrieved from http://web.media.mit.edu/~anmol/political-2.pdf

The "Friends and Family" Study: Progress Report and Initial Results

Nadav Aharony, Wei Pan, Alex (Sandy) Pentland
MIT Media Lab
Contact: {nadav, panwei, pendland}@media.mit.edu

Overview

The Friends and Family study in the MIT Media Lab is a long-term mobile phone-based experiment that transforms a graduate family community into a living lab for social science investigation. Data from this study, collected via Android-based phones equipped with our software platform for passive data collection, will be used to look at issues including individual and group identity, real world decision making, social diffusion, social health, and boundaries of privacy. In the talk we will briefly review the study, which has been running since March 2010, and update on the current status of the experiment, give an overview of our collected dataset, and highlight some preliminary results. We focus our initial analysis on patterns surrounding mobile applications (apps), as well as the results of a social-mechanism based in intervention that we have conducted with the study population . The goal of this talk is to engage the WIDS community with discussion of discuss our approach of living-laboratory experiments, our platform for conducting such experiments, and the potential of the unique dataset we have already assembled.

For the first analysis, we look at participants' app installation patterns and investigate the roles of different networks, inferred from Bluetooth proximity and self-reported surveys, in the spreading of apps. We find that face-to-face interactions have a stronger correlation with the number of shared apps between individuals than self-perceived ties. By the time of the WIDS workshop we plan to have a more thorough analysis of these patterns as well as additional results from other components of the study.

Introduction

Today's mobile phones are powerful computing and sensing platforms. We are investigating ways to help users leverage individual as well as aggregated data to improve their lives. Additionally, we are investigating how this data can contribute to the understanding of societal and especially community-related issues.

The Friends and Family study (FunF) is an experiment in the form of a living lab, with participants' everyday behavior patterns sampled via mobile phones and other data collection mechanisms. The pilot phase of the study ran from March to July 2010 with 55 participants, and the expanded second phase of the study will begin in September 2010 with around 130 participants. The data collected pertains to both the physical and digital realms and includes information on face-to-face interactions, mobility, phone communication networks, and online social network activity. The study team also has direct access to the participants in the forms of questionnaires, interviews, and various experimental interventions, giving the FunF study access to a tight-knit physical community at an unprecedented scale and depth. Considering the study will run at least 18 months, the dataset generated from the study will shed light on a wide range of behavioral, social, and health-related topics.

The study touches on many aspects of life, from social dynamics to health to purchasing behavior to community organization. The two high-level topics that unify these varied aspects are: **(a)** how people make decisions, especially the social aspects involved in decision making, and **(b)** how we can empower people to make *better* decisions using personal and social tools.

Study Components

The study is composed of four main components:

Android Phone Sensing Platform (FunF System): This is the core of the study's data collection. Android OS-based mobile phones are used as in-situ social sensors to map users' activity features, proximity networks, media consumption, and behavior diffusion patterns. The phones are augmented with our

software, which periodically senses and records information such as cell tower IDs;wireless LAN IDs;proximity to nearby phones and other Bluetooth devices; accelerometer and compass data; call and SMS logs; statistics on installed phone applications, running applications, media files, and general phone usage; and other accessible information. The system also supports integration of user-level apps, such as the alarm clock app we developed for additional data collection and potential use in interventions.

Surveys: Each participant has to complete surveys at regular intervals, currently set at weekly and monthly. These include self reports about their perception of their social relationships, groups, and interactions, logging of various types of activities and mood, and standard scales that examine different personality traits and states (e.g. the Big Five Personality Test [4]).

Purchasing Behavior: Information on purchases is collected through receipts and credit card statements submitted at the participants' discretion. This component targets a specific set of categories: child-related, entertainment, and dining expenses.

Facebook Data Collection Platform: Participants can optionally install a Facebook application to log different Facebook activities.

Initial Analysis of Mobile App Installations

In the pilot study, the 55 participants have installed around 870 unique apps over a period of 3 months. (not counting any apps that come bundled with the phone or the OS version). We discovered that people who spend more time in face-to-face interaction are more likely to share common apps. In fact, in our dataset, pairs with face-to-face interaction share on average two more common apps on their phones compared with pairs with little face-to-face interactions (avg of 2.7 apps in common vs 4.9 apps in common in the latter case). Those face-to-face interactions might include group activities, religion-related interactions, time spent withsignificant others and many other possibilities. However, we also observed that the self-reported friendships do not result in an increase in the number of common shared apps. We believe our results provide strong evidence on app diffusion patterns: apps do spread via socialinteraction. In particular, the diffusion of apps relies much more on the face-to-face interaction ties than the self-perceived friendship ties. Therefore, one should be cautious in using declared friendship networksto infer the spreading of smart-phone apps and for applying viral marketing strategies, since the face-to-face interaction seems to have a stronger correlation with app diffusion.

(a)	Group 1	Group2
BT Co-Location Closeness Range	[0,10]	(10,2000]
Mean #Common Apps / Pair	2.7253	4.9
ANOVA: F=74.48, p<0.0000001		
K-S test: True, p=7.8e-19		

(b)	Group 1	Group2
Self Reported Closeness Range	[0,1]	(1,10]
Mean #Common Apps / Pair	4.75	4.05
ANOVA: F=4.97, p<0.026		
K-S test: True, p=0.0045		

Table 1: Summary results for (a) Bluetooth proximity closeness and (b) Self reported closeness.

Social Mechanism Intervention

Between October – December 2010 we have conducted the "FunFit" game, an intervention aimed at investigating different motivation mechanisms relating to social health and wellness. The idea of FunFit was to test the effects of social capital (or social pressure, depending how it is viewed) for helping individuals achieve the goal of being more fit and active. Participants were divided into three groups based on different conditions, one group of participants received feedback on their own performance and were rewarded based on it, a second group was still rewarded based on the individual performance, but was also able to see the information about two other participants, while the third group saw the performance of two other participants and also were rewarded based on the peers' performance, and not their own. We will briefly mention some of the interesting initial results coming out of this intervention.

Information Flow in Networks — Trendsetters, Bellwethers and Shepherd Dogs

Yaniv Altshuler[1] , Alex (Sandy) Pentland[1]

Human Dynamics Group
MIT Media Lab
yanival@mit.edu , sandy@mit.edu

1 Abstract

The rapid expansion of "social network research" is an exciting and unique phenomenon taking place in the last decade. Networks, from their essential definition, serve as an infrastructure for communication — either explicit (e.g. phone calls or emails), implicit (e.g. social signals), or a combination thereof (e.g. online social service such as *Facebook* or *Flicker*). As such, a fundamental aspect that must be addressed in order to truly comprehend networks, is the particular way they interact with and influence information that flows through them. Specifically, when addressing this issue, our goals should be threefold.

First, we must aspire for the *understanding* the underlying mechanisms that control information flow in networks. This should be done by developing analytic models that describe the properties (both local and global) of this process.

Second, using these models (or other methods that will be developed based on them) to generate *predictions* concerning the evolution of information propagation processes. This may include for example the ability to predict the probability that some trend or idea will epidemically spread throughout a given network, . Another example would be the assessment of the *openness* of networks to the proliferation and assimilation of new information or knowledge that are introduced to them by some of their members.

Third, we are interested in developing techniques for efficiently *intervening* in this process. For example, such intervention may take the form of external stimuli that may be injected to the system in order to generate (or strengthen) certain social links, and subsequently facilitate the assimilation of future information that would be introduced to the network (i.e. the creation of "Shepherd dogs", that increase the cohesion of the network). Alternatively, by efficiently recognizing the implicit "roles" of members of the network, certain members can be engaged (e.g. by informative or monetary means) in a way that influence the global behavior of the entire network (i.e. "Bellwethers" and "Trendsetters").

In order to formally discuss the ability of a new idea or a new piece of information to gain popularity over a network, we formally define two terms : "*appeal factor*" and "*persistence factor*", denoting an idea's *local* spreading and deletion expected probabilities, respectively. Using these terms we show that the ability of an idea to epidemically spread throughout a network can be analytically

predicting, resulting in a predictor that tightly thresholds between decaying and prospering spread processes. We then demonstrate this method using several datasets containing information for real world online social networks.

We then discuss the challenges that are involved with the implementation of this technique, and suggest ways of overcoming them. Specifically, we propose a way to generate fast-to-calculate local-information-based predictors that may be able to approximate the results that can be achieved by a theoretical optimal algorithm. We conclude by presenting preliminary results concerning the implementation of such methods using real world online social networks.

Network Manipulation (with application to Political issues)

Panagiotis Takis Metaxas
Department of Computer Science
Wellesley College
Wellesley, MA 02481, USA
pmetaxas@wellesley.edu

I. Introduction

We live in an increasingly interconnected world, one in which a growing number of people turn to the web to make important medical, financial and political decisions [1]. As more people use the Web's search engines daily as their primary source for locating information on many important issues, search engines are in the position to influence what is perceived as relevant information through their mechanism of ranking web pages. However, as studies have shown [2], interested groups and individuals can also make use of web spamming mechanisms to trick search engines in ranking their pages higher than those of their rivals.

The battle for controlling the messages in cyberspace is spreading over many ideological, cultural, and political issues where controversial positions vie for the public support. For example, consider issues such as abortion legality and morality, children vaccination risks, creationism vs. evolution, homosexuality, etc. [3]. Nowhere this battle is more obvious than when it comes to issues related to national elections. Obviously, the stakes here are high. If one is able to influence the public elect officials that will be favorable to his/her agenda, this will have far-reaching implications.

The term "Web Spam" or Adversarial Web Search [4] is broadly used to describe misinformation planted on the Web. Google-bombs are probably the best known examples of Web spamming, because of their broad coverage in the press [5]. Web spammers are creating misinformation using "text spam" and "link spam", while using "cloaking" to cover their tracks. This generic categorization of practical actions does not explain why they are successful, or why Google Bombs may or may not be be successful.

Recently, spamming techniques have been introduced in Social Media [6], making it more appropriate to talk about "Network Manipulation" rather than just Web spam. The basic techniques of this manipulation, however, can be traced in the long history of propagandistic techniques in society [2].

In this presentation proposal we provide an overview of the technical side of Network Manipulation, and discuss its connection to Social Propaganda. Further, we describe how it has been extended in the area of Social Media and discuss some of its successes and failures when it comes to political issues, especially related to congressional elections. We end with a discussion of what the search engines have said to have done so far, and what is likely that they have done without admitting it.

Ranking Techn.	Net. manipulation	Soc. Propaganda
Doc Similarity	keyword stuffing	glittering generalities
Site popularity	link farms	bandwagon
Page reputation	mutual admiration societies	testimonials
Anchor text	Google bombs	card stacking
Real-time search	Twitter bombs	plain folks

Table I
Ranking techniques by search engines are listed along with the response of the web graph manipulators and their corresponding propagandistic techniques.

II. What is Network Manipulation

Network Manipulation is the attempt to modify the Web graph and/or a social network, and thus influence online network tools in ways beneficial to the manipulators. The modification of a network is in terms of altering its structure and/or its contents. The online network tools that manipulators try to influence are typically search engines and online social media.

One can explain most of the major technological developments in the area of search engine technologies as their attempt to stop the successes of Web Graph manipulators (See Table I). For example, Google's attempt to combat link farms (groups of interlinked web sites controlled by the same entity) with the introduction of the famous Page Rank algorithm was countered by the introduction of "mutual admiration societies" (organized groups of manipulators who have achieved high reputation independently for unrelated themes) [2]. In terms of propagandistic techniques, this corresponds to "testimonials" often used by advertising companies: A famous actor playing the doctor on TV urges the audience to buy a particular pain killer, as if he is an expert in medicine.

Each of the network manipulation techniques is implemented with altering the structure or the contents of the network's components. For example, in creating mutual admiration societies, the so-called "black hat" search engine

optimization companies organize themselves exchanging links [7]. To create Google bombs, they announce the terms to be targetted as anchor text and the links to support, sometimes even in the open [8], as we discuss next.

Ever since the appearance of the "miserable failure" hoax that produced search results that included President George W. Bush (and later, President Barack Obama [9]), Google bombs have attracted a lot of attention in the media, since they appear that with little effort, net manipulators can game the sophisticated algorithms of the search engines. This practice of "gaming" the search engines is implemented with mislabeled anchor text techniques (corresponding to the "card stacking" propagandistic technique; see Appendix**??**), in which web site masters and bloggers use the anchor text to associate an obscure, negative term with a public entity [10]. In particular, during the 2006 US midterm congressional election, a concerted effort to manipulate ranking results in order to bring to public attention negative stories about Republican incumbents running for Congress took openly place under the solicitation of the progressive blog, MyDD.com (My Direct Democracy) [8].

Can these efforts be stopped? Search engines have tried to counter this bad publicity by announcing initially a plethora of features for ranking, and recently more sophisticated algorithms than Page Rank [11]. These changes seemed to bear fruits in the 2008 congressional elections where very few spamming sites rose to the top of search results [12].

It appears, however, that these supposedly "new, sophisticated" (and secret) algorithms would not scale: They are likely pre-computed search results on a white-list of search query terms that would likely be bombed.

In particular, these new features and algorithms appeared to be highly effective in the announced launching of 98 Google bombs [13] during the 2010 congressional elections, as we predicted [14]. Both in 2008 and 2010, searches related to congressional candidates would bring up in the top six results a ranked list of the same sources: one or more of the campaign sites of the candidates, their official web pages, their wikipedia page, and google images (Figure 1). This was consistent across the candidates independently of their visibility or of the fact that they were under attack by thousands of political spammers. Moreover, the relative location of each result in a period of 29 weeks remained remarkably steady (Figure 2).

At the same time, they proved to be completely ineffective in at least two examples of network manipulation that were under their radar screen: The Decor-My-Eyes site [15] that successfully used bad publicity to rank high, and the JCPenney case [16] that used anchor text manipulation of link farms. After these cases became known, their relative position changed dramatically moving downwards by dozens of locations per day.

These counter examples provide strong evidence that we are not talking about new sophisticated algorithms in effect,

Figure 1. Percentage of times a site appeared in a particular position in the top-10 search results.

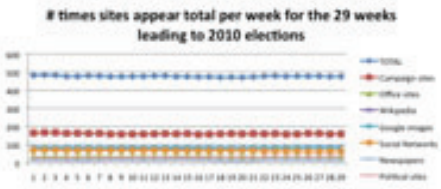


Figure 2. Relative change in position of collections of sites during the 29 week period preceding the 2010 congressional elections.

but for old-fashion, hand-crafted list of blacklisted sites and white-listed terms.

ACKNOWLEDGMENT

The author would like to thank Eni Mustafaraj, Era Vuksani, Ljubica Ristovska and Dana Bullister for collecting some of the date presented in this paper. This work was partially supported by a Brachman-Hoffman Fellowship.

REFERENCES

[1] Pew Foundation, "Pew internet and american life project." http://www.pewinternet.org, 2008.

[2] P. Metaxas, "Web spam, social propaganda and the evolution of search engine rankings," *Lecture Notes in Bus. Info. Proc. (LNBIP)*, accepted; to appear 2010.

[3] M. Hindman, K. Tsioutsiouliklis, and J. Johnson, "Googlearchy: How a few heavily-linked sites dominate politics on the web," in *Annual Meeting of the Midwest Political Science Association*, April 3-6 2003.

[4] C. Castillo and B. D. Davison, "Adversarial web search," *Foundations and Trends in Information Retrieval*, vol. 4, pp. 377–486, June 2010.

[5] T. McNichol, "Your message here." New York Times, Jan. 22 2004.

[6] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: the underground on 140 characters or less," in *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, (New York, NY, USA), pp. 27–37, ACM, 2010.

[7] srainwater, "Nigritude ultramarine faq." http://www.nigritudeultramarines.com/, 2004.

[8] T. Zeller Jr., "Gaming the search engine, in a political season." New York Times, Nov. 6 2006.

[9] D. Sullivan, "Obama Is "Failure" At Google & "Miserable Failure" At Yahoo." http://searchengineland.com/yahoo-obama-is-a-miserable-failure-16286, January 22., 2009.

[10] T. McNichol, "Engineering google results to make a point." New York Times, Jan. 22 2004.

[11] S. Hansell, "Google keeps tweaking its search engine." New York Times, Jun. 3 2007.

[12] P. Metaxas and E. Mustafaraj, "The battle for the 2008 us congressional elections on the web," in *Proceedings of the Web Science 2009 Conference*, (Athens, Greece), March 2009.

[13] C. Bowers, "Call to action: Google-bombing the election." Daily Kos, http://www.dailykos.com/story/2006/10/22/133437/99, Last retrieved on Nov. 23, 2010.

[14] S. L. Stirland, "Google is latest weapon vs. gop." Politico, http://www.politico.com/news/stories/1010/43767.html, Last retrieved on Nov. 23, 2010.

[15] D. Segal, "A bully finds a pulpit on the web." New York Times, Nov. 20 2010.

[16] D. Segal, "The dirty little secrets of search." New York Times, Feb. 12 2011.

[17] D. Welch, "Power of persuasion - propaganda," *History Today*, vol. 49, no. 8, pp. 24–26, 1999.

[18] A. M. Lee and E. B. Lee(eds.), *The Fine Art of Propaganda*. The Institute for Propaganda Analysis. Harcourt, Brace and Co., 1939.

III. Appendix: On Propaganda Theory

We offer here a brief introduction to the theory of propaganda detection. For more information, see [2].

There are many definitions of propaganda, reflecting its multiple uses over time. One working definition we will use here is

Propaganda is the attempt to modify human behavior, and thus influence people's actions in ways beneficial to propagandists.

Propaganda has a long history in modern society and is often associated with negative connotation. This was not always the case, however. The term was first used in 1622, in the establishment by the Catholic Church of a permanent Sacred Congregation *de Propaganda Fide* (for the propagaton of faith), a department which was trying to spread Catholicism in non-Catholic Countries [17]. Its current meaning comes from the successful Enemy Propaganda Department in the British Ministry of Information during WWI. However, it was not until 1938, in the beginning of WWII, that a theory was developed to detect propagandistic techniques. For the purposes of this paper we are interested in ways of detecting propaganda, especially by automatic means.

First developed by the Institute for Propaganda Analysis [18], classic Propaganda Theory identifies several techniques that propagandists often employ in order to manipulate perception.

- **Name Calling** is the practice of giving an idea a bad label. It is used to make people reject and condemn the idea without examining the evidence. For example, using the term "miserable failure" to refer to political leaders such as US President George Bush can be thought of as an application of name calling.
- **Glittering Generalities** is the mirror image[1] of name calling: Associating an idea with a "virtue word", in an effort to make us accept and approve the idea without examining the evidence. For example, using the term "patriotic" to refer to illegal actions is a common application of this technique.
- **Transfer** is the technique by which the propagandist carries over the authority, sanction, and prestige of something respected and revered to something he would have us accept. For example, delivering a political speech in a mosque or a church, or ending a political gathering with a prayer have the effect of transfer.
- **Testimonial** is the technique of having some respected person comment on the quality of an issue on which they have no qualifications to comment. For example, a famous actor who plays a medical doctor on a popular TV show tells the viewers that she only uses a particular pain relief medicine. The implicit message is that if a famous personality trusts the medicine, we should too.
- **Plain Folks** is a technique by which speakers attempt to convince their audience that they, and their ideas, are "of the people," the "plain folks". For example, politicians sometimes are seen flipping burgers at a neighborhood diner.
- **Card Stacking** involves the selection of facts (or falsehoods), illustrations (or distractions), and logical (or illogical) statements in order to give an incorrect impression. For example, some activists refer to the Evolution Theory as a theory teaching that humans came from apes (and not that both apes and humans have evolved from a common ancestor who was neither human nor ape).
- **Bandwagon** is the technique with which the propagandist attempts to convince us that all members of a group we belong to accept his ideas and so we should "jump on the band wagon". Often, fear is used to reinforce the message. For example, commercials might show shoppers running to line up in front of a store before it is open.

The reader should not have much trouble identifying additional examples of such techniques used in politics or advertising.

[1] Name calling and glittering generalities are sometimes referred to as "word games."

Social Relevance

Enkh-Amgalan Baatarjav and Ram Dantu
Department of Computer Science and Engineering
University of North Texas
Denton, Texas, 76203, USA
Email: {eb0050, rdantu}@unt.edu

I. Introduction

One of the first online social networking (OSN) sites was SixDegrees. Since its launch in 1997, there have been a wide range of OSN sites targeting different interests[1]. We can find OSN sites for business connection, dating, photo sharing, video sharing, music broadcasting, microblogging, mainstream network, bookmarking, etc. These are just some of many OSN sites that build their themes based on their users' social network. Today, two of the most popular OSN sites are Facebook and Twitter; Facebook has more than 500 million active users and Twitter has more than 105 million registered users[2].

OSN sites have revolutionize way we communicate and share information. For instance, 5,000 status updates[1] were posted on Twitter in 2007. However, the average number of status updates per day increased dramatically during the next few years. The 300,000 status updates in 2008 rose to 2.5 million the following year, and after that 35 million at the end of 2009. But by just the beginning of 2010, there were 50 million status updates. In other words, there were 600 status updates being posted every second. This number can only increase as more and more people adopt this communication medium. 96% of the Millennial generation, an amount estimated to be 80 million people, have already joined a social networking site[3]. A study conducted between October 2008 and February 2009 by Inside Facebook[4] shows that the fastest growing demographic on Facebook is women who are 55 years of age, with the growth rate being 175.3%. This is an indication that OSN sites are being embraced by many different age groups as a way to communicate.

Even though online social networking sites have been thoroughly integrated into our everyday lives, some people are still sceptical about sharing their personal information online because of insufficient privacy and security features. The main goal of our research is to create a privacy management system that only allows for a specified set of a user's friends/followers to have access to the user's status updates. In this paper, we explore how to find a socially relevant set of a user's followers. We make the assumption that social relevance can be defined by the number of messages that are exchanged between two different parties: the more messages that are exchanged, the closer the social relevance between the two parties is. However, the question that remains is what other attributes can help to define social relevance? Here we will analyze two attributes activity matching and vocabulary usage similarity matching between users and their followers.

[1]On Twitter network, status updates are called tweets.

II. Methodology

To find attributes that can define social relevance, we analyze activity and vocabulary usage between users and their followers. First, we rank followers based on social relevance, which is defined by the number of exchanged messages. Second, we find activity patterns for users and their followers, ranking the followers from the ones who have the most similar activity pattern with the users who have the most dissimilar activity pattern. For vocabulary usage, we also rank the followers based on vocabulary similarity to their users' vocabulary of words. Finally, we compare the result of the social relevance ranking with the activity pattern ranking and vocabulary usage similarity.

A. Activity Pattern Analysis

Activity patterns are analyzed in three time domains: hour, week, and month. Hour, week, and month domains are divided into 24, 7, and 12 sub time intervals. On each time interval, the average number of status updates is calculated to represent the activity pattern.

For the next step, we interpolate each activity pattern and give a score to each follower based on how closely matched the follower's activity pattern is with their user's activity pattern. Finally, we rank the followers based on the score and compare the ranking with the social relevance ranking. Detailed study on this work is in [5].

B. Vocabulary Usage Simiarity

We use the cosine coefficient approach to analyze vocabulary usage similarity (VUS) between users and their followers. Some advantages of the cosine coefficient approach are that it eliminates negative vector terms and handles non-binary values.

To apply the cosine coefficient, we first create the vector space of common vocabulary of words used between a user and each one his/her followers. Next, each term words in the vector space (1) is multiplied by importance of each term word:

$$T_i = w_1 * W_1 + w_2 * W_2 + ... + w_n * W_n \qquad (1)$$

Based on Zipf law of vocabulary frequency distribution, the importance of a term is defined by the frequency of a word in the vocabulary; if a word used a high frequency, the importance of the word is small. The importance of term is calculated by transforming normalized word frequency:

$$w_i = \exp^{-\frac{v_i}{\sum_{j=1}^{n} v_j}} \quad (2)$$

, where v_i is vocabulary frequency.

Same as activity pattern analysis, we rank all followers of a user based on VUS and compare the results with the social relevance ranking. The result is shown in the next section.

III. Discussion of Results

To evaluate activity pattern analysis and vocabulary usage similarities, we compare both of the result rankings with social relevance ranking:

$$RD(U_i) = \mid SRR(U_i) - ResultR(U_i) \mid \quad (3)$$

, use a histogram of rank differences (RD) between users positions on the social relevance ranking (SRR) and users' positions on either activity pattern ranking (APR) or vocabulary similarity ranking (VSR). If a follower is ranked similarly on both SRR and APR, their differences should be fairly low. In Eq. 3, taking the absolute value of the result is optional.

In this study we use two set of data: one set consists of 3,652,148 status updates and 2,312 users and followers. The second data set is provided by the Microsoft Research Project[2].

A. Activity Pattern Ranking

The rank difference histogram of SRR and APR is shown on Fig. 1. It shows that a significant portion of the followers are ranked similarly on both rankings (the difference is close to zero), suggesting that the activity pattern analysis could provide a useful attribute to define social relevance between users and their followers.

B. Vocabulary Similarity Ranking

The cumulative result of rank difference between SRR and VSR is shown in Fig. 2. From the result, we can observe that there is a pattern of increasing frequency as we move close to zero on X-axis. It indicates that there is a positive correlation between common vocabulary usage and social relevance.

IV. Conclusion and Future Work

Online social networking has become popular among all generations. It is used for broadcasting and sharing information. Even though it is an efficient method of communication, there are still some privacy issues involving who on the social graph is able to view what information. Existing social networking sites are lacking in the privacy control needed to manage this. One approach is to find socially relevant followers with whom users want to share their information. In this preliminary study, we delve into non-obvious ways to define social relevance, and we discover that the activity pattern and vocabulary similarities can be good representatives to define social relevance between users. The result of both attribute studies shows that there are positive correlations. In our future work, we will explorer more ways to find social relevance and combine different attributes to reduce RD (Eq. 3).

[2]http://research.microsoft.com/downloads

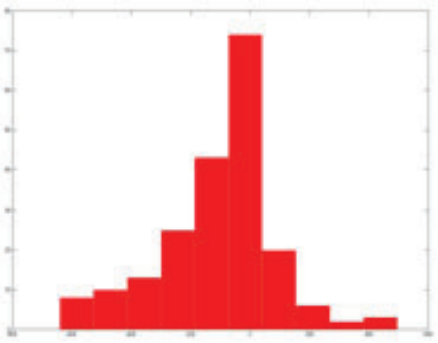


Fig. 1. This histogram demonstrates the difference between the APR and the SRR. A difference of 0 means that both rankings are the same for a particular follower, while a large difference signifies very different rankings. The X-axis represents the difference measurement, while the Y-axis represents the number of users with that difference measurement. It is evident that ranking methods correlate fairly well, with a large portion of users retaining a small difference measure.

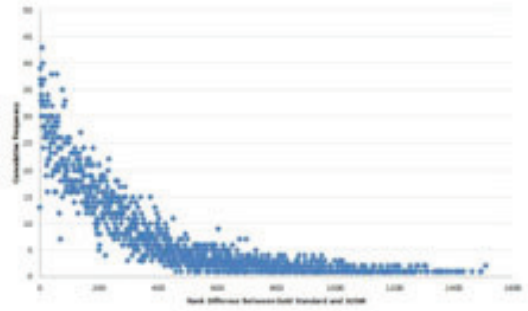


Fig. 2. Cumulative result of ten users who have the highest number of conversations in the corpus. X-axis represents the difference between RSS and VSR, and Y-axis represents frequency of the difference.

References

[1] D. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1-2, November 2007. [Online]. Available: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

[2] N. Bilton, "Chirp, twitters first developer conference, opens its doors," *The New York Times*, April 2010, [Online; accessed 26-April-2010].

[3] E. Qualman, *Socialnomics: How Social Media Transforms the Way We Live and Do Business*. Hoboken, New Jersey: John Wiley and Sons, Inc, 2009.

[4] J. Smith, "Fastest growing demographic on facebook: Women over 55," *Inside Facebook*, 2009.

[5] E.-A. Baatarjav, A. Amin, R. Dantu, and N. K. Gupta, "Are you my friend?" in *Proceedings of the 7th IEEE conference on Consumer communications and networking conference*, ser. CCNC'10. Piscataway, NJ, USA: IEEE Press, 2010, pp. 554–558. [Online]. Available: http://portal.acm.org/citation.cfm?id=1834217.1834342

Identification of Social Network Actors: A Fuzzy based Context-Dependent Approach

Mohamed Fazeen[†1], Ram Dantu[†2], and Parthasarathy Guturu[‡3]
[1]mohamedfazeen@my.unt.edu, [2]rdantu@unt.edu, [3]guturu@unt.edu
[†]*Department of Computer Science & Engineering, College of Engineering, University of North Texas.*
[‡]*Electrical Engineering, College of Engineering, University of North Texas.*

***Abstract*—In this paper, we present a method to classify different social network actors such as leaders (e.g., news groups), lurkers, spammers and close associates with a context-dependent approach in Twitter domain. This method is a two stage process with a fuzzy-set theoretic (FST) approach to evaluate the strengths of network links (actor-actor relationships) followed by a simple linear classifier to separate the actor classes. Since the method uses mostly contextual information such as actor profiles *etc.*, it may be termed as a context-dependent approach. The research was conducted on a Twitter database of 441234 actors, 2045804 links, 6481900 tweets, and 2312927 total reply messages. This context-dependent analysis reveals strong clustering of actor data based on their types, and hence can be considered as a superior approach when data available for training the system is abundant.**

I. Introduction

Recent years have witnessed a proliferation of social networks with popular applications such as Twitter, Flickr, YouTube, LiveJournal, Orkut, and Facebook. With this rapid pace of growth in social networks (SN), there has also been a growing interest in the Internet research community in the SN analysis to address various aspects of social networking issues.

As a step towards addressing the problem of SN privacy, we present the classification algorithms for the problem of identification of types of actors (individuals or organization) in a twitter network. We categorize the twitter actors into 4 types: a) Leaders, who start tweeting, but do not follow any one there after, though they could have many followers, b) Lurkers, who are generally inactive, but occasionally follow some tweets, c) spammers, the unwanted tweeters, also called as twammers, and d) close associates, including friends, family members, relatives, *etc.* A context-dependent classification approach proposed here to addresses this problem for situations where an abundant amount of tweet data is available; here we employ a fuzzy classification scheme in contrast to the stochastic estimation methods [1].

We present in Section II our approach to context-dependent classification of twitter actors. In Section III, details of experimentation and results are presented. Finally, summary and conclusions are presented in Section IV.

II. Context-Dependent Classification

Social networks are formed by social groups of people that are linked by social bond or relationship. In a group, one can follow another or one can be followed by the other. In the twitter jargon, those two types of individuals are called followers and followees, respectively. Unlike in the case of emails, the mutual relationships between individuals/groups in the twitter network can be tracked by monitoring the tweets. Even though it is possible to compile email message and response pairs, the sparseness of data there makes it difficult to estimate the strength of relationship between the corresponding individuals. In the twitter (generally, SN) domain, it is possible to estimate the SN link (relationship) strengths by using the followee-follower message statistics.

Since most of the tweets occur between close associates, the twitter data for this group is generally overabundant compared to the other three groups, and this data imbalance poses problems for an identification of the actors, particularly those belonging to sparsely populated classes. Hence, we follow a two stage process. In the first stage, we estimate the strength of links in the social network and eliminate a large number of actors with strong social bond as they naturally classify into the group of close associates. Then, in the second stage, we perform a linear classification of the four actor types mentioned in Section I, using number of tweets and the followee-follower ratio as two features considering only those tweeters with weak link strength (less than 15% of maximum strength) between them.

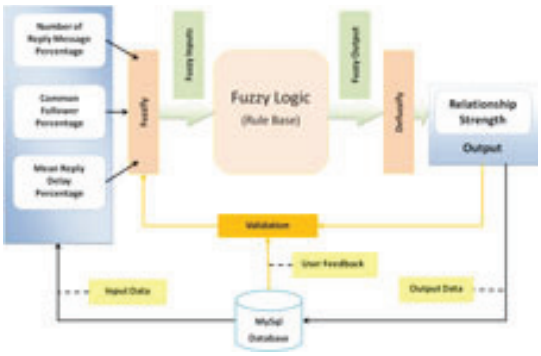


Fig. 1. Proposed Fuzzy System Architecture for Link (Relationship) Strength Evaluation

In the first stage, we focus exclusively on the fuzzy methodology, and do the implementation using jFuzzyLogic [2], an open source code in the Java language for the fuzzy control language (FCL) defined by the International Elec-

trotechnical commission (IEC)'s standard 1131-7 [3]. Three parameters "Reply Message Percentage", "Common Follower Percentage", and "Normalized Mean Reply Delay" have been considered as indicators of the keenness with which a tweeter is followed, and hence used to constitute the input set of our system depicted in Fig. 1.

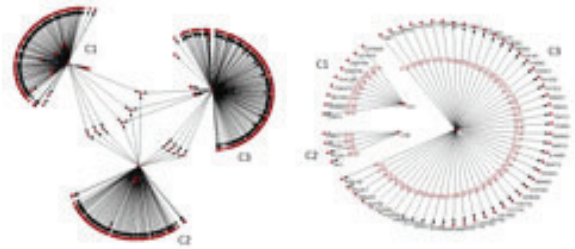
TABLE I
CUMULATIVE TWITTER DATABASE STATISTICS

Actors	Links	Tweets Messages	Tweet Replies
441234	2045804	6481900	2312927

III. EXPERIMENTAL RESULTS

A. Data Collection Procedure

The main requirement for this research is availability of a good data set that includes details of all the activities in the twitter network. Therefore an efficient crawling program was written to collect data from the live twitter network. Table I summarizes the statistics of this database. However, since it is difficult to visualize such a huge network, we show only the results of 500-node subnet in subsection III-B.



(a) Connectivity of a 500-node twitter network with all Links shown.

(b) Connectivity of the twitter network with only links having relationship strength above 15%.

Fig. 2. Changes in twitter network clustering with thresholding of links

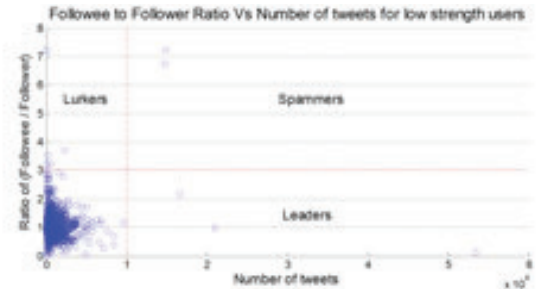


Fig. 3. Linear Separation of Leaders, Lurkers, Associates, and Spammers

B. Results on Link Strength Determination and Context-Dependent Classification

We determine the link strengths of a 500-node twitter network by applying the fuzzy logic based classification method discussed in Section II. This sub-network is visualized in Fig. 2. First, it can be observed that the network nodes form strong clusters, and the cluster structures don't change much when weak links are removed; this indicates that the same set of tweeters communicate frequently with one another though some are more involved than the other in tweeting. Next, since it is easy to infer that the tweeters with strong connectivity are close associates, we need to apply our classification to the tweeters with low relationship strength. From the Fig. 2(b), it is clear that a threshold of 15% for the link strength, is good enough to separate out the tweeters who are unambiguously close associates. Hence, we apply the simple linear classification algorithm using number of tweets and the followee-follower ratio as two components of the pattern vector only to the tweeters with the link strength below this threshold. Clear separation of the four tweeter classes as depicted in Fig. 3 suggests that this method holds promise for an effective identification of leaders, lurkers, spammers, and associates. For validation, we hand-labeled these records by going through the profile information and the tweets contents. By considering the hand labels of the records as the ground truth, we tally the results of linear classification on the validation set with the ground truth. Validity of our classification approach has been established by a perfect tally.

IV. SUMMARY AND CONCLUSION

In this paper, we present a classification method for twitter network actor identification. It employs a fuzzy logic approach to estimate inter-actor relationship strengths in the first step and then a linear classifier to separate out the four actor classes.

This research enforces the conventional wisdom that spammers follow a large number of people (followees), but they themselves are followed by very few people. Specifically, as evidenced from the results of Section III-B, spammers are defined by the accounts that make more than 10,000 tweets in a 10-day interval (or equivalently over an average of 1000 tweets a day) and have a followee-to-follower ratio of 1.5 to 1 or more. The twitter leaders, on the other hand, can be distinguished by their high rate of tweeting, large number of followers, but a few, if any, followees, and hence by a followee-to-follower ratio much below 1. Close associated are marked by strong connectivity to their followers, low to moderate number of tweets (1000) per day, and small to moderate (less than 3) followee-follow ratio. Finally, lurkers is a rare class of tweeters, who follow many people, but they themselves rarely post or reply any tweets.

REFERENCES

[1] M. Fazeen, R. Dantu, and P. Guturu, "Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches," *Social Network Analysis and Mining*, pp. 1–14, 2011, 10.1007/s13278-011-0017-9. [Online]. Available: http://dx.doi.org/10.1007/s13278-011-0017-9

[2] P. Cingolani, "jfuzzylogic, open source fuzzy logic library and FCL language implementation," Available: http//jfuzzylogic.sourceforge.net/html/index.html, 2010.

[3] I. E. Commission, "Technical committee no. 6: Industrial process measurement and control. sub-committee 65 b: Devices. iec 1131-programmable controllers," Available: http://www.fuzzytech.com/binaries/ieccd1.pdf, 1997.

Abstract Submission for Interdisciplinary Workshop on Information and Decision in Social Networks
May 31 - June 1, 2011
MIT

Title: Social Media and the 25 January Revolution: Social Firestorm or Tempest in a Teapot?

Author: Kimberly Glasgow

Email Address: kimberly.glasgow@jhuapl.edu

Affiliation: The Johns Hopkins University Applied Physics Laboratory

Abstract:
The potential impact and power of social media to support, sustain, and perhaps ignite massive social protest have taken the world by storm as the dramatic events currently reshaping the Middle East unfold. Yet social media data, such as Twitter data, remain problematic as sources of information about these phenomena. They may contain inaccuracies, and they are vulnerable to deliberate manipulation, for both symbolic 'support" motivations and for more nefarious reasons. Discriminating evidence of the true social processes and networks in play from the surrounding "chaff" is a challenge. This talk describes work to investigate a social media data set (from Twitter), posted ostensibly from Egypt during the recent protests that led to the fall of the Mubarak government. Can we isolate reflections of the social processes or networks involved in information flow, influence, or decision making in times of protest, civil unrest, or crisis? What is the impact of these data issues on our ability to draw conclusions from social media?

Contact :

Kimberly Glasgow
The Johns Hopkins University Applied Physics Laboratory
Applied Information Science Group (AISD)
11100 Johns Hopkins Road
Laurel, MD 20723
(443) 778-7444
kimberly.glasgow@jhuapl.edu

That's What (Best) Friends Are For*

David Liben-Nowell
Department of Computer Science
Carleton College
Northfield, MN
dlibenno@carleton.edu

Over the last decade, the burgeoning networks research community has paid considerable attention to structural properties of social networks. Such properties—triadic closure, heavy-tailed degree distributions, small-world phenomena, and homophily, to name just a few—have proven to be remarkably robust in newly accessible large-scale data on networks. Of course, untangling the causes of these apparently ubiquitous properties from the data is harder. What behavioral mechanisms lead to these structural facts? And what we can infer about individuals' views of friendship from these structural phenomena?

In recent work with two evolutionary psychologists (Peter DeScioli and Robert Kurzban) and two computer scientists (Elizabeth Koch and me), we have been addressing questions about how people choose friends and prioritize among those friends. (The deepest question stemming from the evolutionary psychology perspective is, simply: why do people have friends at all?) For our analysis, we collected a large sample of about 10 million profiles from the MySpace social network. Most MySpace users have a *Top Friends* module in their profiles, in which an individual selects a small subset (usually eight) of his or her friends and organizes them into a ranked order of the individual's choice. These rankings allow the exploration of finer-grained distinctions among social relationships: not just "is Alice a friend of Charlie?" but instead "does Charlie rank Alice higher or lower than he ranks Bob?"

Different classes of behavioral hypotheses give rise to very different graph-theoretic structures in the *best-friend network*. We discover that the MySpace best-friend network is most consistent with the class of hypotheses in which an individual's relative ranking of two candidate friends is tied to some measure of the individual's dyadic relationship with those two candidates, rather than some measure of the monadic quality of the candidates themselves. For example, an individual tends to prefer the candidate friend who is geographically closest to the individual. But an individual tends to prefer the candidate who is globally *less* popular than another candidate, suggesting that qualitatively different dynamics than preferential attachment are at play in best-friend networks. Our best predictor of whether Alice or Bob will be named by Charlie as his best friend is how highly Alice and Bob rank Charlie. This observation provides support for *alliance hypotheses* ("we have friends so that they will take our side in a potential conflict") about the ultimate explanation for why we have friends at all.

*Supported in part by NSF grant CCF-0728779. Nearly all of this research was performed jointly with Peter DeScioli, Elizabeth N. Koch, and Robert Kurzban; those results are reported in "Best Friends: Alliances, Friend Ranking, and the MySpace Social Network," *Perspectives on Psychological Science* 6(1):6–8, January 2011.

Fertility decisions and their sensitivity to social networks and family policies

Thomas Fent, Vienna Institute of Demography, Austrian Academy of Sciences, email: thomas.fent@oeaw.ac.at

Belinda Aparicio Diaz, Vienna Institute of Demography, Austrian Academy of Sciences, email: belinda.aparicio.diaz@oeaw.ac.at

Alexia Prskawetz, Institute of Mathematical Methods in Economics, Vienna University of Technology, email: afp@econ.tuwien.ac.at

We investigate the effectivity of family policies in the context of the structure of a society. We use an agent based model to analyse the impact of policies on individual fertility decisions and on fertility at the aggregate level. Our results indicate that both fixed and income dependent child support have a positive and significant impact on fertility. In addition, the specific characteristics of the social network and social influence will not only relate to fertility but also influence the effectivity of family policies. Policymakers aiming to adapt a specific policy mix that has proved successful from one country to another one ignoring differences in the social structure may fail. Family policies can only be successful if they consider the characteristics of the society they are assigned for.

We aim to resolve the confusion and disagreement about the effectivity of family policies by explicitly adressing their direct and indirect effect on fertility. The direct effect eases the load of having children for instance by providing institutional childcare or financial benefits. The indirect effect rests on the assumption that many people imitate or consult their friends, siblings, or parents in choosing their intended fertility. Policies causing a modest effect on fertility at the individual level may have a large impact at the macro level due to such peer effects.

The crucial features of our model are the agents' heterogeneity with respect to age, income, parity, and intended fertility, the social network which links the agents to a small subset of the population and the influence mechanism acting via that network. We use data from the Gender and Generation Survey (GGS) to estimate the distribution of the desire for children given the agents age and parity. We define the probability π_i^m that agent i wants at least m additional children and use the logit model

$$\text{logit}(\pi_i^m) = \beta_0^m + \beta_1^m x_i + \beta_2^m p_i \tag{1}$$

for each m to estimate the according probabilities for the agent population.

The agents own consumption, $c_{i,t}$, is assumed to be a concave function of household income, $c_{i,t} = \sigma\sqrt{w_{i,t}}$, the consumption level of $n_{i,t}$ children is$c_{i,t}^{(n_{i,t})} = n_{i,t}\,\tau\sqrt{w_{i,t}}$. and the disposable income becomes $y_{i,t} = w_{i,t} - c_{i,t} - c_{i,t}^{(n_{i,t})}$. If intended fertility exceeds actual parity,

$$f_{i,t} > p_{i,t}, \tag{2}$$

and the disposable income is equal or greater than the costs of an additional child,

$$y_{i,t} \geq \tau\sqrt{w_{i,t}} \iff \sqrt{w_{i,t}} \geq \sigma + (n_{i,t}+1)\tau, \tag{3}$$

the agent is exposed to the biological probability (fecundity) of conception. In case of a live birth a new agent is generated. After the child's transition to adulthood the new adult agent gets assigned her income level z_i determining her household income $w_{i,t} = w_{i,t}(z_i, x_{i,t})$, her social

network, and her fertility intentions. Thereafter she evaluates her fertility intentions with respect to (2) and (3).

The policy maker may provide a mix of fixed family allowances, b^f, and benefits proportional to income, $b^v w_{i,t}$. Then, the necessary condition for an additional child becomes

$$\sqrt{w_{i,t}} \geq \sigma + (n_{i,t}+1)\left(\tau - \frac{b^f}{\sqrt{w_{i,t}}} - b^v\sqrt{w_{i,t}}\right).$$

The agents are linked to a set of agents with whom they communicate about their fertility intentions and realisations. We refer to this group as an agent's social network or peer group. The similarity of agents' characteristics has an impact on the probability of being chosen into an agents social network. Moreover, we assume a certain degree of network transitivity or clustering, i.e. the tendency that two agents who are connected to a common third party establish a mutual relationship over time. Each agent i has an intended fertility $f_{i,t}$, defined as the sum of current parity $p_{i,t}$ and the intended additional children which may be altered due to social influence imposed by the peer group. We assume that with probability pr_3 (pr_4) intended fertility increases (decreases) due to the influence exerted by a peer with parity greater (less) than the agents intended fertility. Then, we compute π_i^+ (π_i^-), the number of agents j who are linked to i with parity greater (less) than the intended fertility of agent i, i.e. $p_{j,t} > f_{i,t}$ ($p_{j,t} < f_{i,t}$) to obtain the probabilities to be positively or negatively influenced by at least one agent from the peer group, $p_{i,t}^+ = 1-(1-pr_3)^{\pi_i^+}$ and $p_{i,t}^- = 1-(1-pr_4)^{\pi_i^-}$. The probability of being only positively (negatively) influenced becomes $(1-p_{i,t}^-)p_{i,t}^+$ (respectively $(1-p_{i,t}^+)p_{i,t}^-$) and the probabiliy of being positively and negatively influenced is $p_{i,t}^+ p_{i,t}^-$. Then, the probabilities to increase, decrease, or keep the intended fertility constant are

$$\begin{aligned}
p_i(f_{i,t+1} = f_{i,t}+1) &= (1-p_{i,t}^-)p_{i,t}^+ + \gamma p_{i,t}^+ p_{i,t}^- \\
p_i(f_{i,t+1} = f_{i,t}-1) &= (1-p_{i,t}^+)p_{i,t}^- + (1-\gamma)p_{i,t}^+ p_{i,t}^- \\
p_i(f_{i,t+1} = f_{i,t}) &= (1-p_{i,t}^+)(1-p_{i,t}^-).
\end{aligned}$$

These adaptations of fertility intentions capture the indirect effect of family policies.

We run the simulation model for 100 time steps (years) with a fixed set of parameters and record completed cohort fertility, intended fertility, and the fertility gap (the difference between intended and completed cohort fertility) on the aggregate level. Afterwards we generate another initial population and run the simulation again with a different set of parameters. We study the impact of fixed and income dependent family allowances on intended fertility, on the realisation of intended fertility and on the resulting completed cohort fertility. In particular we investigate whether the structure of a society represented by parameters specifying the social network and the social influence mechanism has the potential to alter the role of family policies.

Our simulations reveal a positive impact of both fixed and income dependent family allowances on completed cohort fertiliy and on intended fertility and a negative impact of fixed and income dependent child support on the fertility gap. However, several network and social influence parameters have the ability not only to influence fertility itself but also to influence the effectivity of family policies, often in a detrimental way. For instance, while a higher degree of homophily among network partners has a positive effect on fertility (intentions and realisations), family policies may be less effective in such a society.

We further conclude that empirical cross-country comparisons of different types of family policies need to be interpreted with caution for two reasons. Firstly, the empirical impact of a certain policy depends on the subset of policies investigated and comprehensive experiments taking into consideration any possible policy mix are not feasible in the real world. Secondly, many empirical studies do not account for differences in the social structure in the countries under consideration.

Adopting longitudinal network analysis to investigate the emergence of shared leadership

Dr Cécile Emery

Department of Management

London School of Economics and Political Science

Houghton Street

London WC2A 2AE

Tel: +44 (0)20 7106 1169

Fax: +44 (0)20 7955 7424

c.emery@lse.ac.uk

How do people choose a leader for their group? Traditional models of leadership have recently been challenged by a growing sentiment that they failed to completely represent and understand the complex, dynamic, distributed, and contextual nature of leadership (McKelvey, 2008; Uhl-Bien & Marion, 2009). The model of shared, or distributed, leadership goes in such direction as it envisions "leadership as an emergent property of a group or network of interacting individuals" (Bennett, Harvey, Wise, & Woods, 2003: 7). Shared leadership assumes that leadership does not reside in a single individual but can be "dispersed among some, many, or maybe all of the members" (Gronn, 2002: 429). It is an emergent team property that results from the distribution of leadership roles and activities across team members, allowing the possibility for multiple leaders to emerge (Carson, Tesluk, & Marrone, 2007; Mehra, Smith, Dixon, & Robertson, 2006). Shared leadership can be represented as a network of leadership perceptions (hereafter "leadership network") in which nodes and arrows symbolize individuals and leadership nominations/perceptions respectively (Carson et al., 2007; Mehra et al., 2006). If assessed over time, a dynamic leadership network captures the emergence of shared leadership within a group or organization.

The objective of this paper is to introduce an innovative methodology capable of performing a dynamic assessment of leadership networks, a none-trivial exercise impossible to realize until recently. Thanks to new advances in social network techniques, it is now possible to conduct a longitudinal analysis on leadership networks by specifying actor-oriented models (Snijders, 2005, 2009). Actor-oriented models were conceptualized to statistically assess how social networks evolve over time (Snijders, 2005, 2009; Snijders, van de Bunt, & Steglich, 2010), and are run using the software SIENA (*Simulation Investigation for Empirical Network Analysis*; Ripley & Snijders, 2010). SIENA models were developed for a variety of longitudinal networks. Empirical studies count investigations on the effects of the Big Five personality traits on longitudinal friendship networks (Selfhout et al., 2010), the evolution of interorganizational networks (van de Bunt & Groenewegen,

2007), dynamic advice networks (Lazega, Mounier, Snijders, & Tubaro, 2010), and the reciprocal effects of self-view as a leader and leadership emergence (Emery, Daniloski, & Hamby, 2011). This methodological and analytic strategy has the potential to rise our understand of the emergence of shared leadership in group as it captures the complex and dynamic process of leader emergence, simultaneously captures different levels of analysis - the individual, the dyad, and the group (Livi, Kenny, Albright, & Pierro, 2008), combines different leadership theories (leader-centered, follower-centered, similarity-hypothesis, and relational leadership) in the same analysis, and statistically tests for the processes hypothesized to impact the evolution of leadership networks. As an illustrative example, this paper provides an exploratory study designed to explore the role of narcissism on the evolution of two types of leadership networks: task versus relationship oriented leadership.

Using Metrics to Enable Large-Scale Deliberation

Mark Klein
MIT Center for Collective Intelligence
m_klein@mit.edu

Humanity now finds itself faced with a range of highly complex and controversial challenges – such as climate change, the spread of disease, international security, scientific collaborations, product development, and so on - that call upon us to bring together large numbers of experts and stakeholders to deliberate collectively on a global scale. Collocated meetings can however be impractically expensive, severely limit the concurrency and thus breadth of interaction, and are prone to serious dysfunctions such as polarization and hidden profiles (Sunstein 2006). Social media such as email, blogs, wikis, chat rooms, and web forums provide unprecedented opportunities for interacting on a massive scale, but have yet to realize their potential for helping people *deliberate* effectively, typically generating poorly-organized, unsystematic and highly redundant contributions of widely varying quality. Large-scale argumentation systems represent a promising approach for addressing these challenges, by virtue of providing a simple systematic structure that radically reduces redundancy and encourages clarity. They do, however, raise an important challenge. How can we ensure that the attention of the deliberation participants is drawn, especially in large complex argument maps, to where it can best serve the goals of the deliberation? How can users, for example, find the issues they can best contribute to, assess whether some intervention is needed, or identify the results that are mature and ready to "harvest"? Can we enable, for large-scale distributed discussions, the ready understanding that participants typically have about the progress and needs of small-scale, collocated discussions?

This paper will address these important questions, discussing (1) the strengths and limitations of current deliberation technologies, (2) how argumentation technology can help address these limitations, and (3) how we can use attention-mediation metrics to enhance the effectiveness of large-scale argumentation-based deliberations.

Sarah M. G. Otner, Sarah.Otner@post.harvard.edu
London School of Economics (**LSE**), UK
"Emerging Expertise: Status and Influence in Electronic Networks of Practice"

Leadership is the exertion of influence in order to "guide, structure, and facilitate relationships in a group," (Yukl, 1998, p. 3). Traditional leadership originates from formal position or designated authority. In contrast, "distributed leadership" derives from a "broader, mutual influence process independent of any formal role or hierarchical structure and [as] diffused among the members of any given social system," (de Rue & Ashford, 2010, p. 627). Such "emergent" leadership is earned from followers through incremental contributions and accumulated influences (cf. Yoo & Alavi, 2004). Given that leadership intrinsically incorporates such patterns of unequal influence, deference, and respect, and that those observable patterns are status structures (Ridgeway & Walker, 1995), then leadership inherently involves status differentiation (Ridgeway, 2004). Mastering this process of distinguishing individuals according to their status characteristics is fundamental to understanding how individuals gain influence and then emerge as leaders in organizations.

In organization studies, the literature concerning communities of practice (CoP; Brown & Duguid, 1991) provides a foundation for understanding distributed leadership and new organizations. Computer-mediated communication has produced hybrid community-networks, or "electronic networks of practice" (ENoP). ENoP are "computer-mediated social spaces where individuals working on similar problems self-organize to help each other and share knowledge, advice, and perspectives about their occupational practice or common interests," (Wasko, Teigland, & Faraj, 2009, p. 254). ENoP can be a significant value driver for an organization; "these types of institutions are especially important for promoting value by creating 'new combinations'," (Moran & Ghoshal, 1996, p. 42). Indeed, the management of ENoP is "a delicate and highly strategic internal capability" (Kane, 2009, p. 41) and a "core competency" (Gilbane Group, 2008). Therefore, a primary motivation is to discover how individuals emerge as leaders in ENoP, and also the consequences of the same.

Previous research (cf. Organisciak, 2009; Brabham, 2008b; Nov, 2007; Lakhani & Wolf, 2005) has demonstrated that direct financial gain is not prominent among reasons that individuals contribute to online communities; however, many of the other drivers – including improvements to personal reputation, achievement of status, and particularly career advancement – facilitate *indirect* financial gains. Considering these motivations reveals a core problem of applying extant literature to this novel context. From a generalized exchange perspective, the community relies on both the signaling and the sanctioning functions of reputation in order to keep individuals contributing. In contrast, from a resource perspective, individuals have no incentive to amass reputation *ad infinitum* if they cannot or do not avail of it. Therefore, it is central to understand how individuals "spend" their reputations to achieve status and, subsequently, to wield influence. Indeed, disentangling these "overlapping but partially autonomous" (de Cremer & Sedikides, 2008, p. 66) constructs is a priority of current network research (see discussion in Wong & Boh, 2010). The present research contributes to the "burgeoning literature on legitimacy, reputation, and status in management research" (Bitektine, 2011, p. 152; Morrison, 2010) through examining status construction and characteristics in a new organizational form, the ENoP.

Sarah M. G. Otner, Sarah.Otner@post.harvard.edu
London School of Economics (**LSE**), UK

Research Questions

- Why, how, and on whom is status conferred in ENoP?
- How do high status individuals gain influence, and of what form?
- What are the consequences of that influence – for individuals, for the network, and for the organization?

Keywords: Online Communities / Networks of Practice, Status, Influence, Emergent Leadership

REFERENCES

Bitektine, A. 2011. Toward a theory of social judgments of organizations: The case of legitimacy, reputation, and status. *Academy of Management Review*, 36(1): 151-179.

Brown, J. S., & Duguid, P. 1991. Organizational learning and communities of practice: Toward a unified view of working, learning, and innovation. *Organization Science*, 2(1): 40-57.

de Cremer, D., & Sedikides, C. 2008. Reputational implications of procedural fairness for personal and relational self-esteem. *Basic and Applied Social Psychology*, 30: 66-75.

de Rue, D. S., & Ashford, S. J. 2010. Who will lead and who will follow? A social process of leadership identity construction in organizations. *Academy of Management Review*, 35(4): 627-647.

Moran, P., & Ghoshal, S. 1996. Value creation by firms. *Academy of Management Best Paper Proceedings*: 41-45.

Morrison, E. 2010. OB in *AMJ*: What is hot and what is not? *Academy of Management Journal*, 53(5): 932-936.

Ridgeway, C. L. 2004. Status characteristics and leadership. In D. van Knippenberg, & M. Hogg (Eds.), *Leadership and Power: Identity Processes in Groups and Organizations*: 65-78. London: Sage.

Ridgeway, C. L., & Walker, H. A. 1995. Status structures. In K. S. Cook, G. A. Fine, & J. S. House (Eds.), *Sociological Perspectives on Social Psychology*: 281-310. Boston: Allyn and Bacon.

Wasko, M., Teigland, R., & Faraj, S. 2009. The provision of online public goods: Examining social structure in an electronic network of practice. *Decision Support Systems*, 47(3): 254-265.

Wong, S.-S., & Boh, W. F. 2010. Leveraging the ties of others to build a reputation for trustworthiness among peers. *Academy of Management Journal*, 53(1): 129-148.

Yoo, Y., & Alavi, M. 2004. Emergent leadership in virtual teams: What do emergent leaders do? *Information and Organization*, 14: 27-58.

Yukl, G. A. 1998. *Leadership in Organizations* (4th ed.). Upper Saddle River, N.J.: Prentice Hall.

TEMPORAL DIMENSIONS OF ORGANIZATIONAL NETWORK STABILITY: AN EXAMPLE IN THE CONTEXT OF PROJECT TEAMS

Eric Quintane – eric.quintane@usi.ch
Institute of Management – The University of Lugano
and School of Behavioral Science – The University of Melbourne

Philippa E. Pattison – pepatt@unimelb.edu.au
School of Behavioral Science – The University of Melbourne

Garry L. Robbins - garrylr@unimelb.edu.au
School of Behavioral Science – The University of Melbourne

Joeri M. Mol - jmol@unimelb.edu.au
Department of Management and Marketing – The University of Melbourne

Organizational network research focuses on the patterns of stable, long-term relationships. This long-term perspective has proven useful in order to examine the unfolding of social processes that are stable over time. Yet, many organizations face environments that require them to adapt and change over very short periods of time. These short-term changes are usually not captured in organizational network analysis, and the literature is silent on how they may relate to long-term network stability or to organizational stability. In this paper, our aim is to broaden our understanding of the temporal dimension of network processes. We argue that stable social processes exist in different time frames, ranging from short to long-term, and are not subsumed into long-term processes. Specifically, we focus on *reciprocity*, *centralization* and *closure*, and the different time frames in which they can be embedded. Using the example of the e-mail communications patterns of two project teams, we examine the temporal dynamics of organizational networks by considering short-term regularities in interaction structures as well as regularities observed over longer periods. We use the recently developed Relational Event Model to examine these network processes over multiple overlapping time frames using continuous data (without the need for cross sectional aggregation). Our results show that team interaction structures exhibit regularities not only in the long-term, but also in the short-term, thus demonstrating that stability is not confined to long-term social interactions. For instance, *closure* has a strong propensity to have regular occurrences in the short term, which reveals a form of coordination that is temporally bound and goes beyond the traditional transitivity argument. We discuss the role of interaction regularities across different timescales in maintaining stability and flexibility in organizations as well as implications for organizational network research.

Key terms: Stability and Flexibility, Organizational networks, Temporal dimensions, Social structure in teams, Sequences of interactions.

Social learning and the dynamic cavity method

Yashodhan Kanoria*, Andrea Montanari† and Omer Tamuz‡

March 8, 2011

Abstract

In many contexts, agents 'learn' behavior from interaction with friends/neighbors on a network. We call this phenomenon 'social learning'. We will focus on models of repeated interaction, with agents 'voting' in a series of rounds on some issue of interest. Votes in the initial round are based on 'private signals', whereas votes in future rounds incorporate knowledge of previous votes cast by friends.

We consider two different models of iterative learning. A very simple model is 'majority dynamics' where agents choose their vote based on the majority of neighbors' votes in the previous round. We analyze this model on regular trees [KM11]. At the other extreme is iterative Bayesian learning: a fully rational model introduced by Gale and Kariv (2003). We introduce new algorithms for this model, challenging a widespread belief that it is computationally intractable [KT11]. A new technique we develop – the *dynamic cavity method*, serves as a key tool for both models.

1 Majority dynamics

A voter sits on each vertex of a regular tree of degree k, and has to decide between two alternative opinions. At each time step, each voter switches to the opinion of the majority of her neighbors. We analyze this majority process when opinions are initialized to independent and identically distributed random variables.

In particular, we bound the threshold value of the initial bias such that the process converges to consensus. In order to prove an upper bound, we characterize the process of a single node in the large k-limit. This approach is inspired by the theory of mean field spin-glass and can potentially be generalized to a wider class of models. We also derive a lower bound that is nontrivial for small, odd values of k.

As part of our analysis, we introduce a new tool, the dynamic cavity method. This method yields an exact characterization of single node 'vote' trajectories for dynamic processes on trees.

2 Efficient Iterative Bayesian learning

We consider a set of agents who are attempting to iteratively learn the 'state of the world' from their neighbors in a social network. Each agent initially receives a noisy observation of the true state of the world. The agents then repeatedly 'vote' and observe the votes of some of their peers, from

*Department of Electrical Engineering, Stanford University. Email: ykanoria@stanford.edu

†Department of Electrical Engineering and Department of Statistics, Stanford University. Email: montanar@stanford.edu

‡Weizmann Institute. Email: omer.tamuz@weizmann.ac.il

which they gain more information. The agents' calculations are Bayesian and aim to myopically maximize the expected utility at each iteration.

This model, introduced by Gale and Kariv (2003)[GK03], is a natural approach to learning on networks. However, it has been criticized, chiefly because the agents' decision rule appears to become computationally intractable as the number of iterations advances. For instance, a dynamic programming approach (part of this work) has running time that is exponentially large in $\min(n, (d-1)^t)$, where n is the number of agents.

We provide a new algorithm to perform the agents' computations on locally tree-like graphs. Our algorithm uses the dynamic cavity method to drastically reduce computational effort. Let d be the maximum degree and t be the iteration number. The computational effort needed per agent is exponential only in $O(td)$ (note that the number of possible information sets of a neighbor at time t is itself exponential in td).

Under appropriate assumptions on the rate of convergence, we deduce that each agent is only required to spend polylogarithmic (in $1/\epsilon$) computational effort to approximately learn the true state of the world with error probability ϵ, on regular trees of degree at least five. We provide numerical and other evidence to justify our assumption on convergence rate.

We extend our results in various directions, including loopy graphs. Our results indicate efficiency of iterative Bayesian social learning in a wide range of situations, contrary to widely held beliefs.

References

[GK03] Douglas Gale and Shachar Kariv, *Bayesian learning in social networks*, Games and Economic Behavior **45** (2003), no. 2, 329–346.

[KM11] Yashodhan Kanoria and Andrea Montanari, *Majority dynamics on trees and the dynamic cavity method*, to appear in Annals of Applied Probability. Preprint at `http://www.stanford.edu/~ykanoria/majority_camerarready12.pdf`, 2011.

[KT11] Yashodhan Kanoria and Omer Tamuz, *Efficient bayesian social learning on trees*, Draft at `http://arxiv.org/abs/1102.1398`, Feb. 2011.