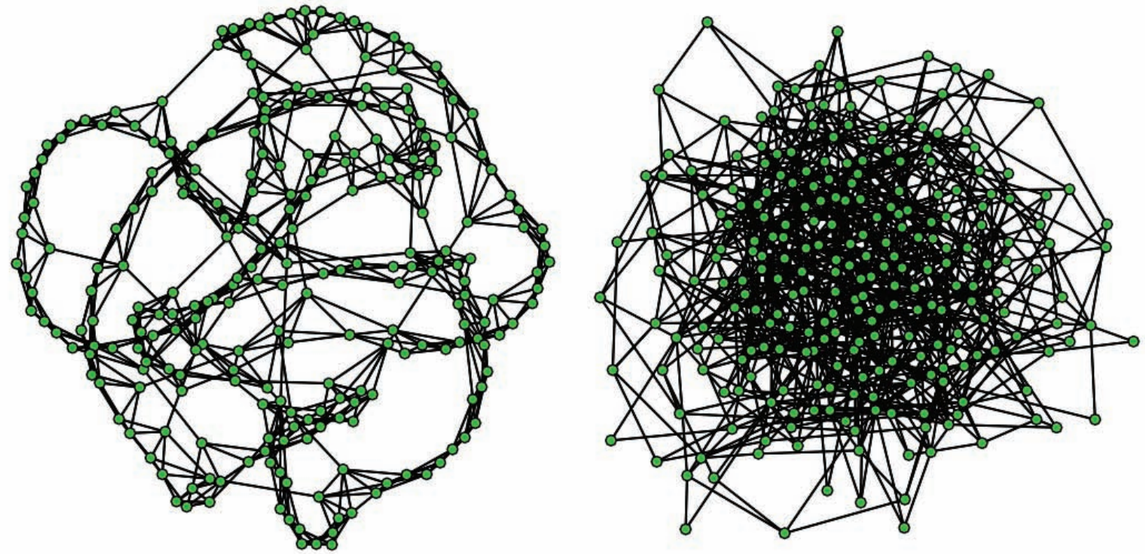
July, 2011

CONNECTIONS

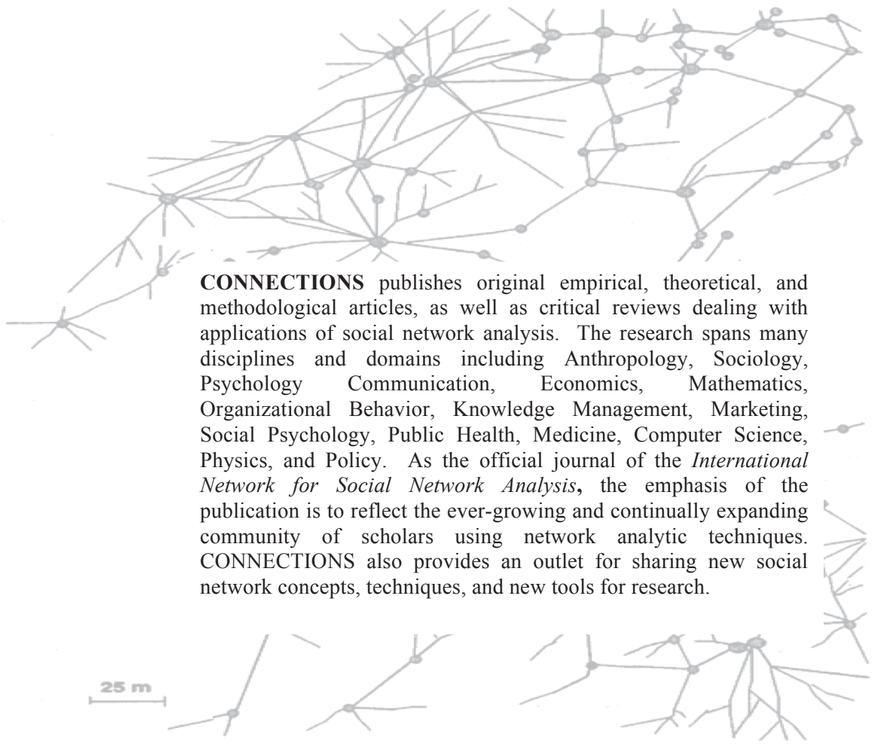
Volume 31 • Issue 1

CONNECTIONS

July 2011 *Volume 31 • Issue 1*



Official Journal of the International Network for Social Network Analysis



CONNECTIONS publishes original empirical, theoretical, and methodological articles, as well as critical reviews dealing with applications of social network analysis. The research spans many disciplines and domains including Anthropology, Sociology, Psychology Communication, Economics, Mathematics, Organizational Behavior, Knowledge Management, Marketing, Social Psychology, Public Health, Medicine, Computer Science, Physics, and Policy. As the official journal of the *International Network for Social Network Analysis***,** the emphasis of the publication is to reflect the ever-growing and continually expanding community of scholars using network analytic techniques. CONNECTIONS also provides an outlet for sharing new social network concepts, techniques, and new tools for research.

Front Cover: *The two images are from the enclosed article "*Linking network structure and diffusion through stochastic dominance" by PJ Lamberson. *The two networks represent the same number of nodes and edges, but different degree distributions. The degree distribution of the network on the right is a mean preserving spread of the degree distribution of the network on the left. In other words, there is greater variation in the number of connections that nodes have in the network on the right. For most contagion processes, such as the spread of a disease or a rumor, we expect the "infection" to spread more easily in the network on the right.*

International Network for Social Network Analysis

CONNECTIONS is the official journal of the **International Network for Social Network Analysis** (INSNA). INSNA is a scientific organization made up of scholars across the world. Updated information about the INSNA's annual conference (**Sunbelt Social Network Conferences)** can be found on the website at *www.insna.org*.

INSNA Board Members

President:	John Skvoretz
Vice President:	Katherine Faust
Treasurer:	Thomas Valente
Founder:	Barry Wellman
Members:	George Barnett, Urik Brandes, Carter Butts, Mario Diani, Laura Koehly, David Lazer, Phillipa Pattison, Garry Robins, Marijtje Van Duijn, and Barry Wellman

INSNA Committees

Finance Committee - Chaired by Treasurer Tom Valente

Program/Conference Committee – Co-chaired by Sunbelt hosts Rebecca Davis, Laura Koehly, and Thomas Valente

Web Committee - Chaired by webmaster and Chief Information Officer Benjamin Elbirt

Publications Committee - Composed of current and former editors of *Social Networks*, *CONNECTIONS* and *Journal of Social Structure* (JOSS) to oversee the INSNA's relations with the publications, selection of *CONNECTIONS* 'and *JOSS's* future editors and to coordinate the publications so that they are complimentary rather than in competition with one another. To insure openness to new ideas, one or more additional members will be selected by the President.

INSNA Awards 2011

- Student Award:

  The 2011 best student paper award was given to Lea Ellwardt, Christian Steglich, and Rafael Wittek for their paper titled "The Co-evolution of Gossip and Friendship at Work: Studying the Dynamics of Multiplex Social Networks."

- Simmel Award:

  The 2011 winner is Kathleen Carley who accepted the award as the keynote speaker at the 2011 Sunbelt Social Networks Conference XXXI in St. Pete Beach, Florida, USA.

- William D. Richards, Jr. Software Award (biennial):

  The winners of the INSNA's 2011 William D. Richards Software award are Steven P. Borgatti, Martin G. Everett and Linton C. Freeman for UCINET: Software for Social Network Analysis.

- INSNA i2 Citation Award:

  The 2011 winner of the INSNA citation award sponsored by i2 was awarded for the paper "An Introduction to Exponential Random Graph (p*) Models for Social Networks," by Garry Robins (University of Melbourne), Pip Pattison (University of Melbourne), Yuval Kalish (University of Melbourne), and Dean Lusher (University of Melbourne) appearing in Social Networks 29(2): 173-191, 2007.

CONNECTIONS

Manuscripts selected for publication are done so based on a peer-review process. See instructions to authors for information on manuscript submission. The journal is edited and published by the Connections Editorial Group:

Editorial Headquarters
University of Southern California, Institute of Prevention Research
1000 Fremont Ave., Unit #8, Building A, Room 5133, Alhambra, CA 91803
Tel: (626) 457-4139; fax: (626) 457-6699

Email *dimitriscc@gmail.com* or *kcoronges@gmail.com* for questions or change in address.

International Network for Social Network Analysis
Hardcopy circulation of Connections is sent to all members of INSNA, the International Network for Social Network Analysis, which has over 1300 members. Subscription to CONNECTIONS can be obtained by registering for INSNA membership through the website*: www.insna.org*. Standard annual membership fee is US$60 ($40 for students). Wherever possible, items referenced in articles (such as data and software) are made available electronically through the INSNA website. In addition, the website provides access to a directory of members' email addresses, network datasets, software programs, and other items that lend themselves to electronic storage.

Sunbelt Social Network Conferences
Annual conferences for INSNA members take place in the United States for two years and in Europe every third year. The Sunbelt Conferences bring researchers together from all over the world to share current theoretical, empirical and methodological findings around social networks. Information on the annual Sunbelt Social Network Conferences can also be found on the INSNA website. Sunbelt XXXII will be held in Redondo Beach, California in March 12-18, 2012.

Instructions to Authors

CONNECTIONS publishes original empirical, theoretical, tutorial, and methodological articles that use social network analysis. The journal publishes significant work from any domain that is relevant to social network applications and methods. Commentaries or short papers in response to previous articles published in the journal are considered for publication. Review articles that critically synthesize a body of published research are also considered, but normally are included by invitation only. Authors who wish to submit a commentary, book review, network image or review article are welcome to do so.

Submitting Manuscripts

Authors are required to submit manuscripts online to the Editor, Dimitris Christopoulos at *dimitriscc@gmail.com*. Expect a notice of receipt of your manuscript via email within one week. Feedback from the editor and reviewers will be sent to the corresponding author within six months after receipt. Revised or resubmitted manuscripts should include a detailed explanation of how the author has dealt with each of the reviewer's and Editor's comments. For questions or concerns about the submission process, authors should contact the editor.

Manuscripts must be in MS Word format and should not exceed 40 pages including tables, figures and references. Manuscripts should be arranged in the following order: title page, abstract, corresponding author contact information, acknowledgments, text, references, and appendices. Abstracts should be limited to 250 words. Please embed all images, tables and figures in the document. Format and style of manuscript and references should conform to the conventions specified in the latest edition of Publication Manual of the *American Psychological Association*. Please remove all embedded formatting and disable any automatic formatting when possible. A figure and its legend should be sufficiently informative that the results can be understood without reference to the text. Please submit two copies of your article, one of which should is anonymised with all reference to the original author/s removed. The journal follows a double-blind review process for research articles. Every issue, we select an image from an accepted article to appear on the front cover of the journal.

CONNECTIONS

July, 2011 **Volume 31 ● Issue 1**

ARTICLES

Linking Network Structure and Diffusion Through Stochastic Dominance

PJ Lamberson
Sloan School of Management,
MIT, Cambridge, Massachusetts, USA

Abstract
Recent research identifies stochastic dominance as critical for understanding the relationship between network structure and diffusion. This paper introduces the concept of stochastic dominance, explains the theory linking stochastic dominance and diffusion, and applies this theory to a number of diffusion studies in the literature. The paper illustrates how the theory connects observations from different disciplines, and details when and how those observations can be generalized to broader classes of networks.

Correspondence concerning this article should be addressed to PJ Lamberson, MIT Sloan School of Management, E62-441, 77 Massachusetts Ave, Cambridge, MA 02139, pjl@mit.edu.

1. Introduction

Network structure affects the speed and extent to which information, disease, behavior, and innovations spread (Abrahamson & Rosenkompf, 1997; Newman, 2002; Sander, Warren, Sokolov, Simon & Koopman, 2002; Young, 2003; Cowan & Jonard, 2004; Centola, Willer & Macy, 2005; Ohtsuki, Hauert, Lieberman & Nowak, 2006). Often there is an abrupt transition from those networks in which the diffusion process dies out completely to those in which it envelops the network. Our understanding of the relationship between network structure and diffusion is built on observations from different disciplines, which employ different techniques and consider different families of networks leaving an array of similar results with no formal connection.

Broadly speaking, there seems to be a greater tendency towards diffusion in networks that are "more random." Underlying this observation are a range of studies employing the Watts-Strogatz family of networks (Watts & Strogatz, 1998), which interpolate between random and regular networks. However, without a general theory we cannot extrapolate from these studies to conclude that "increasing randomness" promotes diffusion in all networks.

Recently, several scholars have demonstrated how stochastic dominance can be used to order networks according to their proclivity to sustain diffusion (Jackson & Yariv, 2005, 2007; Jackson & Rogers, 2007; López-Pintado, 2008; Galeotti, Goyal, Jackson Vega-Redondo & Yariv, 2010; Lamberson, 2010). These results explain the propensity for diffusion in the more random Watts-Strogatz networks and provide a means for understanding diffusion in more general networks. The theory reveals that increasing randomness does not always increase diffusion; instead, the relationship is conditional on the form of local reinforcement in the diffusion process.

Use of the concept of stochastic dominance (Rothschild & Stiglitz, 1970) has largely been confined to the theoretical economics and finance literatures, and as such may be unfamiliar to scholars in many of the growing number of fields that study networks. Additionally, proofs of the main results relating stochastic dominance and diffusion rely on a mean-field approach borrowed from statistical mechanics, which may be unfamiliar to many network scientists. The aim of this article is threefold. First, we wish to introduce the concepts of stochastic dominance to a broader audience of scholars interested in network analysis. Second, we seek to explain and provide intuition for the recent results relating stochastic dominance and diffusion without requiring the technical expertise to parse the mean-field arguments. And third, we indicate how the stochastic dominance results connect independent observations from a variety of network diffusion studies as well as when and how these observations can be generalized.

2. Stochastic Dominance

To impose order on the bewildering number of possible networks, theorists have constructed a long list of methods for measuring and categorizing them. Networks can be bipartite, star-shaped, scale-free, regular, Eulerian, Hamiltonian, small-world, connected, planar, or sparse. Each network has a girth, diameter, cyclomatic number, chromatic number, centralization, density, average path length, average degree, clustering coefficient, and fraction of transitive triples (see the books by Diestel (2000) and Jackson (2008) for definitions). Among all of these features, the degree distribution of a network plays a key role in diffusion.

The *degree* of a node in a network is simply the number of edges connected to that node. The *degree distribution* of the network is the probability distribution P defined by setting $P(d)$ equal to the fraction of degree d nodes in the network. One way to quantify the randomness of the network is by examining the heterogeneity in degree across nodes as captured by the spread of the degree distribution. We can order distributions according to their spread using the notion of second order stochastic dominance.

A distribution P *second order stochastically dominates* (SOSD) a distribution P' if:

$$\sum_{C=0}^{D}\left(\sum_{d=0}^{C}P(d)\right)\leq\sum_{C=0}^{D}\left(\sum_{d=0}^{C}P'(d)\right) \qquad (1)$$

for all D. The dominance is strict if the inequality is strict for at least some D. Equivalently, P SOSD P' if for every nondecreasing concave function $u\colon\mathbb{R}\to\mathbb{R}$,

$$\sum_{d=0}^{D_{max}}u(d)P'(d)\leq\sum_{d=0}^{D_{max}}u(d)P(d) \qquad (2)$$

When P and P' have the same mean, second order stochastic dominance is equivalent to the more familiar notion of a *mean preserving spread*. A distribution P_X of a random variable X is a mean preserving spread of a distribution P_Y of the random variable Y if $P_X = P_{Y+Z}$ where $E[Z|Y]=0$. In other words, X is Y with added random noise. This is why mean preserving spreads, and the equivalent relation of stochastic dominance, capture the intuitive notion of "more random." The equivalence of second order stochastic dominance and mean preserving spreads is captured in the following theorem:

Theorem 1. *If P and P' have the same mean then P* SOSD *P' if and only if P' is a mean preserving spread of P* (see e.g. Jackson 2008).

We say that a network Γ second order stochastically dominates a network Γ′ if the degree distribution for Γ SOSD the degree distribution of Γ′, and similarly Γ′ is a mean preserving spread of Γ if the same relationship holds for their degree distributions. We can think of a network that is a mean preserving spread of another as having the same average degree but greater variation in degrees.

For social network analysts, second order stochastic dominance and mean preserving spreads will be reminiscent of measures of network centralization, and degree centralization in particular (Wasserman and Faust, 1994, p.180). For example, Høvik and Gleditsch (1970) describe group level centralization as a measure of dispersion of centrality across nodes, and Snijders (1981) recommends variance in degree as an appropriate network level measure of centralization. From this perspective, a mean preserving spread corresponds to a decrease in centralization, while if a network Γ SOSD Γ′, then Γ will have higher centralization than Γ′.

Figure 1 depicts two networks Γ and Γ', and Figure 2 plots their degree distributions (Figure 1 and all network computations were made in R using the statnet package (Handcock, Hunter, Butts, Goodreau & Morris, 2003)). Both networks have the same number of nodes and edges, and thus the same average degree, but Γ′ is a mean preserving spread of Γ (equivalently, Γ SOSD Γ′). Second order stochastic dominance can be used to order commonly studied network families: assuming the same average degree, a scale-free (or power law) network is a mean preserving spread of an exponential network, which is a mean preserving spread of a Poisson network, which is a mean preserving spread of a regular network.

3. Diffusion

Many studies address the impact of network structure on diffusion (Abrahamson & Rosenkompf, 1997; Watts, 1999; Chwe, 2000; Morris, 2000; Pastor-Satorras & Vespignani 2001a, 2001b; Newman, 2002; Sander et al., 2002; Young, 2003; Cowan & Jonard, 2004; Centola et al., 2005); however, the critical role of stochastic dominance was identified relatively recently (Jackson & Yariv, 2005, 2007; Jackson and Rogers, 2007; López-Pintado, 2008; Galeotti et al., 2010; Lamberson, 2010). In nearly all network diffusion studies, individual nodes of a network are more likely to transition from a status quo state to a new state of interest when more of their neighboring nodes have made that transition; there is some type of local reinforcement. For example, a person is more likely to become infected with a disease when more of their contacts have the disease, or a person is more likely to adopt a new technology

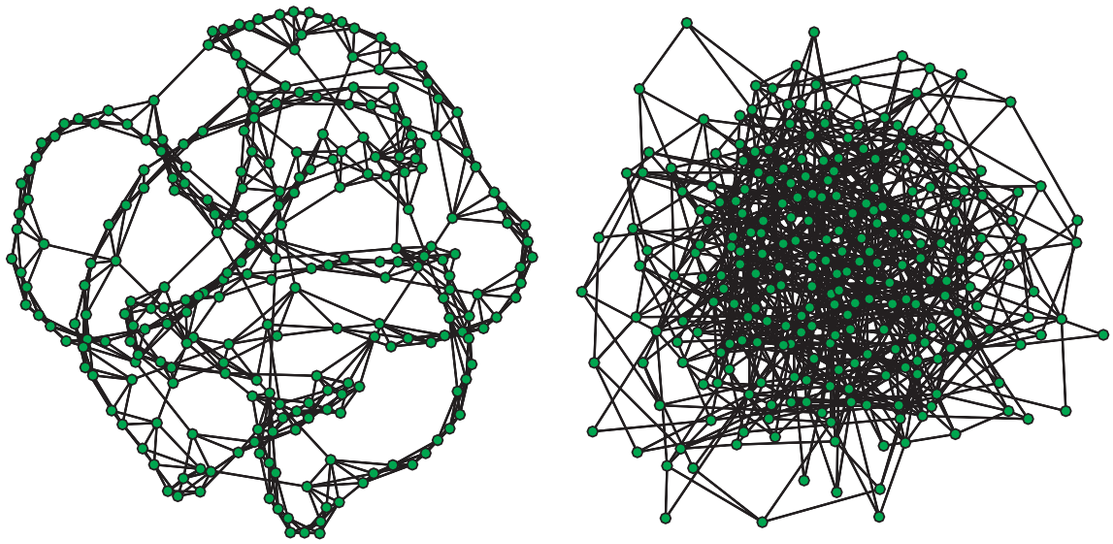


Figure 1. Two networks, $\boldsymbol{\Gamma}$ (left) and $\boldsymbol{\Gamma'}$ (right) with the same number of nodes and edges, but different degree distributions. Γ' is a mean preserving spread of Γ. Both networks are Watts-Strogatz networks with $\boldsymbol{N = 250}$ and $\boldsymbol{k = 3}$ (see the section *Watts-Strogatz Small-worlds* below). For Γ, $\boldsymbol{p = .05}$, and for Γ', $\boldsymbol{p = .9}$.

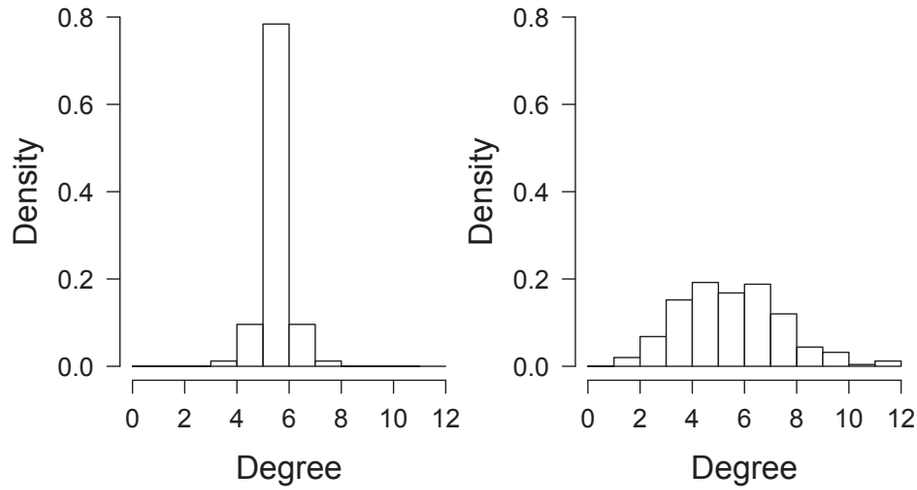


Figure 2. The degree distributions for the networks Γ (left) and Γ' (right) shown in Figure 1.

When more of their friends and family have adopted the technology. The form of local reinforcement differs across models and sometimes depends on parameters within a particular model. Common diffusion mechanisms include *contagion* or *contact* models (Pastor-Satorras & Vespignani 2001a, 2001b), *threshold* models (Valente 1996), and *social learning* models. Stochastic dominance has been applied to understand network diffusion under all three of these models (for contagion see Jackson and Rogers, 2007 and López-Pintado, 2008; for thresholds see Jackson & Yariv, 2005, 2007; for social learning see Lamberson, 2010). Below we discuss focus on contagion and threshold models.

We first consider a variation of the basic susceptible-infected-susceptible (SIS) model of infection (Bailey, 1975) developed by López-Pintado (2008). In this model every individual is in one of two states, susceptible or infected. Susceptible agents that come into contact with infected agents run the risk of becoming infected. An individual can recover from the infection, but once she does she is immediately susceptible to becoming infected again. There is no mortality in the model.

In an SIS model with no network structure, if each contact between an infected individual and a susceptible individual leads to a new infection with probability β, and the probability that an infected individual recovers in each time step is γ, then the disease will spread from an initial infection if:

$$\frac{\beta}{\gamma} > \frac{1}{N} \qquad (3)$$

where N is the total population size.

To add network structure to the model, we adjust the probability of infection to reflect the agents' number of neighbors and number of infected neighbors. The probability that a susceptible agent of degree d with x infected neighbors becomes infected in each small time step is βf(d,x). We call the function f the contagion function and the ratio λ=β⁄γ the effective spreading rate (López-Pintado calls f the contagion function, but we find that terminology somewhat confusing). This specification allows for significant flexibility. For example, by using the contagion function f(d,x)= x/d, we can capture the situation in which agents respond to the fraction of their neighbors that are infected rather than the specific number. The contagion function f could also capture a threshold rule, for example by setting f(d,x)= 1 if x exceeds some threshold, and zero otherwise (to incorporate heterogeneous thresholds, we need a richer framework, such as the one developed by Jackson & Yariv (2007) that we detail below). As in the SIS model without network structure, there is a diffusion threshold λ^* such that if λ>λ^* then the infection will spread from an initial infection to a non-zero steady state; if λ≤λ^* then the infection will die out.

Figure 3 illustrates the diffusion threshold for the two networks shown in Figure 1. The figure plots the average percent infected nodes of each network for different values of the effective spreading rate. The diffusion threshold for Γ' is lower than the diffusion threshold for Γ, and so diffusion occurs more easily in Γ'.

The main result relating stochastic dominance and diffusion is shown in Theorem 2 (López-Pintado 2008).

Theorem 2. *The diffusion threshold for a network* Γ *will be lower than that for a network* Γ' *and thus an infection is more likely to spread through* Γ *than* Γ', *if:*

- Γ *is a* MPS *of* Γ' *and* $d^2 f(d,1)$ *is convex for all* $d \geq 1$, *or*
- Γ' *is a* MPS *of* Γ *and* $d^2 f(d,1)$ *is concave for all* $d \geq 1$.

The theorem says that diffusion is more likely to occur in networks with more variability in the number of connections if $d^2 f(d,1)$ is convex or less variability in connections if $d^2 f(d,1)$ is concave. The result relies on a mean-field approach, which requires several implicit assumptions regarding the diffusion dynamics. Here, we focus on the qualitative implications of the theorem and the broad similarities between the many diffusion models to which it applies in

one form or another, so we will not digress into the technical conditions. For a precise statement and proof see the article by López-Pintado (2008).

To give some intuition behind the theorem, notice that increasing the degree of an agent in the network increases not only the probability that she is infected, but also the probability that she infects someone else. This leads to the d^2term. Thus, the effect of one degree d agent with one infected neighbor on the total infected population varies like $\beta d^2 f(d,1)$. Since we are interested in the diffusion rate in Γ relative to the rate in Γ′, and the β term appears in both, we can ignore it and consider only $d^2 f(d,1)$. Now, we care about the expectation of this measure of infectivity over all agents in the network, so we integrate this against the degree distribution to obtain:

$$\sum_{d=1}^{D_{max}} d^2 \, f(d,1) P(d) \qquad (4)$$

By equation (111), if $d^2 f(d,1)$ is concave and Γ′ is a mean preserving spread of Γ (as in the third bullet of the theorem), then equation (4) is smaller if we replace P with P' i.e. the infection spreads more in Γ than Γ′. The logic for the second bullet is similar, noting that if $d^2 f(d,1)$ is convex then $-d^2 f(d,1)$ is concave.

A useful corollary to Theorem 2 is:

Corollary 2.1. *If* Γ *is a mean preserving spread of* Γ′ *and* f *depends only on* x *(not on* d*), then the diffusion threshold for* Γ *is lower than that for* Γ′ (López-Pintado 2008).

Thus, if the likelihood that any agent becomes infected depends only on the number of infected contacts she has, independent of her total number of contacts, increasing the variation of the degree distribution makes diffusion easier. This follows immediately from Theorem **2** by setting f equal to a positive constant c in the second bullet and observing that $d^2 c$ is convex.

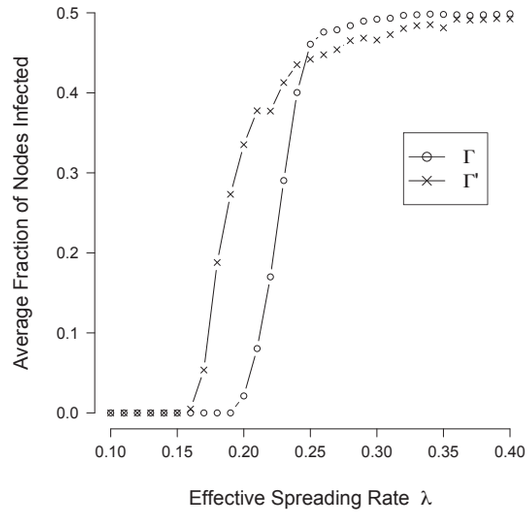


Figure 3. The average fraction of infected nodes for the two networks, Γ and Γ′, shown in Figure 1 for different values of the effective spreading rate $\boldsymbol{\lambda = \beta/\gamma}$ from 10 realizations of the SIS diffusion model with $\boldsymbol{f(d,x) = x}$.

Figure 3 illustrates an example of Corollary 2.1. The contagion function f(d,x)=x depends only on x, so Corollary 2.1 applies. Since Γ' is a mean preserving spread of Γ, Corollary 2.1 implies that the diffusion threshold for Γ' is lower than that for Γ, as the computations depicted in figure 3 confirm.

Jackson and Yariv (2005, 2007) prove a similar result using a threshold model of diffusion. In this model, rather than "becoming infected" we think of individuals as choosing whether or not to adopt a new technology. Suppose that each individual i has a cost c_i associated with adopting the new technology and that the benefits to adopting the technology are specified by a function, $v(d_i, x)$. Here, d_i is the degree of individual i and individual i expects each of her neighbors to be adopters of the new technology with independent probability x. Let F be the cumulative distribution of the costs c_i and define $H(d,x)$ by $H(d,x) = F\big(v(d,x)\big)$. Jackson and Yariv (2007) prove the following theorem, which is analogous to Theorem 2:

Theorem 3. *A network* Γ *will generate more diffusion than a network* Γ' *if:*

- Γ *is a* MPS *of* Γ' *and* $H(d,x)$ *is non-decreasing and convex in* d, *or*
- Γ' *is a* MPS *of* Γ *and* $H(d,x)$ *is non-increasing and concave in* d.

For similar results under different models of diffusion see the articles by Jackson and Rogers (2007), Galeotti et al. (2010), and Lamberson (2010).

4. Network Structure and Diffusion

In this section we describe several models and results on network diffusion and demonstrate how they can be understood through the lens of stochastic dominance and in particular as versions of Theorems **2** and **3** or Corollary 2.1.

4.1. Watts-Strogatz Small Worlds

Some of the most commonly modeled networks are the Watts-Strogatz small-worlds (Watts & Strogatz, 1998). This family of networks is parameterized by a single variable p, that ranges from zero to one. To construct the network corresponding to a given value of p, begin with a ring lattice in which each of N nodes is connected to its k closest neighbors. For each node n of the network, consider each edge connected to that node and with probability p disconnect the opposite end of that edge and reconnect it to another node chosen uniformly at random from all of the nodes not already connected to n. For p=0, the corresponding network is the original regular ring lattice and for p=1 the network is random. For intermediate levels of p the resulting network exhibits two characteristics of many empirical networks, low average path length (the so-called "small-world" phenomenon) and high clustering. The corresponding degree distributions range from a delta distribution when p=0 to a Poisson distribution when p=1. The networks depicted in Figure 1 are Watts-Strogatz networks with N=250 and k=3; Γ (left) has p=.05, and Γ' (right) has p=1. Figure 2 plots their degree distributions.

Since for every p the number of nodes and the number of edges remains the same, the average degree is constant. However, because increasing p increases the heterogeneity of degree across the nodes, if $p > p'$ then the Watts-Strogatz network with parameter p is a mean preserving spread of the Watts-Strogatz network with parameter p'. In light of this ordering and Theorem **2**, many properties of networks that vary monotonically with p in the Watts-Strogatz family may hold more generally for arbitrary networks under the mean preserving spread relation.

For example, in the seminal paper by Watts and Strogatz (1998) in which the small-world family is introduced, the authors consider a standard diffusion model as an illustration of the significance of their construction for dynamic processes. Initially all of the population is healthy and at time zero one infected individual is introduced. Individuals recover at a fixed rate and during each unit of time infect each of their neighbors with probability r. Watts and Strogatz observe through simulation that the critical infectiousness, above which an epidemic sweeps the network and below which the disease vanishes, decreases with p. We can see this as a consequence of stochastic dominance. Since in this case the likelihood that an agent becomes infected depends only on the number of her infected neighbors, independent of her total number of neighbors, Corollary **Corollary 2.1** implies that the Watts-Strogatz networks with a higher p will have a lower diffusion threshold and thus be more prone to diffusion of the infection.

4.2. Concurrency and Disease Spread

Kretzschmar and Morris (1996) investigate the effect of concurrent partnerships on the spread of sexually transmitted diseases. In their model the network of connections represents sexual partnerships and is constantly in flux as new partnerships are formed and old partnerships are dissolved. However, while specific partnerships change, the degree distribution remains relatively unchanged (coincidentally this more closely fits the assumptions of the mean-field approximation than a static network). When an

infected individual is in a sexual relationship with a susceptible individual the disease is transmitted with a fixed probability.

Kretzschmar and Morris consider a population level measure of concurrency (the number of relationships that an individual carries on simultaneously) which they call the index of concurrency and denote κ_3. Since any two edges that connect to the same node in the sexual contact graph at a given time correspond to concurrent relationships, concurrency is related to the degree distribution of the graph. In particular, their measure is approximately:

$$\kappa_3 = \frac{\sigma^2}{\mu} + \mu - 1 \qquad (5)$$

where μ and σ are the mean and standard deviation of the degree distribution respectively. They examine the effect of varying the level of concurrency on the extent that a simulated disease spreads from a single infection and find that the number of agents infected in a fixed time grows exponentially with κ_3.

The relationship between varying levels of κ_3 and stochastic dominance is ambiguous because κ_3 is not monotonically related to μ; however, for the family of networks that Kretzschmar and Morris examine, μ remains fixed. Thus, increases in κ_3 can be accounted for by increases in σ and therefore correspond to mean preserving spreads. Because the probability that an agent becomes infected depends only on the number of her infected partners, Corollary 2.1 applies, so the increased diffusion with increased concurrency is explained by the stochastic dominance relation.

4.3. Increasing Returns and Winner-Take-All Markets

Besides the spread of disease, diffusion models are often used to represent the adoption of products or innovations. In many situations an agent might prefer a product that has been purchased more by other consumers. When this occurs, the market is said to exhibit *increasing returns* (Arthur, 1994). For example, a consumer can expect that more software will be developed for a more popular hardware platform, thus making that hardware platform more desirable. In this case, the increasing returns are global; it is the overall level of adoption in a population that affects the availability of hardware. In other cases, the increasing returns may be local. For example, a professor might prefer to use a computer with the same operating system as her coauthors so that she can more easily share files with them. In this case it is only the choices of the individuals that are "near" the agent in some social sense that affect the agent's purchasing decision, so we call the returns *local*. Arthur (1989) shows that when consumers choose among a set of products based on global increasing returns eventually one of the products will come to dominate the market completely. This outcome is popularly referred to as the *winner-take-all* outcome (Frank, 1996).

When consumers choose based on local increasing returns, multiple products can split the market (Janssen & Jager, 2003; Lee, Lee & Lee, 2006). We observe this "local bias" in reality when, for example, most professors in one field use Macintosh computers while most in another use PCs. The tendency of the market to converge to a winner-take-all outcome or a shared market depends on the structure of the social network. For example, consider the model developed by Lee et al. (2006). A new technology is introduced in two variants, A and B. Consumers choose whether to adopt the new technology at all, and when they do, whether to adopt variant A or variant B. Both global and local increasing returns (Lee et al. refer to them as indirect and direct network effects) drive consumer adoptions. Specifically, an agent i's utility from choosing variant A is:

$$U_i(A) = a_i + \alpha x_{i,A} + \beta \pi_A \qquad (6)$$

where a_i is an agent specific preference for variant A, $x_{i,A}$ is the number of i's neighbors using variant A and π_A is the proportion of all adopters in the population choosing variant A. The α and β terms are weights to adjust the relative strength of the global and local returns. The analogous utility for variant B is:

$$U_i(B) = ab_i + \alpha x_{i,B} + \beta \pi_B \qquad (7)$$

A consumer adopts the new technology when her utility from one of the variants is greater than 0 (one can think of a_i and b_i as costs of adoption, so a consumer adopts when her utility overcomes these costs), and then she chooses whichever variant offers her the greatest utility. Agents are allowed to periodically switch variants if they find that their preference ordering has reversed. Lee et al. simulate the purchases of a population of agents on a Watts-Strogatz family of networks and examine the effect of the network parameter P on the probability that the market converges to a winner-take-all outcome.

For the moment, consider only the diffusion of a single variant, say A, and ignore the global increasing returns $\beta\pi_A$. In this case, the probability of an agent adopting depends only on the number of her neighbors that adopt, so applying Corollary 2.1, we would expect diffusion to occur more easily for higher values of P. The same argument can be applied to variant B, which causes the market to be more unstable when p is higher. The effect is exacerbated by the global increasing returns making a winner-take-all outcome more likely in networks with a higher value of p.

This is exactly the result that Lee et al. find by simulation. With their parameter choices, moving from networks with values of $p < .2$ to those with $p > .5$ changes the frequency of a winner take all outcome from close to zero to nearly one. The theory implies that these results would also hold for other types of networks ordered by second order stochastic dominance.

5. Reversing the Inequality

In all of the examples we have discussed and most examples in the literature, moving from one network to a mean preserving spread of that network increases the likelihood of diffusion. Based on these observations alone, one might conclude that a mean preserving spread of the degree distribution always leads to greater diffusion, but with the theory of stochastic dominance in hand we can see that the effect is conditional on the form of the contagion function.

Corollary 2.1 partially explains the observations. In order to find a case where a mean preserving spread of the degree distribution decreases diffusion, the contagion function must depend not only on the number of infected contacts that an individual has, but also on the degree of the individual. Moreover, suppose that the contagion function is of the form $f(d,x) = xd^{-\alpha}$. A mean preserving spread results in a decrease in diffusion when $d^2 f(d,1) = d^{2-\alpha}$ is concave for all $d \geq 1$, which only holds for values of α strictly between one and two. This is illustrated in **Figure 4**, which plots the percent change in the critical threshold between a Watts-Strogatz network with $p = 0$ ($N = 1000$, $k = 10$) and with $p = 1$ as the exponent α in the contagion

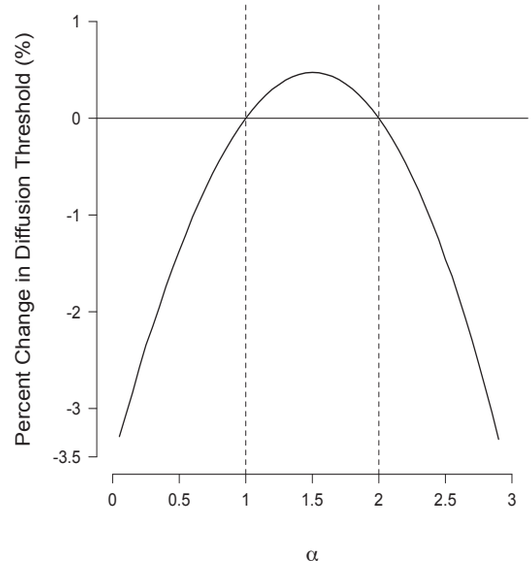


Figure 4. The difference in the critical threshold between a Watts-Strogatz network with $\boldsymbol{p = 0}$ ($\boldsymbol{N = 1000}$, $\boldsymbol{k = 10}$) and with $\boldsymbol{p = 1}$ as the exponent $\boldsymbol{\alpha}$ in the contagion function $\boldsymbol{f(d,x) = xd^{-\alpha}}$ is varied. When $\boldsymbol{\alpha}$ is between one and two, increasing the randomness of the network raises the diffusion threshold and thus makes diffusion less likely. For all other values of $\boldsymbol{\alpha}$, increasing randomness lowers the diffusion threshold.

function is varied. The critical diffusion threshold was calculated for fifty networks of each type using the formula from López-Pintado (2008). For values of α between one and two, moving from a regular lattice to a random network increases the diffusion threshold (by .47% at the most). For any other value of α, the diffusion threshold is lower in a random network than in a regular one, and there is no limit on the magnitude of this effect as α is increased. Of course the contagion function may not be of the form $f(d,x) = xd^{-\alpha}$, but in a sense there are far fewer contagion functions under which a mean preserving spread decreases diffusion than there are under which the opposite relationship holds.

6. Conclusion

Stochastic dominance helps to explain, connect, and generalize the many observations relating network structure to diffusion. Ongoing research continues to expand our understanding of these relationships and their implications for different and more general diffusion mechanisms (Galeotti et al., 2010) as well as policies for exploiting these relationships (Galeotti & Goyal, 2008).

References

Abrahamson, E., Rosenkopf, L., 1997. Social network effects on the extent of innovation diffusion: A computer simulation. *Organization Science* 8 (3), 289–309.

Arthur, W. B., 1989. Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal* 99, 116–131.

Arthur, W. B., 1994. *Increasing Returns and Path Dependence in the Economy*. Ann Arbor, Michigan: The University of Michigan Press.

Bailey, N., 1975. *The Mathematical Theory of Infectious Diseases and its Applications*. London: Charles Griffin and Company Ltd.

Centola, D., Willer, R., Macy, M., 2005. The emperor's dilemma: A computational model of self-enforcing norms. *American Journal of Sociology* 110 (4), 1009–1040.

Chwe, M., 2000. Communication and coordination in social networks. *Review of Economic Studies* 67 (1), 1–16.

Cowan, R., Jonard, N., 2004. Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control* 28 (8), 1557–1575.

Diestel, R., 2000. *Graph Theory*, 2nd Edition. New York: Springer.

Frank, R. H., Cook, P. J., 1996. *The Winner-Take-All Society*. New York: Penguin.

Galeotti, A., Goyal, S., 2008. A theory of strategic diffusion. Unpublished.

Galeotti, A., Goyal, S., Jackson, M. O., Vega-Redondo, F., Yariv, L., 2010. Network games. *Review of Economic Studies* 77 (1), 218–244.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Morris, M., 2003. statnet: Software tools for the Statistical Modeling of Network Data. Seattle, WA, version 2.0. http://statnetproject.org.

Høvik, T., Gleditsch, N. P., 1970. Structural parameters of graphs: A theoretical investigation, *Quality and Quantity* 4 (1), 193–201.

Jackson, M. O., 2008. *Social and Economic Networks*. Princeton, New Jersey: Princeton University Press.

Jackson, M. O., Rogers, B. W., 2007. Relating network structure to diffusion properties through stochastic dominance. *The B.E. Journal of Theoretical Economics* 7 (1), Article 6.

Jackson, M. O., Yariv, L., 2005. Diffusion on social networks. *Économie Publique* 16, 69–82.

Jackson, M. O., Yariv, L., 2007. Diffusion of behavior and equilibrium properties in network games. *American Economic Review* 97 (2), 92–98.

Janssen, M., Jager, W., 2003. Simulating market dynamics: Interactions between consumer psychology and social networks. *Artificial Life* 9 (4), 343–356.

Kretzschmar, M., Morris, M., 1996. Measures of concurrency in networks and the spread of infectious disease. *Mathematical Biosciences* 133 (2), 165–195.

Lamberson, P. J., 2010. Social learning in social networks. *The B.E. Journal of Theoretical Economics* 10 (1) (Topics), Article 36.

Lee, E., Lee, J., Lee, J., 2006. Reconsideration of the winner-take-all hypothesis: Complex networks and local bias. *Management Science* 52 (12), 1838–1848.

López-Pintado, D., 2008. Diffusion in complex social networks. *Games and Economic Behavior* 62 (2), 573–590.

Morris, S., 2000. Contagion. *Review of Economic Studies* 67 (1), 57–78.

Newman, M., 2002. The spread of epidemic disease on networks. *Physical Review E* 66, 016128.

Ohtsuki, H., Hauert, C., Lieberman, E., Nowak, M., 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature* 441, 502–505.

Pastor-Satorras, R., Vespignani, A., 2001a. Epidemic dynamics and endemic states in complex networks. *Physical Review E* 63 (6) 66117.

Pastor-Satorras, R., Vespignani, A., 2001b. Epidemic spreading in scale-free networks. *Physical Review Letters* 86 (14), 3200–3203.

Rothschild, M., Stiglitz, J., 1970. Increasing risk: I. A definition. *Journal of Economic Theory* 2, 225–243.

Sander, L., Warren, C., Sokolov, I., Simon, C., Koopman, J., 2002. Percolation on heterogeneous networks as a model for epidemics. *Mathematical Biosciences* 180 (1-2), 293–305.

Snijders, T., 1981. The degree variance: An index of graph heterogeneity. *Social Networks* 3, 163–174.

Valente, T. 1996. Social network thresholds in the diffusion of innovations. *Social Networks* 18 (1), 69–89.

Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.

Watts, D., 1999. Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology* 105 (2), 493–527.

Watts, D., 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99 (9), 5766 – 5771.

Watts, D., and Strogatz, S., 1998. Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.

Young, H. P., 2003. The diffusion of innovations in social networks. In: Blume, L. E., Durlauf, S. N. (Eds.), *The Economy as an Evolving Complex System III*. Oxford: Oxford University Press.

PJ Lamberson is a Senior Lecturer in the System Dynamics Group at the MIT Sloan School of Management and a Visiting Scholar at the Northwestern Institute on Complex Systems (NICO) and the Northwestern University Kellogg School of Management.

Transitivity & Matrix Operations in Digraphs

Andy Kishida
ARK Internationals, Kobe, Japan

Abstract
The concept of transitivity is of great importance not only in mathematics but also in social sciences. In the present article, a theorem for enumerating non-vacuously transitive triads in digraphs is derived in terms of matrix operations. The theorem is then used to derive two more new theorems related to transitive digraphs and transitive closure. The first theorem provides a criterion for a transitive graph by means of matrix operations. The second theorem shows an alternative way of obtaining a transitive closure that is related to the idea of reachability.

The author would like to thank John. P. Boyd of the University of California, Irvine for a useful discussion.

Note: The present article is based on earlier work the author presented at the Sunbelt XXX Conference, Riva Del Garda (TN), Italy.

Correspondence concerning this article should be addressed to Andy Kishida, ARK Internationals, 7-16-5 Shimoyamate-dori, Chuo-ku, Kobe, Japan 650-0011; email address: akishida-gen@kobe.zaq.jp.

1. Introduction

The concept of transitivity is of great importance not only in mathematics but also in social sciences. It is no exaggeration to say that in social network research it is one of the most fundamental concepts. In fact, some of the earliest research (Holland & Leinhardt, 1970; Wasserman, 1977; Johnsen, 1985) focused on transitivity and developed its mathematical or statistical models. The concept itself is precisely defined in logic or in relational algebra on which digraph theory is based. A triad D is said to be transitive if for any nodes *i, j, k,* an arc *ik* and another *kj* are present in D, then so is an arc *ij* in D. In the present study, we will focus on matrix operations in digraphs and present some new theorems that are primarily mathematical but also applicable in social network research. Some definitions and terminology will be provided below, but we will assume that the reader is familiar with the basics.

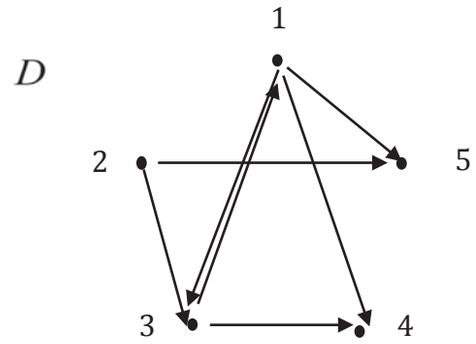
2. Definitions and Terminology

A *digraph* D consists of *points or nodes* $v_1, v_2, \ldots, v_n$ and *lines or arcs* $v_1v_2, v_2v_3, \ldots, v_{n-1}v_n$. The digraph can be expressed as an $n \times n$ *adjacency matrix A(D)* which has n nodes and n^2 cell entries. Each cell entry shows if it has an arc or not as expressed by 1 or 0 respectively. We denote simply A instead of $A(D)$, if no confusion arises. If there is an arc v_iv_j from v_i to v_j, v_i is said to be *adjacent* to v_j. We also write *ij* instead of v_iv_j to denote an arc. All diagonal elements, i.e., *loops*, are usually assumed to be 0 unless the reflexitivity property is assumed. The usual matrix multiplication is denoted by AB while $A \times B$ denotes element-wise multiplication (sometimes called the Hadamard multiplication).

3. Counting Transitive Triads

It is easy to count the number of transitive triads in D if it has a small number of nodes and arcs. As the *order*, the number of nodes, the *size,* and the number of arcs, of D increase, it gets harder to do so by means of visual inspection. It would be desirable to have an alternative method.

A relation R is transitive for any points u, v, w of R, if uRv and vRw, then uRv. In directed graphs, transitivity is defined as follows: A triad is said to be transitive if for any nodes *i, j, k,* an arc *ik* and another arc *kj* are present in D, then an arc *ij* is present in D (Wasserman & Faust, 1994) In algebra, if ρ is transitive, then $\rho \circ \rho \subseteq \rho$ holds where the operation $\circ$ denotes composition (Boyd, 1988). Similarly if a matrix A is transitive, then $A \otimes A \subseteq A$ holds where the operation $\otimes$ is the Boolean multiplication. The following theorem allows us to count the total number of non-vacuously transitive triads in D by means of matrix operations, which is both simple and useful in the following sections.



$$A(D) = \begin{pmatrix} 00111 \\ 00101 \\ 10010 \\ 00000 \\ 00000 \end{pmatrix} \quad A^2(D) = \begin{pmatrix} 10010 \\ 10010 \\ 00111 \\ 00000 \\ 00000 \end{pmatrix}$$

$$Tv(D) = A^2(D) \times A(D) = \begin{pmatrix} 00010 \\ 00000 \\ 00010 \\ 00000 \\ 00000 \end{pmatrix}$$

Figure 1. Digraph D and its matrix operations for counting transitive triads (illustration of Theorem 1).

Theorem 1. *Tv(D)* = $A^2 \times A$ *where* A^2 denotes *AA* by the usual matrix multiplication. An α_{ij} entry of the *Tv(D)* matrix gives the number of transitive triads with an arc ij such that the removal of this arc makes the triad intransitive.

Proof 1. Each entry of A^2 gives the number of length 2 sequences between *i*, *j* via, say, *k* with or without an arc, and *A* is an adjacent matrix whose *i, j* entries are either 1 or 0. If the α_{ij} entry in *A* is present, then $\alpha_{ij} = 1$. If the $\alpha^{(2)}_{ij}$ entry in A^2 is present, then $\alpha^{(2)}_{ij} \geq 0$. Therefore, If A^2 is multiplied by the Hadamard product of *A*, i.e., $\alpha^{(2)}_{ij} \times \alpha_{ij} \geq 0$, the result is a matrix whose *i*, *j* entries are the number of non-vacuously transitive triads each of whose arc starts from point *i* and ends at *j*. Let *Tv(D)* = $[e_{ij}]$. Then we have the following corollary:

Corollary 1. $\mathbf{\Sigma_{i,j}}\ [e_{ij}]$ gives the total number of non-vacuously transitive triads in a given *D*. Figure 1 illustrates this theorem.

4. Criterion for a Transitive Graph by Means of Matrix Operations

A triad is *transitive* if a *sequence* of length 2, v_iv_j and v_jv_k exists in *D*, then v_iv_k exists in *D*. A digraph *D* is a *transitive graph* if for any nodes, *i, j, k*, an arc *ij* and an arc *jk* are in *D*, then an arc *ik* is in *D*. The following theorem shows a criterion for *D* to be a transitive graph by means of matrix operations.

Theorem 1. If $A^2(D) = Tv(D)$, then *D* is a *transitive digraph.*

Let $\alpha^{(2)}_{ij}$ denote the total number of sequences of length 2 from a node *i* to a node *j* in *D*. Let e_{ij} denote the total number of non-vacuously transitive triads from a node *i* to a node *j* in *D* as in Corollary 1.1. We exclude its diagonal elements.

Proof 2. $\alpha^{(2)}_{ij}$ gives the number of length 2 sequences *ik* and k*j* and with or without an arc *ij* in *D*. *D* consists of transitive triads and intransitive triads. These appear in *A*. The e_{ij} entry gives the number of length 2 sequences with an arc *ij* in *D*, which is a transitive triad. Thus, if every sequence of length 2 of *D* has the arc *ij,* then *D* is a transitive digraph so that. $A^2(D) = Tv(D)$. If *D*, however, has intransitive triads, then $A^2(D) \neq Tv(D)$. As a matter of fact, $\mathbf{\Sigma_{i,j}}\ [\alpha^{(2)}_{ij}] > \mathbf{\Sigma_{i,j}}\ [e_{ij}]$ if $A^2(D) \neq Tv(D)$. Thus, the total number of transitive triads $\mathbf{\Sigma_{i,j}}\ [e_{ij}]$ in *D* should correspond to that of $\mathbf{\Sigma_{i,j}}\ [\alpha^{(2)}_{ij}]$ if *D* is a transitive graph.

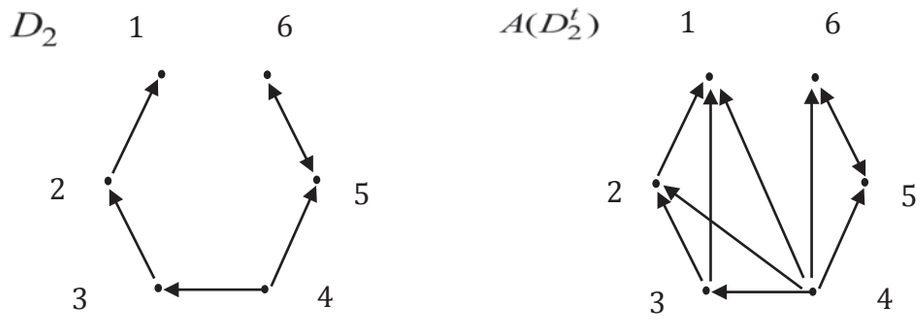
5. Transitive Closure

Transitive closures are usually constructed when we try to obtain a reachability matrix. In social network research, for instance, we would like to know if some messages can reach some people in an organization. The idea of transitive closure is also important in computer programming. Nowadays there are several ways of constructing a transitive closure more efficiently than using matrix methods. The following theorem, however, shows an application of the previous theorems, which allows us to obtain a transitive closure.

The transitive closure D^t of a given digraph *D* is a minimal transitive digraph containing *D* and has the same set of points as *D*. D^t can be constructed from *D* by using the algorithm of Harary, Norman and Cartwright (1965). Here is an alternative way to get $A(D^t)$. By applying Theorem 1, we can find a *transitive closure* by means of matrix operations as follows. First calculate A^2, then obtain a new matrix $M_1 = (A^2 + A)\#$ where # denotes the Boolean operations. Check if M_1 is a *transitive digraph* or not by using Theorem 2. If not, calculate $M_2 = (M_1^2 + M_1)\ \#$ and check if M_2 is a *transitive digraph*, and continue this operation until you get a *transitive digraph*. It is summarized as the following theorem:

Theorem 3. $A(D^t) = M_n = (M_{n-1}^2 + M_n)\ \#$ where $n = 2,3,\ldots$ and $M_1 = (A^2 + A)\#$

Proof 3. *An ij* entry of A^2 is the number of length-2 sequences from a node *i* to a node *j*. Adding the original matrix *A* to A^2 in an element-wise manner by using the Boolean operations creates an *ij* path if the triad *i, k, j* is intransitive, but keeps the *ij* path if the triad *i, k,*



$$A(D) = \begin{pmatrix} 000000 \\ 100000 \\ 010000 \\ 001010 \\ 000001 \\ 000010 \end{pmatrix} \qquad A^2(D) = \begin{pmatrix} 000000 \\ 000000 \\ 100000 \\ 010001 \\ 000010 \\ 000001 \end{pmatrix} \qquad Tv_1 = A^2 \times A = \begin{pmatrix} 000000 \\ 000000 \\ 000000 \\ 000000 \\ 000000 \\ 000000 \end{pmatrix}$$

$$M_1 = (A^2 + A)\# = \begin{pmatrix} 000000 \\ 100000 \\ 110000 \\ 011011 \\ 000011 \\ 000011 \end{pmatrix} \qquad (M_1)^{tr=0} = \begin{pmatrix} 000000 \\ 100000 \\ 110000 \\ 011011 \\ 000001 \\ 000010 \end{pmatrix} \qquad (M_1)^2 = \begin{pmatrix} 000000 \\ 000000 \\ 100000 \\ 210011 \\ 000010 \\ 000001 \end{pmatrix}$$

$$Tv_2 = [(M_1)^2 \times M_1] = \begin{pmatrix} 000000 \\ 000000 \\ 100000 \\ 010011 \\ 000000 \\ 000000 \end{pmatrix} \qquad M_2 = ((M_1)^2 + M_1)\# = \begin{pmatrix} 000000 \\ 100000 \\ 110000 \\ 111011 \\ 000011 \\ 000011 \end{pmatrix} \qquad (M_2)^{tr=0} = \begin{pmatrix} 000000 \\ 100000 \\ 110000 \\ 111011 \\ 000001 \\ 000010 \end{pmatrix}$$

$$((M_2)^{tr=0})^2 = \begin{pmatrix} 000000 \\ 000000 \\ 100000 \\ 210011 \\ 000010 \\ 000001 \end{pmatrix} \qquad Tv_3 = [(M_2)^2 \times M_2\ ] = \begin{pmatrix} 000000 \\ 000000 \\ 100000 \\ 210011 \\ 000000 \\ 000000 \end{pmatrix}$$

Figure 2. D_2 and matrix operations for obtaining $A(D_2^t)$ (illustration of Theorem 3).

j is transitive in A, which results in the matrix $(A^2 + A)$#. Then using Theorem 2, we test if the resultant matrix $M_1 = (A^2 + A)$# is a transitive graph or not. If so, we stop here. If not, we repeat the above procedure by calculating $M_2 = (M_1^2 + M_1)$ #. The reader should be referred to Harary et al. (1965) for the sake of comparison. The present method allows us to derive $A(D^t)$ without calculating A^n all the way to A^{p-1}. Figure 2 illustrates this theorem.

5.1. Summary of the Matrix Operations for Transitive Closure

1. First we calculate A^2 and $Tv_1 = A^2 \times A$. Then we set the main diagonal of A^2 and Tv_1 0.
2. We compare A^2 with $(Tv_1)^{tr=0}$. Since $A^2 \neq (Tv_1)^{tr=0}$, D is not transitive.
3. Next we calculate $M_1 = (A^2 + A)$# and set its main diagonal 0 so that we make the matrix $(M_1)^{tr=0}$.
4. Then we calculate $((M_1)^{tr=0})^2$ and $Tv_2 = [((M_1)^{tr=0})^2 \times ((M_1)^{tr=0}]$ and set the main diagonals of $((M_1)^{tr=0})^2$ and Tv_2 0. Since $((M_1)^{tr=0})^2 \neq Tv_2$, the digraph whose matrix is M_1 is not transitive.
5. Therefore, we calculate $M_2 = ((M_1)^2 + M_1)$ # and set its main diagonal 0 so that we make the matrix $(M_2)^{tr=0}$.
6. Likewise, we calculate $(M_2)^2$ and $Tv_3 = [(M_2)^2 \times M_2]$ and set the main diagonals of $(M_2)^2$ and Tv_3 0. Since $(M_2)^2 = Tv_3$, we stop here. The digraph $A(D_2^t)$ whose matrix is M_2 is transitive and is the transitive closure of D_2.

References

Boyd, J. P. (1988). Social semigroups: A unified theory of scaling and blockmodelling as applied to social networks. Fairfax, VA: George Mason University Press.

Harary, F. (1972). Graph theory. Reading, MA: Addison-Wesley.

Harary, F., Norman, R.Z., and Cartwright, D. (1965) Structural models: An introduction to the theory of directed graphs. New York: John Wiley and Sons, Holland, P.W., and Leinhardt, S (1970). A method for detecting structure in Sociometric data. American Journal of Sociology 70, 492-513.

Johnsen, E.C. (1985). Network macrostructure models for the Davis-Leinhardt set of empirical sociometrices. Social Networks. 7, 203-224

Wasserman, S. (1977). Random directed graph distributions and triad census in social networks. Journal of Mathematical Sociology. 5, 61-86.

Wasserman S., and Faust K., (1994). Social network analysis methods and applications. Cambridge: Cambridge University Press.

Andy (Akishige) Kishida, Ph.D. did his graduate work at the University of California, Irvine. Currently, he is a researcher at Ark Internationals.

Matter Over Mind?
E-mail Data and the Measurement of Social Networks

Eric Quintane
Institute of Management, University of Lugano, Switzerland
School of Behavioral Science, University of Melbourne, Australia

Adam M. Kleinbaum
Dartmouth College, Tuck School of Business, Hanover, New Hampshire, USA

Abstract
Organizational network scholars have not yet fully exploited the information revolution for data on intra-organizational social networks. To encourage research using electronic data, we analyze the correspondence between e-mail and survey measures of the same social network. Substantively, we find that clustering is explained by actor attributes (hierarchy, tenure and group) in the survey measure, but appears to be endogenous in the e-mail measure – that is, relative to electronic traces of observable interactions, survey respondents tend to over-state ties to high-status alters and under-state ties to physically and organizationally distant alters. We conclude that survey data provide information about actors' perceptions of a network and should be used when those perceptions are of substantive interest. In contrast, observational data such as e-mails measure the objective communication structure and are a better data source for research questions that depend on measurement of the actual flow of communications.

The authors would like to thank Galina Daraganova, Dean Lusher, Mikołaj Jan Piskorski, Pip Pattison, Gary Robbins, Toby Stuart, Mike Tushman, Tom Valente, Peng Wang, seminar participants at the Harvard Business School, and conference attendees at INSNA's SunBelt conference for their valuable comments and suggestions. Any remaining errors are our own.

Correspondence concerning this article should be addressed to Eric Quintane, Institute of Management, University of Lugano – Lugano, Switzerland, eric.quintane@usi.ch; phone +41 58 666 44 74.

1. Introduction

In recent years, the interdisciplinary field of network analysis has exploded with activity; biologists, physicists, mathematicians, computer scientists, economists, sociologists and organizational scholars have all made significant contributions to the field (for a review, see Watts, 2004). Much of the empirical work, particularly that published in natural science journals, such as *Science* and *Nature*, has taken advantage of advances in information technology, drawing data from electronic communication sources (e-mail, mobile phone records, instant messaging, SMS, etc); and computational power, analyzing data using more complex algorithms over larger data sets than ever before (Onnela, et al., 2007; Watts & Strogatz, 1998).

Yet, in spite of the rapid development of new methods for network analysis, Marsden, in his recent review (Marsden, 2005) was forced to reiterate his two-decade old observation that network scholars continue to rely primarily on traditional survey-based methods to test and advance substantive theory (Marsden, 1990). Electronic data is only just beginning to make inroads into organizational network analysis. When electronic data is used, it is often for the sake of describing the properties of extremely large networks (e.g., Ebel, Mielsch, & Bornholdt, 2002; Eckmann, Moses, & Sergi, 2004; Kossinets & Watts, 2006), rather than to advance the frontiers of organizational theory (cf. Ahuja, Galletta, & Carley, 2003; Bulkley & Van Alstyne, 2004).

One reason for the paucity of social science research using electronic data lies in the nature of the questions that social scientists study. Sociological and organizational analysis often requires fine-grained information about the type, content, and relevance of social relations – information that may be more easily accessible using smaller scale, more precise, survey questionnaires. Yet, we concur with Lazer et al. (2009), who suggest that there are significant insights to be gleaned from attempting to analyze large electronic datasets from a sociological perspective (e.g., Szell & Thurner, 2010; Wimmer & Lewis, 2010). We suggest that an important step in establishing a link between network surveys and the use of electronic data for substantive network analysis is to provide guidance as to what electronic network data actually signify. From the outset, e-mail data appear to depart substantially from survey data: 1) e-mail exchanges are captured over time and lead to continuous data while survey data is collected at a moment in time; 2) e-mail data is based on observable interactions while survey data is based on participants' responses; 3) finally, e-mails are interactions that occur through one medium of communication while survey provides information about specific relations between individuals. It is clear that these differences matter, but what is not clear is how they are manifest in the social processes that structure e-mail and survey networks. More generally, are networks obtained using survey data and e-mail data as incommensurable as these differences may indicate?

To answer this question and to clarify the meaning of electronic network data, we empirically examine the correspondence between two different measures – e-mail and survey – of the same social network. After gathering both types of data, we use two analytical methods to compare them: the quadratic assignment procedure (QAP) (Krackhardt, 1987b) and Exponential Random Graph Models (ERGMs) (Robins, Pattison, Kalish, & Lusher, 2007). We demonstrate a QAP correlation of 0.35 ($p < 0.01$), a magnitude comparable with previous comparisons between observational and survey measures of social networks, and we concur with those scholars' assessments that "recall of communication links in a network is not a proxy for communication behavior," (Bernard, Killworth, & Sailer, 1981, p. 11). In order to further explore the correspondence in structure of the two measures we expanded our analysis to ERGMs, which enable us to focus on the social processes occurring at a local level. Our results show that the two measures of the network differ in part due to differences in the locus of clustering: we find that the transitive clustering in the survey measure is explained largely by actor attributes (hierarchy, tenure and group), while the

clustering in the e-mail measure – popularity-based structural homophily – is not. This suggests that actors' behavior in our e-mail network is driven by endogenous processes, while actors' recall of their behavior is based on social factors captured by their own and alters' attributes. We interpret this result as showing that e-mail data is a representation of the actual flows of communication in our case study organization, while survey data provides critical information about status attribution and the existence of perceptual divisions in the organization's communication network.

These results are not entirely unexpected, as they are consistent with prior research on network surveys, highlighting the impact of respondents' cognitive processing on measurement (Bernard, Killworth, & Sailer, 1979; Bernard, et al., 1981) and with the subsequent literature exploring the meaning of individual perception of network positions and features (Krackhardt, 1987a). Yet, they go beyond the findings of previous research by showing that the differences between survey and e-mail networks mainly affect the way structure emerges in recall and behavioral networks. In other words, actors' recall of their behavior affects the emergence of grouping structure in survey responses in a way that is different from the development of grouping structure shown through their e-mail communications.

We do not interpret these results to mean that one method of data collection is consistently better than another; on the contrary, we suggest that electronic and survey data should be used for different research purposes. Electronic sources of network data should be seen as valid, legitimate measures of the structure of observable social interactions in organizations; surveys measure actors' perceptions of these interactions. Hence, we argue that while surveys remain the appropriate method for gathering data to answer research questions that depend on the perception of ties by their participants (e.g. trust, friendship, advice or influence networks), research questions that depend theoretically on observable patterns of interactions between individuals (e.g. knowledge exchange, information flows) would be better answered using electronic communication archives (such as e-mail, phone logs, wiki posts, instant messaging, or other media of electronic communication). We conclude by proposing that a better understanding of the similarities and differences between e-mail and survey as sources of data for social network analysis should lead researchers to reduce the amount of bias that they introduce in their results, but also to open new opportunities to do intra-organizational network research that was heretofore unfeasible, as well as the possibility to revisit long-standing research questions using more appropriate data.

2. Email and Survey Measures of Social Networks

Organizational network data is gathered through a variety of methods (see Zwijze-Koning & de Jong, 2005). At the firm level of analysis, data is generally archival in nature and records linkages between firms by their interlocking directors (e.g., Davis, 1991) or by their joint ventures and alliances (e.g., Gulati, 1998; Stuart, 1998); linkages may be inferred from common ties to third parties, such as venture capitalists co-investing in a start-up firm (Sorenson & Stuart, 2001); or ties may result from co-participation in events (Feld, 1981), such as financing syndicates (Podolny, 1993). At the individual level of analysis, similar archival data sources may be used, such as individuals being linked by their co-attendance at social events (e.g., Breiger, 1974), co-authorship of scholarly papers (e.g., Leydesdorff, 1995) or co-patenting activity (e.g., Fleming, Mingo, & Chen, 2007). But by far, the most widely-used method for collecting fine-grained data about interpersonal, or social, networks is the network survey (Marsden, 1990, 2005), in which individuals self-report on their interactions with others.

Most network surveys are distributed at one point (or a few points) in time to a predefined set of individuals. These individuals are asked questions (name generators) that elicit a list of names or are presented with a roster of individuals that they may have specific relations with. The name generator questions enable the researcher to define precisely the type of

relationships that are of interest, by specifying the content (e.g., trust, friendship), the duration (e.g., in the last three months), or the boundaries (geographic, institutional) of the relationship of interest. Respondents are then asked to provide more information about the type of relationship that they have with each alter (e.g., frequency, emotional proximity, medium). In some cases, respondents are also asked to provide information about the relationships among the alters or even about the relationships among all the actors in the network (Krackhardt, 1987a). While there is a wider diversity in the type of questions that can be asked using a network survey, these steps represent the standard template for gathering social network survey data.

By contrast, the collection of e-mail data does not rely on respondent participation. E-mail is a widespread corporate communication medium, which implies that each employee in a potential target organization has a corporate e-mail account that she is expected to use for business purposes. These e-mail accounts are typically hosted on a corporate e-mail server, which automatically keeps a journal of all the e-mail exchanged in the organization. Obtaining access to this journal provides the researcher with a reliable and complete source of electronic interaction data. Compared to survey, e-mail data is inexpensive and unobtrusive to collect, particularly when the population is large. Yet, it involves many challenges, as it is inherently sensitive and difficult to obtain, its use is subject to concerns about the privacy of communications, and research using e-mail data requires new and different skills from more traditional network data collection methods.

While both e-mail and survey data represent social relations between individuals, they differ in many dimensions. A survey network results from the aggregation of egocentric networks, themselves based on the recall of respondents of a specific type of relation. An email network is constituted of the observation of all the interactions, through one communication medium between a group of individuals, as they evolve over time. We identified three key dimensions that capture the differences between e-mail and survey: 1) e-mail data is longitudinal whereas survey data is collected at a moment in time; 2) an e-mail network is based on observation, whereas a survey network is based on the reports of respondents; 3) an e-mail network is composed of interactions, whereas a survey network is composed of relations. In the remainder of this section, we detail these differences; highlight the potential issues that may result from them; and propose strategies to address them.

2.1. Continuous Versus Cross-sectional Data Collection

A typical e-mail dataset is composed of a series of events, each of them indicating that a message was sent from a given e-mail address to a set of e-mail addresses at a specific point in time. While the continuous nature of e-mail data opens many avenues for research, it is, in practice, difficult for a researcher to fully exploit. The first hurdle to be overcome is the sheer volume of e-mail data generated on a daily basis. Kleinbaum et al. (2008) analyze a sample of over 30,000 employees who collectively exchanged as many as 1.28 million e-mails in a single day; Bulkley and Van Alstyne (2007) recorded 125,000 emails among 71 employees in a recruiting firm over 10 months. The data management and manipulation skills to be gained in order to deal with such large datasets differ from those required to analyze survey datasets, even large ones. A second hurdle is the paucity of tools, models and algorithms that deal with continuous relational data. A few examples exist (e.g., Butts, 2008a for modeling; Moody, McFarland, & Bender-deMoll, 2005 for data visualization), but there is a substantial learning curve to master these tools and methods. Finally, there are even fewer theories and concepts that a researcher can use to analyze a continuous dataset and interpret the result based on a time dimension (Ancona, Goodman, Lawrence, & Tushman, 2001). As a result, researchers using e-mail data tend to aggregate the continuous information into one or a few cross-sectional datasets, which is a more familiar format to traditional social network research. Yet, this aggregation is not a straightforward process and decisions made during the aggregation process

could potentially lead to very different networks (Butts, 2009; Grannis, 2010; cf. Kleinbaum, et al., 2008). To reflect this, we focus this paper on comparing a survey network with a collapsed e-mail network and highlight the potential differences between the two networks.

2.2. Observation Versus Recall

Existing literature shows that the network information obtained differs widely, depending on whether it was gathered by observation or recall. Bernard, Killworth and Sailer ("BKS"), in a series of landmark studies (Bernard & Killworth, 1977; Bernard, et al., 1979, 1981; Bernard, Killworth, & Sailer, 1982; Killworth & Bernard, 1979), attempted to empirically assess the correspondence between the networks obtained from network surveys and from a direct observation of behavior. They examined five different networks, measuring each using both a network survey and an observational approach, in which behavioral interactions among the research subjects were directly observed and recorded. BKS argue that network surveys represent the observable, behavioral reality of interaction patterns as filtered through actors' cognition about those interaction patterns. They assume that cognition obscures, confuses, forgets or otherwise distorts the behavioral reality that is reflected in observational data. Examining the correspondence between the two measures of the network, BKS conclude: "People do not know, with any acceptable accuracy, with whom they communicate; in other words, recall of communication links in a network is not a proxy for communication behavior," (Bernard, et al., 1981, p. 11). Subsequent work moved beyond the facile conclusion that network surveys are inaccurate and attempted to explicate the sources of error and bias that lead to this inaccuracy. For example, Freeman, Romney and Freeman (1987) studied participants in a semester-long academic seminar, to show that individuals' cognitive processing of their social interactions leads to survey results that err in the direction of long-term, stable interaction patterns. This work is informed by the substantive literature on biases of perception (reviewed in Bazerman, 2006).

By contrast, as e-mail data are based on a direct observation of the interaction behavior of individuals within their environment. They provide accurate information about when and with whom the interaction occurred, unaffected by the cognitive processes of the respondent. Yet, the absence of a cognitive process to filter out (or add) specific interactions in e-mail also means that there is no indication of the relevance of any given e-mail tie for a specific individual or research question. As such, researchers using e-mail data find themselves confronted with a multitude of interactions that are not readily distinguishable, even when the content of communications is available to researchers (e.g., Aral & Van Alstyne, 2010). The lack of "cognitive pre-processing" by respondents leads to three specific issues that researchers have to address when using e-mail as a source of social network data. Some issues can be addressed methodologically, while other need to be considered when interpreting the results.

2.2.1. The Density issue

E-mail networks are typically much denser than survey networks with a large proportion of the ties being of low intensity (i.e. weak), but the density of e-mail networks does not have the same meaning as that in a survey network. In survey research a high density is usually interpreted as a representation of social cohesion (Friedkin, 2004). In a cohesive group, members have a higher level of social integration and identification with the group, social norms are better defined, trust is established between actors (Coleman, 1988). By contrast, a high level of density in an e-mail network is not necessarily indicative of such social cohesion. It may represent duplicated information paths or a dynamic task structure in which actors communicate with new partners frequently. Furthermore, as density affects many other structural features of networks (Anderson, Butts, & Carley, 1999), the difference in density between an e-mail and a survey network can lead to distinct structural patterns that do not necessarily warrant different substantive conclusions.

2.2.2. The Stable relationship issue

It is unclear how to recognize a stable relationship in e-mail data. We know that survey respondents tend to bias their reports toward stable, long-term patterns of interaction (Freeman, et al., 1987). For e-mail data to offer comparable insights, it too should measure stable relationships, yet, it is difficult to know how to distill a continuous series of discrete communications into some semblance of a stable social relation. At the same time, the need to capture stable relationships must be balanced against the reality that organizations are organic entities that are constantly changing: individuals move between departments, projects mature and evolve and as a result, interaction patterns change fluidly over time. E-mail data offer both the promise of observing accurately the process through which this change occurs but also the danger of losing the forest amidst the trees of overly-granular data.

2.2.3. The Social significance issue

It is equally difficult for researchers to know which observable communications are socially significant. Social significance may be one of the reasons why survey networks differ from e-mail networks: survey respondents implicitly evaluate the social significance of their relations, systematically including some and excluding others, in ways that e-mail network analysts cannot easily do. Take the example of administrative assistants: many professionals exchange frequent e-mails with their administrative assistants. If a researcher were to assume that frequency of communication is a measure of tie strength, she might infer very strong ties between professionals and their assistants, even as the professionals themselves might report, if asked, that those communications lack any social significance because they are purely administrative in nature.

2.3. Interactions Versus Relations

Survey data usually focus on one or a limited number of specified relations (e.g., trust, friendship, advice) that can be defined precisely through the questionnaire and constitute as many networks as there are types of relationships. By contrast, an e-mail tie, absent the complete text of the message, does not contain information about the content of the interaction. An e-mail circulating a joke among employees is the same to the eyes of the researcher as an email announcing a promotion, approving a budget or organizing a night out. Clearly, different types of content are transmitted through e-mails (work, communications, trust, friendship) and though it is conceivable that different social relations are marked by empirical regularities in their e-mail patterns that would allow researchers to infer different underlying relations, we do not yet have a well-established way of distinguishing between them. An e-mail network is thus less specific than a typical survey network in the type of content that it represents. In other words, we know that the pipes exist (Podolny, 2001), but we do not know what travels through them (again, assuming no access to email content, which is not always the case). Further, e-mail is only one medium of communication out of many that could be observed (telephone, instant messaging, face-to-face). Hence, beyond the question of what content flows through the pipes, it is also possible that we are not capturing all the pipes or that different content tends to travel through different pipes.

Yet, we know that an e-mail network is a communication network; in the intra-organizational setting, we assume that most communication is task-related. In that sense, we are expecting that the content of e-mail is constituted mainly of task-related information (Bulkley & Van Alstyne, 2007). The concept of a communication network is nevertheless quite broad: Monge and Contractor describe it as "the patterns of contact that are created by the flow of messages among communicators through time and space. The concept of message should be understood here in its broadest sense to refer to data, information, knowledge, images, symbols and any other symbolic forms," (Monge & Contractor, 2003, p. 3). Correspondingly, the social processes that are reflected by structural positions or configurations in an e-mail network may be interpreted very differently from a survey network. For example, in-degree centrality in an e-mail network (receiving e-mail

from many different senders) might not so readily be interpreted as prominence or prestige as it would in a survey network (Knoke & Burt, 1983). Receiving many e-mails may be an artifact of the particular tasks a person performs for the organization, which may or may not be associated with prestigious positions; for example, administrative assistants have relatively high degree scores that are not necessarily related to their organizational prestige, precisely because their task is to coordinate the activities of others. Other concepts, evocative of information flow, are very applicable to e-mail data. For example, betweenness centrality is conceptualized as the extent to which an individual can control the flow of information in an organization (Freeman, 1979). As such, we argue that interpreting an e-mail network requires a careful interpretation of the type of concepts that the researcher is attempting to explore.

To the extent that e-mail substitutes for other forms of communication there is a risk that e-mail networks would not be a good approximation of the overall observable patterns of communications of actors in an organization. However, prior literature suggests that at least in some organizations, patterns of e-mail interactions are similar to patterns of face-to-face and telephone meetings (Kleinbaum, et al., 2008). Second, the choice of context is key. While e-mail is generally accepted as a day-to-day communication and work tool in most organizations, choosing a research site in which work is done in offices and requires the communication facilities provided by e-mail is important. We do not argue that e-mail captures all interactions that may occur between individuals, but that it is an acceptable proxy for the overall communication patterns between individuals in a specific context

From the interactional nature of e-mail data emerge another set of issues.

2.3.1. The Dependence issue

A full network coming from survey responses is in fact an aggregation of all the egocentric networks of the respondents. Because data collection is conducted privately, we can assume independence of the answers of all the respondents, (though not independence of the actors from the patterns of interactions that surround them). By contrast, the e-mail data collected for each actor are not independent from the other actors. When actor *i* receives an email from actor *j*, she is aware of it and chooses to respond to it or not. In contrast, when actor *i* receives a nomination from actor *j* in a network survey, she is not aware of it and chooses whether to nominate actor *j* in return independently from the nomination that she received. As such, the notion of reciprocity emerging from e-mail data is distinct from reciprocity in a survey network. In a survey network, a reciprocal nomination is indicative of a symmetric relationship. It is socially meaningful in exploring trust, social obligations and social capital (Scott, 1991). In an e-mail setting, reciprocity may result as an artifact of norms of communication or e-mail etiquette, which dictate – in most organizations – that when one receives an e-mail, one should answer it. Better indicators of strong ties might include long messages, frequent exchange, rapid response, or embeddedness within a more complex set of relationships.

2.3.2. The Recipients issue

The dependence issue is compounded when considering that, as a communication tool, e-mail allows the sender to send the same message to multiple recipients, who are usually aware of who else receives the message. Thus a typical e-mail network does not contain an aggregation of purely dyadic relationships stemming from independent respondents, but a variety of dyads originating from interdependent sources of data. Researchers using e-mail data have tended to treat this feature by selecting a threshold number of recipients after which the e-mail is not considered as a personal communication anymore and excluded from the data set (e.g., Kossinets & Watts, 2006 excluded e-mails with more than four recipients). Yet, including multiple recipients on an e-mail is tantamount to expanding a dyadic interaction to include third parties which, as Simmel (1902) argued, complicates the matter significantly.

Furthermore, the choice of including additional recipients in a given e-mail might reflect distinct social processes (Engel, 2009). As such, the study of an email network that comprises solely e-mails with one recipient may lead to different results from a network that aggregates e-mails sent to up to four recipients.

Taken together, these observations suggest that survey and e-mail networks should differ substantially. In the remainder of this paper, we offer what we believe to be the first empirical study that explicitly compares electronic communication archives with survey data for social network analysis. Using data from both a standard sociometric survey and from the e-mail communications among the same sample of people in the same organization, we investigate the correspondence between the network as measured by survey and by e-mail and address the issues presented above. In doing so, we attempt to understand whether a network based on e-mail data and a network based on survey data are as incommensurable as can be anticipated.

We find a correlation that is similar to that of previous comparisons between behavioral and recall measures of social networks. We further explore the sources of the differences between these network measures using exponential random graph models. We find that the two measures of the network differ mainly due to differences in the locus of clustering: we find that clustering is largely an endogenous process in the email measure while it is explained by actor attributes (hierarchy, tenure and group) in the survey measure. We interpret this result as showing that the lack of correspondence in the global structure of the networks is due to different mechanisms occurring at a local level, with email data representing information flows while survey data provides information about attribution of status and social divisions (Krackhardt, 1987a). In the final section, we conclude that e-mail is a valid and informative source of behavioral data for social network analysis and discuss the implications of this new data source for the field of organizational network analysis.

We must stress, however, that the distinction we make is one of degree, not of kind. We report a substantial, if moderate in magnitude, correspondence between the e-mail and survey networks in our organization. While the differences suggest that survey data provide insight into actors' perceptions and attributions of the social environment, the similarities make clear that these perceptions are deeply rooted in the interactional reality that we observe in the e-mail data.

3. Data and Methods

To empirically assess e-mail data, we gather two measures – e-mail and survey – of the social network among a set of individuals in a medium-sized childcare agency operating in the greater New York area. The organization had 135 total employees; we focus on the 31 who are based in the central office. We chose this particular organization as the research setting because we believe it to be a context in which e-mail is likely to be a reliable measure of the overall communication structure. Physically, the 31 employees in the organization's administrative department (i.e., our sample) are all located in the same building, but are dispersed across three corridors on two floors of the building. Additionally, most offices are in closed rooms – not open spaces – which make face-to-face interactions relatively infrequent. The nature of the organization's work requires that many electronic documents be transferred on a daily basis, including budgets, purchase orders and employees' selection information; because of this, administrative employees are expected to use e-mail as part of their work. In interviews, management affirmed the pervasive use of e-mail, which is universally accepted as a part of the way work gets done in the organization; additionally, while many employees use outside e-mail accounts for personal communications, all internal traffic was believed to occur through the corporate e-mail system.

Our data consist of three parts. The first data set includes information on employees' accounts of their work communication networks which we gathered using a web-based survey instrument. The survey asked the respondents to name up to

ten individuals within the organization with whom they had work interactions (see the full survey instrument in Appendix 1). Because our aim is to compare an e-mail measure of the network with the survey measures that are widespread in the field, we based our survey instrument substantially on the network survey items from the General Social Survey (GSS); the final instrument is similar to those used by most network scholars. We used a free-recall rather than a roster/recognition name generator due to concerns by the project sponsor that the presence of a roster in our survey would induce non-response or incomplete response; free recall methods are likely to be at least as reliable (though perhaps not as complete) as roster methods for network surveys (Ferligoj & Hlebec, 1999; Hlebec & Ferligoj, 2001), especially in a moderately-sized population and with relatively simple survey questions (Butts, 2008b). Additionally, although we limited respondents to a maximum of 10 alters, this constraint was binding on no more than two respondents; the median respondent cited 7 alters. Our analysis here focuses on the four groups of the administrative department: human resources, operations, finance and programs integration. We chose to focus on the administrative department primarily to be responsive to the work context in which actors exist: while the program staff spends significant time in the field, the administrative staff, like most knowledge workers (Drucker, 1959) in the economy, does office-based work at the organization's headquarters. Additionally, we asked respondents to identify which communication media (e-mail, telephone and/or face-to-face) they used in their communications with each alter.

For our second data set, we harvested the e-mail communications of these same employees over a three-month observation window co-terminating with the period during which the survey was administered. E-mail data were received in the form of log files produced by the corporate Exchange server and sent from the organization's information technology department to one of the researchers. The files were then parsed using a software application that was custom-built in Java onto a MySQL database. Because we have chosen to limit our analysis by the boundaries of the organization, actors outside the organization and all communications between them and actors within the organization were removed from the sample. Mass mailings, defined as messages with more than four recipients, were also removed. Mass mailings typically consist of factual information that must be broadcast to multiple people simultaneously; as such, they are unlikely to contain socially meaningful interpersonal interactions. The choice of a particular threshold is inherently arbitrary; we chose a threshold of four because it eliminates the most obvious mass mailings while preserving over 93% of e-mails in our sample and because it is similar to choices made by other scholars (Kleinbaum, et al., 2008; Kossinets & Watts, 2006); however, our results are robust to other threshold choices.

For comparability between the two measures of the network, we collapse the entire three-month observation window of the e-mail data into a single cross section. In our survey instrument, we did not specify a time-frame in order to measure stable, long-term patterns of interaction in the survey data (Freeman, et al., 1987); correspondingly, we capture stable patterns of interaction by aggregating data across the three-month observation window. Setting the observation window to be too short risks systematically omitting stable, long-term ties between people who communicate on a regular, but infrequent basis; conversely, setting the observation window to be too long risks including ties that have since dissolved. In the organization we study, three months appears to offer the optimal balance between stability and fluidity. For analytic tractability, we also dichotomize each measure of the network, counting as a tie any reported interaction in the survey measure and one non-mass e-mail in the e-mail measure.

Our third data set contains attribute information for the individuals in the sample, including each person's group assignment and hierarchical level in the formal organizational structure, and age. All three data sets are linked through the use of encrypted ID numbers for each employee, which

serve to strictly disguise employees' identities from the researchers.

3.1. Sample Selection

We received completed surveys from 23 of the 31 members of the administrative group. Respondents were indistinguishable from non-respondents in terms of their department within the organization, age, and e-mail volume, but were slightly more senior in the organization than non-respondents. Non-respondents were removed from the data set and all analyses were performed on the subset of 23 individuals for whom we have complete data.

3.2. Comparing the Networks Using the Quadratic Assignment Procedure

Our first analysis is a correlational comparison of the two measures of the network using the quadratic assignment procedure (QAP) (Krackhardt, 1987b), as implemented in UCINET (Borgatti, Everett, & Freeman, 2006). QAP is the appropriate method to compare networks: traditional estimation procedures assume independence across observations and would therefore yield incorrect standard errors because of the interdependent structure of network data (Simpson, 2001); QAP avoids this problem by employing a bootstrapping methodology to compute the expected distribution of dyadic-level correlation measures between two networks under a hypothesis of fixed structure in each network but random alignment of nodes (Hubert & Schultz, 1976; Zhao & Robins, 2006).

3.3. Contrasting the Measures Using Exponential Random Graph Models

After assessing the overall similarity of the two measures of the network, we move to systematically explore the differences between them. Exponential random graph modeling (ERGM or p* modeling) is a powerful methodology for the examination of both local network microstructure and actor attributes to determine what factors lead some actors to be tied to one another while others are not. ERGMs come from a long tradition of statistical modeling of social networks (for an introduction and review, see Robins, et al., 2007). They are based on the statistical representation of an observed network using an autologistic model at the dyad level of analysis: the dependent variable is the presence or absence of an individual tie between two actors which is modeled as a function of effects including the local structure of the network surrounding the two actors that are involved in the tie as well as the individual attributes of the actors themselves (Robins, Pattison, & Wang, 2006; Snijders, Pattison, Robins, & Handcock, 2006). Unlike simpler logit models, the autologistic form of ERGMs ensures that careful account is taken of dependencies of observations typical in network data (Anderson, Wasserman, & Crouch, 1999).

The general form of the model for multiple networks is:

$$\Pr(Y = y \mid X = x) = \frac{\exp\left[\sum \lambda_A Z_A(y,x)\right]}{\kappa} \quad (1)$$

where x is the survey network and y is the e-mail network; A is the parameter corresponding to local network configuration; λ_A are the parameter estimates; Z_A (x) is the network statistic counting the frequency of subgraph A in the graph x; κ is a normalizing quantity to ensure that the probability is a proper probability distribution (see Robins, Pattison, & Wang, 2009).

We analyze the e-mail and survey measures of the network using ERGMs with higher-order parameters for directed graphs (Robins, et al., 2006; Snijders, et al., 2006) applied to multiple networks (Pattison & Wasserman, 1999) using the XPnet software package (Wang, Robins, & Pattison, 2006). We chose to model two pairs of networks: Model 1 establishes a baseline and includes only explanatory variables related to local network structure: *Arc* indicates the overall propensity for two randomly-selected individuals to interact, controlling for the other parameters in the model. *Reciprocity* indicates the propensity for the interaction within a directed dyad to be reciprocated. *Survey–Email* reflects the propensity of observing an e-mail tie conditional on the presence of a survey citation or vice versa; unlike our other parameters, *Survey–Email* is

jointly estimated across the two networks in each model. Additionally, we include effects that reflect potential endogenous local network structural mechanisms, including the propensity for various star-like forms (out, in and mixed stars) and triangle-like structures (transitive and acyclic; see Robins, et al., 2009).

In Model 2, we add to this baseline attributes about the individual actors – group, hierarchical level and tenure[1] – as explanatory variables. The organizational hierarchy covariates are parameterized as interactions between *Arc* (i.e. the existence of a directed tie) and the sender's or, separately, the recipient's position in the organizational hierarchy, which ranges from 1 (rank-and-file employee) to 4 (executive office). Thus, a positive coefficient would indicate a tendency for highly-ranked employees to send more (sender effect) or receive more (receiver effect) ties (either survey or email) than lower-ranked employees. Similarly, a positive coefficient for organizational tenure would indicate a tendency for long-tenured employees to send more or receive more ties. Finally, organizational structure covariates are parameterized as interactions between *Arc* or *Reciprocity* and whether the actors are in the same group (1) or different groups (0). Thus, a positive coefficient would indicate that ties are more likely to occur (or to be reciprocated) within groups than across groups.

Within each model, we look to compare the coefficient in the model applied to e-mail data with the corresponding coefficient in the model applied to survey data: to the extent that the coefficients differ, there will be substantive structural differences between the e-mail and the survey measures of the social network. We also look across models to see both the main effects of actor attributes and the effect on structural parameters of controlling for actor attributes. By using QAP analysis to describe the overall, global structures of these two measures of the network; and ERGMs to understand their microstructural differences, we are able to make detailed, fine-grained comparisons between the survey and e-mail measures of the network.

[1] In our primary models, we include the untransformed tenure of sender and recipient in years; for robustness, we separately modeled log-transformations of tenure and found substantively similar results.

4. Results

4.1. Communication Media

In our survey, we asked respondents who they communicated with, as well as which communication media they used with each alter. In our sample of work relations among a small administrative department, most communicating dyads use both face-to-face and e-mail communication. For robustness, we separately analyzed a data set that counted as "tied" only those dyads who claimed to communicate via e-mail and found that the correspondence was no higher. While we believe this was a valuable check on the robustness of our results, we prefer to use all communicating dyads in our primary survey data set because the density of the network is higher, making it more readily comparable with the e-mail data set (see details in the next section); this increases our confidence that our results are not a manifestation of density differentials. This nevertheless indicates that our respondents were not particularly good at remembering with whom they exchanged emails.

4.2. Descriptive Summary Statistics

We begin our quantitative analysis with some summary statistics describing the two measures of the network, which clearly show some similarity (Table 1). As the density of the e-mail measure is higher than that of the survey measure (33% versus 21%), the average degree (number of communication partners) is also higher: 7.30 versus 4.61 for the survey measure. Correspondingly, this density difference has implications for each actor's global proximity to other actors: the e-mail measure has a diameter that is two steps shorter – each actor is a maximum of three links away from every other actor in the e-mail measure, but as much as five links away in the survey measure. Conversely, though, the total adjacency index – the sum of all actors' maximum distances – is higher in the

e-mail measure (168 versus 106). This may be a function of the number of isolates.

Table 1. Summary Statistics Comparing Email and Survey Measures of the Social Network

	E-mail Measure	Survey Measure
Density	0.33	0.21
Average Degree	7.30	4.61
Network Diameter	3	5
Total Adjacency Index	168	106
Reciprocity	0.70	0.49
Clustering	0.56	0.41
Indegree Centralization	32%	35%
Outdegree Centralization	37%	16%

The networks also differ in terms of their reciprocity, clustering and centralization. As expected, the rate of reciprocity – the proportion of all ties for which a tie also flows in the opposite direction – is much larger in the e-mail measure (0.70) than in the survey measure (0.49). Similarly, the clustering coefficient – a measure of the degree to which the average actor's communication partners also communicate with each other – is higher in the e-mail measure (0.56) than in the survey measure (0.41). We also find a higher level of out-degree centralization – a measure of the extent to which a network is organized around a small cluster of active individuals – in the e-mail measure (37%) compared to the survey measure (16%); their in-degree centralization is similar (32% versus 35%).

4.3. QAP Correlation Analysis

Our QAP analysis yields a correlation between the binary e-mail and survey measures of the network of 0.35 ($p < 0.01$). When we use the valued network, we find QAP correlations as high as 0.50 (additional information about robustness analyses available from the authors). While intuition suggests that this is not a particularly strong correlation, we have few reliable baselines against which to judge the magnitude of these results[2]. To estimate the best baselines we know of, we calculated QAP correlations between the self-reported survey measure and the observational measures of four BKS networks: *Frat*, the network among 58 residents of an undergraduate fraternity house; *Hams*, a network of 44 ham radio operators; *Office*, a network of 40 employees in a social science research firm; and *Tech*, the 37-person network of a graduate program in technology education (Bernard & Killworth, 1977; Bernard, et al., 1979). These are the only data sets available to us that explicitly compare an observed measure of communication with a self-reported measure of the same network and, as such, form an ideal baseline against which to assess the similarity between the two measures of our network.

Across the four binary BKS networks, we calculate QAP correlations ranging from 0.29 to 0.46 between observed and self-reported interactions among the same actors (Figure 1); the 0.35 correlation in our organization falls squarely in the middle of this pack. For robustness, we also calculated QAP correlations in the valued data (additional information available from the authors); by this method, our network exhibits a relatively high correlation between measures. Against these baselines, it appears that the correspondence between the e-mail and the survey measures of our social network is similar to that between observational and recall measures of social networks in prior literature. This result gives us greater

[2] Because the QAP algorithm is highly sensitive to even small changes in network density (Krackhardt, 1987b) and the density of the survey network is substantially different from that of the e-mail measure, the upper bound on the correlation is likely less than one; thus, the judgment with which standard correlations are evaluated is not readily applicable to QAP correlations; this correlation may be stronger than our intuition suggests it is.

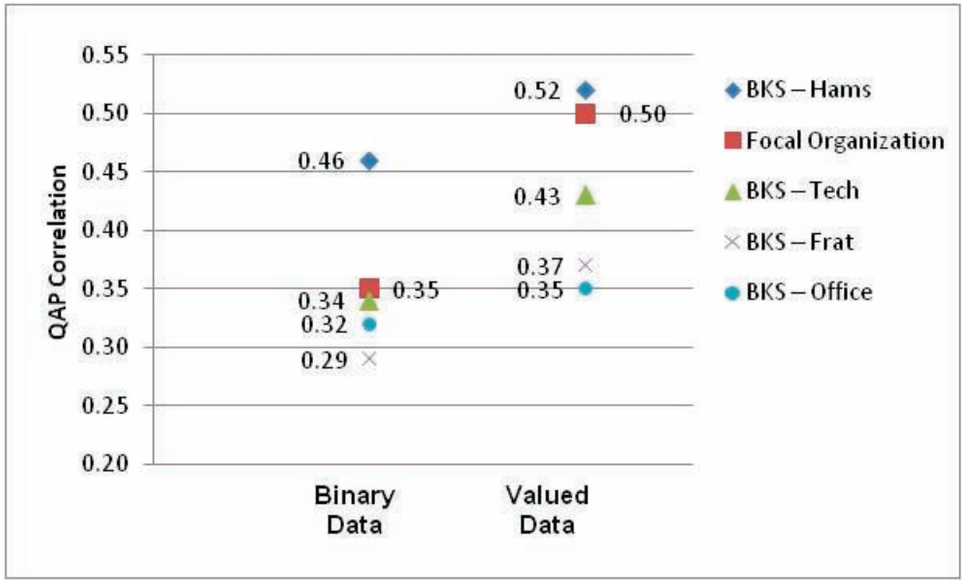


Figure 1. Between-Measure QAP Correlations
QAP correlations of valued data and binary data networks of our focal organization compared to the Bernard, Killworth and Sailor (BKS) datasets. The values represent the QAP correlation between behavioral and recall networks. BKS-Hams, BKS-Tech, BKS-Frat and BKS-Office are the names of the four datasets used in the BKS studies.

confidence that the organization we observe is, in important ways, similar to other organizations that have been studied, but it tells us little about the substantive differences that exist between recall and observational measures of network data.

4.4. Exponential Random Graph Models

Our QAP correlation analysis suggests that although the two measures of the network share a moderate degree of similarity, they also differ in important and meaningful ways; in this section, we explore the sources of those differences using exponential random graph models, sometimes called p* models. Full results are reported in Table 2. To summarize our core results (see Table 2), we find that there is a moderate, statistically significant, degree of similarity between the survey and e-mail measures of the network; that the clustering in the survey measure of the network follows different patterns than the clustering in the e-mail measure of the network; and that while the clustering in the email data is mainly an endogenous process, the clustering in the survey data appears to be driven by the attributes of actors themselves, which influence which alters survey respondents choose to cite. We describe these three core results in turn below.

Consistent with our QAP results, the models indicate that there is a significant, but moderate, degree of alignment between the email and the survey measures of the network. Results on this similarity are captured in the *Survey–Email* parameter in Models 1 and 2, which indicates the propensity of a tie in one network to also exist in the other, net of all other effects in the model; all other coefficients describe differences between the two measures. From Model 1, we see that, net of other effects, people are 3.4 times [=exp(1.23)] more likely to name an individual in one network given mention in the other network.

To confirm that this result indicates substantive similarity and is not a spurious result of

Table 2. Exponential Random Graph Models

	Model 1 No Attributes		Model 2 Attributes	
Parameters	**Survey**	**Email**	**Survey**	**Email**
Arc	-0.52	-4.51*	-2.32	-5.48*
	(1.26)	(0.80)	(1.51)	(0.99)
Reciprocity	1.57*	3.61*	2.72*	4.18*
	(0.46)	(0.50)	(0.99)	(0.82)
Mix2Star	-0.14*	0.06*	-0.08	0.08*
	(0.06)	(0.02)	(0.08)	(0.02)
Popularity Spread	-0.40	-1.69*	-1.13	-1.59*
	(0.53)	(0.46)	(0.72)	(0.57)
Activity Spread	-1.35	1.00*	-0.65	1.36*
	(0.70)	(0.45)	(0.75)	(0.49)
Path Closure [AT-T]	1.12*	-0.49	0.66	-0.85*
	(0.36)	(0.44)	-0.40	(0.41)
Popularity Closure [AT-D]	0.16	1.58*	0.21	1.66*
	(0.35)	(0.45)	(0.38)	(0.48)
Survey-Email	1.23*		0.80*	
	(0.22)		(0.27)	
Receiver Tenure			0.05*	-0.05
			(0.02)	(0.03)
Receiver Hierarchy			0.73*	0.21
			(0.22)	(0.23)
Same Group Arc			1.84*	2.32*
			(0.49)	(0.57)
Same Group Reciprocity			-1.51	-2.28*
			(0.85)	(0.93)

Notes: Standard Errors reported in parentheses
* Significant at $p < 0.05$
Parameters for *Same Ethnicity*, *Same Gender*, *Sender Hierarchical Level, Sender Tenure*, *Sender Age*, and *Receiver Age* were included in all models, but were not significant

differences in network density that arise from different data collection methodologies, we separately modeled e-mail data consisting only of directed dyads who exchanged at least two e-mails (i.e., we excluded dyads in which *i* sent just a single e-mail to *j* during the observation period) with the survey data (Table 3; Models 3 and 4). We selected the threshold of two in order to make the densities of these two measures of the network similar by design (0.21 in the survey measure vs. 0.22 in the 2+ e-mails measure). Results indicate that the *Survey-Email* parameter grows larger in magnitude, from 1.23 in the primary model to 1.54, increasing our confidence that this effect reflects a substantive similarity between the measures and is robust to differences in density.

Table 3. ERGMs with Density Adjusted Email Measures

	Model 3 No Attributes		Model 4 Attributes	
Parameters	**Survey**	**Email (DA)**	**Survey**	**Email (DA)**
Arc	-0.58 (1.27)	-4.58* (0.66)	-1.91 (1.55)	-6.33* (0.92)
Reciprocity	1.50* (0.46)	3.15* (0.45)	2.65* (1.01)	4.29* (0.90)
Mixed–2-Star	-0.13* (0.06)	0.01 (0.04)	-0.08 (0.07)	0.05 (0.03)
Popularity Spread	-0.42 (0.54)	-0.15 (0.44)	-1.16 (0.76)	0.09 (0.46)
Activity Spread	-1.28 (0.75)	0.64 (0.45)	-0.65 (0.74)	0.83 (0.45)
Path Closure	1.07* (0.41)	0.67 (0.41)	0.65 (0.40)	0.27 (0.43)
Popularity Closure	0.17 (0.38)	-0.27 (0.39)	0.21 (0.40)	-0.14 (0.40)
Survey – Email	1.54* (0.23*)		1.17* (0.31)	
Attributes				
Receiver Hierarchical Level			0.70* (0.24)	0.22 (0.23)
Sender Tenure			0.02 (0.03)	0.07* (0.03)
Receiver Tenure			0.05* (0.02)	-0.09* (0.03)
Same Group Arc			1.80* (0.50)	1.73* (0.57)
Same Group Reciprocity			-1.55 (0.84)	-1.58 (0.92)

Notes: Standard Errors reported in parentheses
* Significant at $p < 0.05$
Parameters for Same Ethnicity, Same Gender, Sender Hierarchical Level, Sender Age, and Receiver Age were included in all models, but were not significant.

Our second core finding concerns the locus of clustering in the network. We find evidence of clustering in both the survey and the e-mail data, but the local structures that describe the clusters differ, as evidenced by the structural parameters in Model 1. In the survey measure, clustering occurs along transitive pathways: the probability of *i* nominating *j* is significantly increased when *i* nominates various third parties who also nominate *j*. In other words, triads in the survey measure tend to be closed following a balance principle: if *i* nominates others who nominate *j*,

there is a strong probability that *i* will also nominate *j*. Additionally, the negative *Mix-2-Star* effect reinforces this interpretation as it suggests that transitive paths are unlikely to occur on their own (i.e. without being closed). Thus clustering in the survey data appears to be directed along transitive paths.

By contrast, in the e-mail measure, clustering occurs primarily through a *Popularity Closure* mechanism: the significant positive *Popularity Closure* parameter suggests that popular individuals tend to receive emails from shared alters and to communicate together. This may also be related to the significant negative *Popularity Spread* parameter (-1.69 in Model 1), which indicates that the number of contacts from whom a person receives e-mail (indegree centrality) is more evenly distributed than would be expected, given other effects in the model. That is, while certain individuals may still be more popular than others, these central individuals are more likely to be embedded within dense clusters of relationships. Finally, the positive significant *Mixed-2-Star* (0.06) in the email measure provides additional evidence for this interpretation as it indicates that some individuals behave as information hubs in the network. To summarize our second core finding, there are subtle but important differences in the pattern of clustering between the survey measure of the network and the e-mail measure.

To deepen our understanding of these results and the processes that might have contributed to these structures, we added parameters in Model 2 describing actor attributes, primarily group co-assignments, actors' hierarchical levels, and actor's tenure with the organization. The first finding is that the addition of actor attributes allows us to further tease apart differences between the two measures of the network: the *Survey-Email* parameter is reduced from 1.23 to 0.80. Said differently, when we control for actor attributes as well as local network structure, e-mail interaction patterns explain less of the variation in survey nominations: survey respondents are only 2.2 times [= exp(0.81)] more likely to nominate as a communication partner someone with whom they exchange e-mails when we control for actor attributes – reduced from 3.4 times [=exp(1.23)] in models that exclude actor attributes.

But most interestingly, the introduction of the actor parameters in Model 2 also renders the higher-order structural parameters in the survey measure statistically insignificant. This suggests that the distinctive structural features of the survey measure that are independent of its alignment with the e-mail measure (Model 1) are explained away by the introduction of actor attributes (in Model 2). More generally, actors' recall of their interactions appears to be influenced not only by the actual existence of these interactions, but also by attributes of their communication partners, such as hierarchical level and departmental co-affiliation. The processes that give rise to the structures observed in the survey data in Model 1 appear to have been caused by the actor attribute parameters in Model 2; we discuss the implications of this point more fully below.

Conversely, the existence of structural effects in the e-mail measure over and above the *Email-Survey* parameter and the actor attributes means that the structural processes present in the e-mail measure are not captured well by either the survey measure or the actor attributes alone. The significant negative *Popularity Spread* parameter (-1.59) indicates relative uniformity in in-degree: the distribution of the number of senders for each recipient is homogenous across actors. By contrast, the significant positive *Activity Spread (1.36)* is a sign of heterogeneity of out-degree: some actors are observed to send e-mail to a larger number of alters than others while all actors receive email from similar numbers of alters. This result, too, is unlikely to be an artifact of data collection methods – while each actor's out-degree was limited to 10 in the survey, this constraint was rarely binding – so in practice, neither in-degree nor out-degree was constrained in either data collection method. Further, the fact that this effect only emerges as significant when we control for hierarchical level suggests that there is no main effect of level on in-degree, but that heterogeneity of in-degree occurs at each level of the hierarchy. Additionally, the significant *Mixed-2-Star* (0.08)

effect suggests that some individuals do tend to play important roles in the flow of e-mail communication, by receiving and sending along messages. Finally, in terms of clustering, the introduction of actor attributes does not affect the *Popularity Closure* parameter (1.66) which is still positive and significant, but the *Transitive Closure* parameter (-0.85) is now negative and significant. This confirms the tendency shown in the survey network for hierarchical level and group affiliation to explain transitive closure.

We now turn our attention to the results on the actor attribute variables introduced in Model 2. The *Receiver Hierarchical Level* effect shows that individuals receive more than twice as many survey nominations [exp(0.73) = 2.08] for each step up the four-level hierarchy. In the e-mail measure of the network, the hierarchy effect is not significant: people higher in the organization do not tend to receive more e-mails than those in the rank below them. Furthermore, people at all levels of the hierarchy, on average, receive e-mail from senders who are similarly distributed across the range of hierarchical levels. To confirm that this result is not an artifact of our modeling approach, we also ran independent, single-network models of the survey data and, separately, the e-mail data (additional information available from the authors); in all cases the *Sender Hierarchical Level* and the *Receiver Hierarchical Level* effects were insignificant in the e-mail data. Similarly, the *Receiver Tenure* effect shows that individuals with more tenure in the organization receive more nominations.

In both the survey and e-mail measures of the network, we find a significant *Same Group Arc* effect, indicating that actors both nominate others and send e-mails to others that are in their own department at a higher rate than those from other departments. This suggests that organizational proximity is important both for the recall of communication activity as well as for the observed activity. More surprising is that this attribute is significant in both networks, even when controlling for the *Survey-Email* parameter, indicating that while some dyads may appear in both the e-mail and the survey data, other within-group dyads report communicating in the survey, but are not observed to e-mail, while still others exchange e-mail but do not report communicating in the survey. It is possible that this may be understood as a medium substitution effect, perhaps moderated by physical proximity, whereby some dyads communicate mostly face-to-face while others communicate mostly by e-mail (while still others do both frequently).

The significant negative *Same Group Reciprocity* effect in the email network shows that actors are less likely to reciprocate emails from a group member than from a member of a different group. Again, medium substitution provides one possible explanation for this curious result: if *i* sends an email to *j* within the same group, *j* may come to talk directly with *i* instead of replying via email, as groups are generally co-located. Another possibility is that there are some within-group e-mails that are purely announcements and require no reply, but we believe that most such "broadcast" e-mails were eliminated by our inclusion criterion of four or fewer recipients. Alternatively, it may reinforce our earlier note of the presence of some individuals with an important role in redistributing email communications. The reduced in-group reciprocity may suggest that the redistribution activity of these individuals tends to span group boundaries.

4. Discussion

We began this paper by observing that in spite of recent advances in methods for collecting and analyzing large data sets of electronic communications, the organizations field has been reluctant to adopt e-mail data for substantive network analysis (Lazer, et al., 2009). Although organizational scholars are well-equipped with many ways to explain this collective inertia (e.g., Christensen & Bower, 1996; Tripsas, 1997; Tushman & Anderson, 1986), we suggest that one reason for the field's reluctance may have to do with theoretical and empirical ambiguity about how to interpret a network of electronic communications. To directly assess the similarities and differences between network data drawn from e-mail and from surveys, we gathered data on the communications network of an organization

using both methods and compared them quantitatively. Overall, our results bring us to the conclusion that people's recall and perception of their communication patterns is explained by a social process that differs substantively from their actual communication patterns.

The comparisons show that the networks correspond to only a moderate extent, with QAP correlations of 0.35. Our ERGM results suggest that e-mail and survey measures of one social network have some similarities, but also have predictable differences. In summary, we demonstrate three core results. First, we show that, at least in the organization we study, the correspondence between survey and e-mail measures of the network is significant, if moderate in magnitude; second that survey data and e-mail data both exhibit clustering, but that the processes that give rise to clustering in the survey data differ from the processes of clustering in the e-mail data; and third, that the higher-order structural parameters describing the survey measure cease to be statistically significant when we account for actor attributes, while they remain significant in the email measure. We elaborate on these core results and their implications below.

First, similar to Bernard, Killworth and Sailer before us, we find that the correspondence between observational and recall measures of social networks are moderate in magnitude, with QAP correlations no higher than 0.35. While we reiterate Krackhardt's interpretive warning, we must nevertheless conclude that actors' recall of their social network differs significantly from their observed pattern of interactions. The remainder of our analysis served to explore the nature and origins of these differences.

With respect to survey data, we find significant effects for higher-order structural parameters related to transitivity; but that when we control for hierarchy, tenure and group affiliation, these effects disappear. Said differently, the higher-order parameters that we observe in Model 1 appear to be driven by the actor attributes in Model 2. This result has at least two important implications for social network research. First, it implies that we may have elucidated the process that underlies the creation of the social structural pattern we observe in Model 1: at the individual level, actors tend to over-state their ties to high-ranking, long-tenured or proximate alters; these individual processes give rise to the local microstructures of transitivity that are significant in Model 1, which, in turn, give rise to the global structure that we observe in the survey network. Second, the fact that actor attributes, such as hierarchy, tenure and grouping, play such a dominant role in determining survey nomination patterns suggests the underlying reason why survey data differ from observational data: because some ties are more salient to actors than others. Simply put, survey respondents tend to over-state their ties with high-status people. Consistent with behavioral decision theory on self-serving bias (Babcock & Loewenstein, 1997), individuals try to enlarge their perceptions of their own role and importance in the organization by systematically attending to contacts with high-status others more than contacts with low-status others. This interpretation is also consistent with our descriptive statistics: at the global level, we found much lower levels of clustering in the survey measure of the network, consistent with a propensity for ties to be directed up the hierarchy rather than at co-workers who are likely to also communicate with one another. Furthermore, one of our preliminary interviews provides anecdotal support for this finding: during a structured interview with one supervisor, he cited other supervisors and directors as communication partners, but neglected to cite the staff that reported to him; when explicitly asked, however, he conceded that he does indeed communicate frequently with his staff, in spite of the fact that he failed to mention them initially. Further, our finding that the magnitude of the *Survey-Email* parameter is lower in Model 2 than in Model 1 may reflect a difference between the actual effect of actor attributes on communications and actors' perceptions of that effect.

In contrast to the survey data, where local microstructure appears to be driven by actor attributes, in the e-mail data, the local microstructure appears to be driven mainly by an endogenous process, where actors'

communications themselves determine the overall structure of the network. We find that actor attributes (grouping) are important in determining the patterns of communications but that there are remaining structural effects that actor attributes do not explain. While these structures may, indeed, be driven by heterogeneity in some unobserved attributes of the actors, we nevertheless need the higher order structural parameters to understand the global structure of the e-mail network. In particular, we note that the main closure mechanism that explains clustering in the e-mail measure is a popularity-based structural homophily effect (Robins, et al., 2009). This effect suggest that the e-mail measure may provide a more genuine representation of the organizational communication process, in which individuals who receive e-mail from the same sources will tend to communicate (i.e. work) together, unfiltered by actors' perceptions of their social environments. We are hardly the first scholars to suggest that behavioral and recall network are different, due to biases inherent in the use of self-reported network data; on the contrary, we build on a solid base of empirical evidence to that effect. Where we depart from that tradition, however, is in explicating the underlying social processes that give rise to these differences. While survey responses highlight the perception of social differences in the groups that actors belong to (in our study based on hierarchy, tenure and group affiliation), email interactions provide a clearer picture of the actual information flows in these same groups and how individuals build complex interaction structures in the process of sharing information.

More generally, this research reinvigorates the need to underscore the differences between recall and behavioral measures of social networks. Surveys measure respondents' perceptions of the network, whereas e-mail data records actual, observed interactions, albeit of a single type. Our results suggest that the two capture different realities of the social structure and processes occurring in the organization. Indeed, the cognitive social structure literature (Krackhardt, 1987a) draws on precisely this distinction in examining deviations between perception and observation (see also Kilduff & Krackhardt, 1994). Importantly, we do not contend that there is one observed "reality" that should be measured; rather, we suggest that scholars must choose whether observable interactions or perceptions of interaction patterns are the relevant set of interactions to bring to bear on their particular research question. For example, in research that posits effects of an actor's network on her subsequent choices (e.g., Casciaro & Lobo, 2008), it is the actor's perception of her network that drives her decision-making, so e-mail data would be inappropriate. In contrast, research that demonstrates effects of an actor's structural position on objective outcomes (e.g., Bulkley & Van Alstyne, 2004), where the actor's perception plays no role, are better served using unbiased, observed network ties, as measured using archival data such as e-mail.

5.1. Implications for Research

Our results have important implications for research on social networks, both in terms of research design and in terms of interpretation. At a research design level, we make the elementary, yet oft-neglected, argument that the match between the research question and the type of data collected to answer it is crucial. We argue that *for certain types of research*, e-mail data is both practicable and suitable for network analysis and that as an observational source of data, it provides a more accurate measure of the actual communication structure of an organization. To the extent that the answer to a research question depends on accurate, unbiased measures of behavioral ties among actors who are heterogeneous in status or location (geographic or organizational), we argue that survey data should be avoided, to the advantage of e-mail data.

Furthermore, we suggest that research that explicitly focuses on infrequent, cross-category communications or weak ties may suffer from under-reporting of those ties in surveys and should be controlled for. For example, in his classic work on the social structure of R&D labs, Allen (1977) shows that communication across intra-organizational boundaries is rare and that status places significant constraints on

communication, as high-status actors rarely communicate with low-status actors. Although empirical support is generally wanting, the intuitive appeal of Allen's work has set expectations for a generation of scholars who read it. More recently, Cross and Parker (2004) report a strong hierarchy effect in network connectivity based on a survey, but they interpret the effect to be caused by information-gathering processes. Our analysis of the survey measure of our social network yields results that are consistent with prior literature, but our comparison with e-mail measures of the network suggests that these results confound – at least partially – real, socially meaningful effects with measurement error introduced by response bias.

5.2. Caveats and Limitations

We must qualify our work by acknowledging that this is but a single case study of one, admittedly small, organization. While we believe that our core finding – that survey and email networks exhibit substantially different clustering processes – is likely to apply to many organizations, our results are not formally generalizable beyond the specific context we study. Nevertheless, based on our results and on previous empirical evidence highlighting biases in survey responses, we suggest that researchers should revisit conclusions that have been reached using recall data to answer questions that require behavioral information.

As in all survey research, non-response is a threat to the validity of our findings and a limitation of this study. We used several approaches to mitigate survey non-response: the survey was originally distributed by our sponsor, a senior executive in the organization, who endorsed the project and personally encouraged participation and we sent multiple individual reminders to non-respondents. The empirical literature on the effect of non-response on social network measures is scant, but the few studies we know of suggest that our data are sufficient to address our question. Costenbader and Valente (2003) examine the effects of non-response on 11 difference centrality measures across 8 different networks; with a response rate similar to ours, they find average correlations between the values in the sampled and complete networks ranging from about .55 to about .97[3]; the measures most similar to those that we employ have correlations ranging from .8 to .97. Kossinets (2006) examines the effect of missing data on global network properties and suggests that response rates of 50-70% are sufficient to achieving unbiased results. To empirically explore the robustness of our results to these sampling concerns, we imputed the missing data based on available data using PNet, then re-ran our analysis on the complete network; results were not substantively different[4]. The combination of support for our methods in the literature and consistent empirical results using imputation increases our confidence, but without certainty, that non-response does not undermine our results.

Another limitation of our study concerns the use of a recall, rather than a roster, name generator in our network survey. The literature offers conflicting opinions on the relative merits of roster versus recall name generators in general. Hlebec and Ferligoj (2001) find that the two methods are equally reliable, but suggest that recall methods may elicit stronger ties than roster methods; for our purposes, it does seem plausible that a roster might have mitigated the bias to under-report ties with distant actors by jogging respondents' memories (Brewer, 2000). Such would have been a more conservative test of the self-serving bias to cite up the hierarchy by separating the memory issue from the motivation issue.

Finally, our study is limited by the use of single sociometric items to measure the social network variables. This limitation is typical of social network research because each additional question adds substantially to the burden on respondents and, therefore, reduces response rates. However, single-item survey measures

[3] Excluding Bonacich's eigenvector centrality measure, which was essentially uncorrelated with the true measure when any data was missing, and which we do not employ in our study at all.

[4] For brevity, we do not include these results in the paper, but they are available from the authors upon request.

appear to be highly reliable when researchers use standard data collection methods (Marsden, 1990), and in particular, the use of survey items that have been tested in prior research, as we did. Furthermore, because our goal was to compare the current standard approach with novel methods using e-mail data, we did not want to use measures that were significantly more complex than those of most network scholars.

6. Conclusions

Organizational network research has, to date, not been able to capitalize on the large amount of electronic data available to researchers. Given the increasing ubiquity of information technology in firms of all sizes, organizational network analysis seems an obvious beneficiary of this unbiased, unobtrusive, and widely available source of data. Yet, little substantive research has employed such electronic data to date. We posit that one obstacle to the more widespread seizure of this opportunity seems to be a lack of understanding about e-mail data and what it reveals about interpersonal relationships. In this paper, we empirically examine the correspondence between e-mail and survey measures of a social network.

We find that the two measures of the network are dissimilar to a large extent, comparable in magnitude to previous such comparisons. We find that at least part of this lack of correspondence comes from different clustering resulting from distinct social processes that occur in communication behavior and in the recall that actors have of this behavior. Actors' recall is influenced by predictable biases that are consistent with prior informant accuracy and behavioral decision theory literature – namely that survey respondents over-state ties to high-status others and under-state ties to physically and organizationally distant others, while their behavior is driven by the pattern of interactions that surrounds them. This substantive difference makes the two measures of the network suitable for answering different sorts of research questions: survey data remains the appropriate method for research that is concerned with actors' perceptions of social structure. But for research that depends on accurate measures of social interactions – particularly among a large or distributed population – our results suggest that survey data could provide misleading results.

In contrast, e-mail data, in spite of the many obstacles currently impeding its widespread use – including the difficulty of negotiating access to e-mail data from the organizations we study and the technical skills required to work with e-mail data, which differ significantly from the skills required to field surveys – provide ubiquitous, practicable, valid measures of social networks.

These findings have important implications for research in organizational sociology, such as ensuring the appropriateness of the type of data used to answer a research question, and exploring alternative interpretation of results coming from measures that have been developed from a behavioral perspective. Finally, as other scholars have recently stressed (e.g., Lazer, et al., 2009) we want to highlight the potential for e-mail data to provide opportunities for research that was previously unfeasible, in addition to offering better data for a variety of existing avenues of research. E-mail data offer the possibility to gather information on observable social interactions among all individuals in small, medium and large companies, but also to provide minute observation of these interactions as they evolve through time. It is our hope that our empirical results will contribute to the inevitable, but slowly-developing, adoption of e-mail data as a widely-accepted source of social network analysis.

References

Ahuja, M. K., Galletta, D. F., & Carley, K. M. (2003). Individual centrality and performance in virtual r&d groups: An empirical study. *Management Science, 49*(1), 21-38.

Allen, T. J. (1977). *Managing the flow of technology: Technology transfer and the dissemination of technological information within the r&d organization*. Cambridge, MA: MIT Press.

Ancona, D. G., Goodman, P. S., Lawrence, B. S., & Tushman, M. L. (2001). Time: A new research lens. *The Academy of Management Review, 26*(4), 645-663.

Anderson, B. S., Butts, C., & Carley, K. (1999). The interaction of size and density with graph-level indices. *Social Networks, 21*(3), 239-267.

Anderson, C. J., Wasserman, S., & Crouch, B. (1999). A p* primer: Logit models for social networks. *Social Networks, 21*(1), 37-66.

Aral, S., & Van Alstyne, M. W. (2010). Networks, information & brokerage: The diversity-bandwidth tradeoff. *SSRN eLibrary*.

Babcock, L., & Loewenstein, G. (1997). Explaining bargaining impasse: The role of self-serving biases. *The Journal of Economic Perspectives, 11*(1), 109-126.

Bazerman, M. H. (2006). *Judgment in managerial decision making* (6th ed.). Hoboken, NJ: Wiley.

Bernard, H. R., & Killworth, P. D. (1977). Informant accuracy in social network data ii. *Human Communication Research, 4*(1), 3-18. doi: doi:10.1111/j.1468-2958.1977.tb00591.x

Bernard, H. R., Killworth, P. D., & Sailer, L. (1979). Informant accuracy in social network data iv: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks, 2*(3), 191-218.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1981). Summary of research on informant accuracy in network data and the reverse small world problem. *Connections, 4*(2), 11-25.

Bernard, H. R., Killworth, P. D., & Sailer, L. (1982). Informant accuracy in social-network data v. An experimental attempt to predict actual communication from recall data. *Social Science Research, 11*(1), 30-66.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2006). Ucinet for windows: Software for social network analysis. Natick, MA: Analytic Technologies.

Breiger, R. L. (1974). The duality of persons and groups. *Social Forces, 53*(2, Special Issue), 181-190.

Brewer, D. D. (2000). Forgetting in the recall-based elicitation of personal and social networks. *Social Networks, 22*(1), 29-43.

Bulkley, N., & Van Alstyne, M. (2007). *Email, social capital, and performance in professional services*. Paper presented at the Annual Meetings of the Academy of Management, Philadelphia.

Bulkley, N., & Van Alstyne, M. W. (2004). Why information should influence productivity. In M. Castells (Ed.), *The network society: A cross-cultural perspective*. Northampton, MA: Edward Elgar Publishing.

Butts, C. T. (2008a). A relational event framework for social action. *Sociological Methodology, 38*(1), 155-200.

Butts, C. T. (2008b). Social network analysis: A methodological introduction. *Asian Journal Of Social Psychology, 11*(1), 13-41. doi: doi:10.1111/j.1467-839X.2007.00241.x

Butts, C. T. (2009). Revisiting the foundations of network analysis. *Science, 325*(5939), 414-416. doi: 10.1126/science.1171022

Casciaro, T., & Lobo, M. S. (2008). When competence is irrelevant: The role of interpersonal affect in task-related ties. [Article]. *Administrative Science Quarterly, 53*(4), 655-684.

Christensen, C. M., & Bower, J. L. (1996). Customer power, strategic investment and the failure of leading firms. *Strategic Management Journal, 17*, 197-218.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology, 94*(Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure), S95-S120.

Costenbader, E., & Valente, T. W. (2003). The stability of centrality measures when networks are sampled. *Social Networks, 25*(4), 283-307.

Cross, R. L., & Parker, A. (2004). *The hidden power of social networks : Understanding how work really gets done in organizations*. Boston, Mass.: Harvard Business School Press.

Davis, G. F. (1991). Agents without principles? The spread of the poison pill through the intercorporate network. *Administrative Science Quarterly, 36*(4), 583-613.

Drucker, P. F. (1959). *Landmarks of tomorrow* ([1st] ed.). New York: Harper.

Ebel, H., Mielsch, L.-I., & Bornholdt, S. (2002). Scale-free topology of e-mail networks. *Physical Review E, 66*(3), 035103.

Eckmann, J.-P., Moses, E., & Sergi, D. (2004). Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences, 101*(40), 14333-14337. doi: 10.1073/pnas.0405728101

Engel, O. (2009). Clusters, recipients and reciprocity: Extracting more value from email communication networks. *SSRN eLibrary*.

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

Ferligoj, A., & Hlebec, V. (1999). Evaluation of social network measurement instruments. *Social Networks, 21*(2), 111-130.

Fleming, L., Mingo, S., & Chen, D. (2007). Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly, 52*(3), 443-475.

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

Freeman, L. C., Romney, A. K., & Freeman, S. C. (1987). Cognitive structure and informant accuracy. *American Anthropologist, 89*(2), 310-325.

Friedkin, N. E. (2004). Social cohesion. *Annual Review of Sociology, 30*(1), 409-425. doi: doi:10.1146/annurev.soc.30.012703.110625

Grannis, R. (2010). Six degrees of "Who cares?". *American Journal of Sociology, 115*(4), 991-1017. doi: doi:10.1086/649059

Gulati, R. (1998). Alliances and networks. *Strategic Management Journal, 19*(4), 293.

Hlebec, V., & Ferligoj, A. (2001). Respondent mood and the instability of survey network measurements. *Social Networks, 23*(2), 125-140.

Hubert, L., & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology, 29*(2), 190-241.

Kilduff, M., & Krackhardt, D. (1994). Bringing the individual back in: A structural analysis of the internal market for reputation in organizations. *Academy of Management Journal, 37*(1), 87-108.

Killworth, P. D., & Bernard, H. R. (1979). Informant accuracy in social network data iii: A comparison of triadic structure in behavioral and cognitive data. *Social Networks, 2*(1), 19-46.

Kleinbaum, A. M., Stuart, T. E., & Tushman, M. L. (2008). *Communication (and coordination?) in a modern, complex organization*. Harvard Business School.

Knoke, D., & Burt, R. S. (1983). Prominence. In R. S. Burt & M. J. Minor (Eds.), *Applied network analysis: A methodological introduction* (pp. 195-222). Beverly Hills: Sage.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks, 28*(3), 247-268.

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science, 311*(5757), 88-90. doi: 10.1126/science.1116869

Krackhardt, D. (1987a). Cognitive social structures. *Social Networks, 9*(2), 109-134.

Krackhardt, D. (1987b). Qap partialling as a test of spuriousness. *Social Networks, 9*(2), 171-186.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., *et al.* (2009). Computational social science. *Science, 323*(5915), 721-723. doi: 10.1126/science.1167742

Leydesdorff, L. (1995). *The challenge of scientometrics: The development, measurement, and self-organization of scientific communications*. Leiden: DSWO Press.

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology, 16*(1), 435-463.

Marsden, P. V. (2005). Recent developments in network measurement. In P. J. Carrington, J. Scott & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 8-30). New York: Cambridge University Press.

Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks*. New York: Oxford University Press.

Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology, 110*(4), 1206-1241. doi: 10.1086/421509

Onnela, J. P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., *et al.* (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences, 104*(18), 7332-7336. doi: 10.1073/pnas.0610245104

Pattison, P. E., & Wasserman, S. (1999). Logit models and logistic regressions for social networks: Ii. Multivariate relations. *British Journal of Mathematical and Statistical Psychology, 52*(2), 169-193.

Podolny, J. M. (1993). A status-based model of market competition. *American Journal of Sociology, 98*(4), 829-872.

Podolny, J. M. (2001). Networks as the pipes and prisms of the market. *American Journal of Sociology, 107*(1), 33-60.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks, 29*(2), 173-191.

Robins, G., Pattison, P., & Wang, P. (2009). Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks, 31*(2), 105-117.

Robins, G. L., Pattison, P., & Wang, P. (2006). *Closure, connectivity and degrees: New specifications for exponential random graph (p*) models for directed social networks*. Unpublished manuscript. University of Melbourne.

Scott, J. (1991). *Social network analysis: A handbook*. London: SAGE Publications.

Simmel, G. (1902). *The sociology of georg simmel* (K. H. Wolff, Trans. 1950 ed.). Glencoe, IL: Free Press.

Simpson, W. (2001). The quadratic assignment procedure (qap). *North American Stata Users' Group Meetings*.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., & Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological Methodology, 36*(1), 99-153. doi: doi:10.1111/j.1467-9531.2006.00176.x

Sorenson, O., & Stuart, T. E. (2001). Syndication networks and the spatial distribution of venture capital investments. *American Journal of Sociology, 106*(6), 1546-1588.

Stuart, T. E. (1998). Network positions and propensities to collaborate: An investigation of strategic alliance formation in a high-technology industry. *Administrative Science Quarterly, 43*(3), 668-698.

Szell, M., & Thurner, S. (2010). Measuring social dynamics in a massive multiplayer online game. *Social Networks, 32*(4), 313-329.

Tripsas, M. (1997). Unraveling the process of creative destruction: Complementary assets and incumbent survival in the typesetter industry. *Strategic Management Journal, 18*(Special Summer Issue), 119-142.

Tushman, M. L., & Anderson, P. (1986). Technological discontinuities and organizational environments. *Administrative Science Quarterly, 31*(3), 439-465.

Wang, P., Robins, G., & Pattison, P. (2006). Xpnet: Pnet for multivariate networks. Melbourne, Australia: The University of Melbourne - School of Behavioural Science.

Watts, D. J. (2004). The "New" Science of networks. *Annual Review of Sociology, 30*(1), 243-270. doi: doi:10.1146/annurev.soc.30.020404.104342

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440-442.

Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: Erg models of a friendship network documented on facebook. *The American Journal of Sociology, 116*(2), 583-642.

Zhao, Y., & Robins, G. (2006, April, 2006). *Multiple networks: Comparing qap and exponential random graph (p*) models.* Paper presented at the Sunbelt XXVI International Social Networks Conference, Vancouver.

Zwijze-Koning, K. H., & de Jong, M. D. T. (2005). Auditing information structures in organizations: A review of data collection techniques for network analysis. *Organizational Research Methods, 8*(4), 429-453.

Eric Quintane is a postdoctoral research fellow in the Institute of Management at the University of Lugano and an honorary research fellow in the School of Behavioral Science at the University of Melbourne. His research focuses on interaction patterns between social actors and their evolution into network structures and social processes.

Adam M. Kleinbaum is an assistant professor at the Tuck School of Business at Dartmouth College. His research examines intraorganizational networks, focusing on the origins of their structure and on their consequences for firm performance.

Democracy at Work: Political Participation

Cynthia Baiqing Zhang
Sociology Department, University of Kentucky

Patricia Ahmed
Sociology Department, University of Kentucky

Abstract
By examining the role of ego network density in political participation, this research extends the literature on civil engagement. Based on the theoretical principles of closure and homophily, the paper shows that the external influence on ego from alters in ego's immediate network is important for individual political behavior. Through the analysis of General Social Survey (GSS) 2004, this paper demonstrates that political participation is related with ego network density positively. The more members in an ego network interact, the more likely they will participate in politics and hence democracy. The paper also discusses the group level analysis of political participation such as that of Putnam's (1995).

We are indebted to Prof. Stephen Borgatti and Prof. Daniel Halgin for their invaluable suggestions and feedback. We also thank the editor Prof. Thomas W. Valente for his insightful review opinions.

Please address correspondence to Cynthia Baiqing Zhang, Sociology Department, University of Kentucky; email: Baiqing.zhang@uky.edu; phone: (859)539-6825

1. Introduction

Democracy, as a form of government and more importantly as a social condition, has been accepted widely as more effective than other forms of political systems in coordinating social life. Although it made its first official appearance in ancient Greece, democracy became familiar to many people as a real life model through the works of Alexis de Tocqueville. In the era of Jacksonian government when most parts of the world were still hesitant about a democratic future of mankind, Tocqueville firmly believed "equality of condition" is the inevitable route of human development. In his famous *Democracy in America* (1835-1840), Tocqueville confidently wrote: "In running over the pages of our history for seven hundred years, we shall scarcely find a single great event which has not promoted equality of condition."

With the flourishing of democracies in the present world, no one would continue to doubt the vitality of the system. However, there is a more important question to be asked. That is, how does democracy work? To answer this question, we investigate political participation. Democracy dictates that the majority rules which in turn requires civic engagement. Therefore, individual political behavior requires special attention to understand democracy. We propose that individuals' ego network is important to their political participation in a civil society. Ego network refers to the network composed of the focal person (ego), the ego's direct links (alters), ties between the ego and the alters and those between the alters.

2. Individual Political Behavior

The literature of political sociology centers on two broad concerns: the relationship among social structures, social action and political institutions (or more generally the interaction between "states" and "societies"), and the social bases of individuals' political behavior (Manza, Hout, & Brooks, 1995). In this paper, we focus on the second theme although recent years have witnessed the rise of the study of the relationship between "states" and "societies".

Within the theme of individual political behavior, one string of previous scholarship focuses on the origins or occurrence of gatherings (e.g. Whyte, 1943; Whyte, 1980), of demonstrations (e.g. Blau & Slaughter, 1971; Eisinger, 1973; Morris, 1981), and riots (e.g. Lieberson & Silverman, 1965; Spilerman, 1970; Snyder, 1979). The micro scholars within this research tradition investigate individual behavior to tease out reasons for political participation (McPhail & Wohlstein, 1983). Some other researchers take organizations and associations as the unit of analysis for political participation. Their research emphasizes the mobilizing ability of these organizations such as labor unions (Cornfield, 1991). A third group of researchers (Anderson & Davidson, 1953; Lipset, 1981 [1960]; Korpi, 1983; Esping-Anderson, 1994) stress the importance of "democratic class struggle" for individual voting behavior in capitalist democracies. In contrast to the irrational behaviorist characterization of individuals in political participation, a later theoretical trend turned to collective identity to account for the participation of individuals in politics (Polletta & Jasper, 2001).

A related yet not completely overlapping literature is volunteering. Volunteering refers to any activity in which time is given freely to benefit another person, group or cause. Theories that explain volunteering by pointing to individual attributes can be grouped into those that emphasize motives or self-understandings on the one hand and those that emphasize rational action and cost-benefit analysis on the other. Other theories seek to complement this focus on individual level factors by pointing to the role of social resources, specifically social ties and organizational activity, as explanations for volunteering (Wilson, 2000).

This paper adds to the literature of political participation by taking a network perspective. By analyzing individuals' ego network, specifically the indicator of ego network cohesion – density, we hope to discover potential mechanisms for individual political participation. To draw a more complete picture, we also look at the mediating factors that

potentially affect individual political participation such as political attitudes.

Borgatti's (1998) typology: individual, individual's ego network, group, and external group in the analysis of social capital, provides the insight into the unit of analysis of political participation. What this paper tries to delineate is not social capital per se but the forces that drive individuals to participate in civil society's democratic activities. However, the way the network embedded forces prompt individual political participation is similar to that of social capital. The group level analysis of social capital such as that of Putnam (1995) is often directly linked to civil engagement. We will discuss this later. The focus of this paper is on the individuals' ego network that plays an important role in individual political involvement.

3. Theorizing Political Participation

3.1. Network Closure

"Closed" networks are characterized by a high degree of connectedness between members of a group and are conducive to the functioning of norms, mutual obligations and sanctions (Coleman, 1990). The intense interaction among group members increases the possibility that members' reputations and behaviors are known by others in the group. Mutual obligations, norms and sanctions are more easily observed in a closed network than in a sparse one. The more interactions a network has, the more cohesive it is. Ego networks are closed networks if there are intense interactions among the ego's alters.

It is possible, however, that intense interactions among one's network can also decrease political participation. That is, while closed networks increase peer pressure, to follow group norms we do not know if this increases pressure for political participation. Peer pressure might very well lead to decreased political involvement if that is mandated by the norms fostered by the group cohesion.

3.2. Homophily

Similarity breeds connection. The homophily principle structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, co-membership, and other types of relationship. The result is that people's personal networks are homogeneous with regard to many socio-demographic, behavioral, and interpersonal characteristics. Homophily limits people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience (McPherson, Smith-Lovin, & Cook, 2001). People seek out similar others to interact. As a result of the interaction, they become more similar. Specifically, people with similar political orientation (such as political attitude and voting behavioral pattern) tend to interact and become more similar in the process.

3.3. Density

Density can be used as a measure of cohesiveness. Mathematically, density is the proportion of links among members of a network divided by the number possible (Scott 2000). For undirected networks, density is $T/[n(n-1)/2]$. "T" is the number of ties in network. "N" is the number of nodes. The higher the density of a network is, the more interactive the members of the group are. In this study, we exclude the links from ego to alters, based on Scott's (2000, p.72) recommendation. Such a choice makes the degree of interaction among alters more distinct. The links between the ego and alters are assumed in an ego network.

3.4. Hypotheses

Based on the theories of closed network and homophily, we hypothesize that density is positively associated with participation in politics. In addition, we anticipate the mediating effect of political attitude and other socio-demographic backgrounds on individual political engagement.

4. Data

We use General Social Survey (GSS) 2004 to evaluate the hypotheses. The GSS contains a standard "core" of demographic and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time trend studies as well as replication of earlier findings. The GSS started in 1972 and completed its 26th round in 2006. The GSS is the largest project funded by the Sociology Program of the National Science Foundation. The GSS has been monitoring social change and the growing complexity of American society.

4.1. Measures: Political Participation

To construct a political participation variable, we conducted principle components analysis (PCA) on various organized and unorganized political measures. Organized participation would include activities coordinated by political parties and associations such as political clubs. Unorganized participation may include sporadic political activities such as boycotting products for political reasons. Using PCA, we constructed a political participation index, using the following measures: signed petition, boycotted products for political reasons, joined demonstrations, belongs to a political party and/or club, discussed politics, attended political meetings, joined an internet political forum, contacted politicians or bureaucrats, and donated money. We chose not to include other potentially relevant variables, such as opposed government due to higher uniqueness values or low factor loadings. The final measure yielded a Crohnbach alpha score of .85.

4.2. Measures: Ego Network Density

Ego network density is based on the questions: "Who are the people with whom you discussed matters important to you?", "Which of these people do you feel especially close to?", and "Are the people named close to each other?" Respondents are required to name up to 5 people they discussed issues with and indicate the extent of closeness among the nominated. Since the question with regard to the relationship between the nominated people (alters) is not based only on the presence/absence of ties, we assigned different values to the answers: "Especially close", "Neither close nor strangers", "Total strangers", "No answer" and "Not applicable" to differentiate ties.

In addition, to calculate the density of individuals' ego network, it is important to exclude ego's ties to alters. We calculated the number of ties in an ego network as the sum of the valued ties among alters. Density is the number of ties divided by ties possible n(n-1)/2.

4.3. Control Variables

The analysis included several control variables that measured attitudinal orientation and socio-demographic backgrounds. The inclusion of some of these control variables, especially socio-demographic ones can be found in previous studies (Musick, 2008; Esping-Anderson, 1994) which show they were correlated with civil engagement, especially volunteering.

For attitudinal orientation, we conducted an exploratory factor analysis of 13 attitude variables. These survey items ask whether people agree or disagree with the statements: "governments do not care what people think," "people do not have any say about what the government does," and "most politicians are only for what get out of politics." The variables also ask the degree of importance of issues such as: "to help worse-off people in America," "to help worse-off people in the rest of the world," "citizens have adequate standard of living," "governments protect right of minorities," "people are given chance to participate in decision," "people keep watch on action of government," "people are active on social or political associations," and "citizens engage in acts of civil disobedience." In addition, the variables question respondents on their interest in politics, the likelihood government administrators are to be corrected when making mistakes, and the scale of corruption in public service in America.

The result of the factor analysis shows that there are two factors that may have the potential to

explain political participation. We recoded the first variable as "indignation" as it indicated the indignation people felt concerning whether governments are functioning to accommodate people's needs and whether people have a say in governmental affairs. We recoded the second variable as "universalism" due to its common theme of universal humanitarian ideals people hold.

We included three continuous measures: education, occupational prestige score and income as SES controls. We took the log of income, since the raw variable was extremely right-skewed. We also included three demographic controls. Gender is a dichotomous variable (1=females; 0=males). Age was measured as a continuous variable. We used a polytomous race measure coded 0 for whites, 1 for blacks and 3 for other races.

4.4. Analysis Procedure

Descriptive statistics were first conducted for the key variables. A second step was to run an ordinary least square regression of political participation on the key variables. Then, interaction terms were added to test if density varies with other socio-demographic and attitudinal variables. Following that is a test of the homogeneity of ego network members using E-Net program (Borgatti, 2006). We also conducted sensitivity analysis to evaluate the robustness of the network variable.

5. Results

5.1. Descriptive Statistics

Table I presents descriptive statistics of the key variables. Density of ego network is moderate. This is understandable if we take into consideration the fact that most of the nominated people are family members, coworkers and friends of the nominating ego. Balance theory assumes that if two people (alters) have strong ties with the same person, the possibility that the two of them get to know each other and have strong ties is very high. One reason is that the alters have a lot of opportunities to meet due to their strong ties to the common friend. Another reason is that they might be similar if they like the same person. As stated previously, similarity leads to interaction. The other variables, except for the three variables obtained through factor analysis (political participation, indignation, and universalism), are conventional variables without much deviance.

5.2. Multivariate Analysis

This section details the linear regression analysis of political participation and the key variables. As predicted by the hypothesis, density is positively related with political participation (Table II Simple Regression of Political Participation on Key Variables). The explored variables explain a moderate percentage (21.1%) of variance in political participation. We discuss the robustness of density as a predictive variable below in the sensitivity analysis section.

Education attainment, universalism, indignation, family income and prestige are positively related with political participation. But universalism is not a significant factor. Being black, other race, female, married and older are negatively related with political participation. But being black and older are not statistically significant. This supports our propositions.

Table I. Descriptive Statistics of Key Variables

Var.	Pol. Part.	Density	Uni.	Indig.	Edu. Attain	Prestige	Income	Black	Other Race	Female	Age	Married
Mean	3.24	0.19	10.85	5.91	13.70	45.09	10.46	0.13	0.07	0.54	3.75	0.53
SD	3.50	0.18	2.56	2.17	2.89	14.16	1.09				0.38	
N	2,812	1,070	1,459	1,464	2,810	2,659	2,482	2,812	2,812	2,812	2,803	2,812

Source: GSS 2004

Table 2. Simple Regression of Network Density on Key Variables and Selected Interaction Terms

Independent Variables	Regression Output (p-value)
Density	0.9748*** (.000)
Education Attainment	0.1617*** (0.000)
Universalism	0.0144 (0.439)
Indignation	0.0651 (0.003)**
Black	-0.0257 (0.860)
Other Race	-0.6018 (0.001)***
Female	-0.2570 (0.005)**
Married	-0.2809 (0.005)**
Family Income	0.1257 (0.013)*
Age	-0.1301 (0.304)
Prestige	0.0074 (0.055)
Constant	0.9543
R^2	0.211
N	944
Significant Interactions	
Density*Educational Attainment	0.2290 (0.042)*
Density*Socio-Economic Index	-0.0663 (0.003)**
Density*Married	-1.1740 (0.043)*

Source: GSS 2004

*Note: We denote p-values as follows: *p<=.05, **p<=.01 and ***p<=.001*

5.3. Two-Way Interaction Analysis

To test if density and other variables have two-way interaction, we added interaction terms to the regression (See Table II, Significant Interactions). The interactions of density and education attainment, density and socio-economic index, and density and being married are significant, indicating that density varies along with education attainment, socio-economic index and marital status. All the other interaction terms are not significant and thus are not included in the table.

5.4. Homogeneity Test

To explain the mechanism that density is positively related with political participation, we conducted a cross-tabulate test with E-Net program. The results show that ego network members are highly homogeneous along race and gender. Age showed some heterogeneity tendency for those over 60. This may be a result of the seniors interacting with family members.

5.5. Sensitivity Analysis

We conducted a sensitivity analysis by regressing different values of the density variable on political participation. The results did not indicate any significant change in the relationship between network density and political participation (See Table III). This suggests that the results are safe with regard to unobserved selection bias. Density is a robust variable.

Table 3. Selected Sensitivity Analysis Results

Mean Density	Regression Output
.099	.9748
.123	.9748
.146	.9748
.175	.9748
.166	.9748
.191*	.9748*
.248	.9748
.250	.9748
.294	.9748
.316	.9748

**Original regression model results*

6. Discussion

There are different arguments about political participation. As mentioned earlier, some

sociologists such as Putnam believe formal organizations and associations are important. As a matter of fact, over the past decade, there has been a resurgence of interest in and research into the connections between associations and democracy. This research lies at the intersection of sociology, political science, and democratic theory, and many of those who have made central contributions (Cohen & Rogers, 1995; Putnam, 2000; Skocpol, 1999) operate at the boundaries between these disciplines. By asking the general question "How do associations enhance democracy?" scholars have brought civil society and groups back into the normative and empirical investigation of democracy.

This renewed attention to the multiple mechanisms operating in the space between economy, intimate private life, and formal state structures is welcome. In contrast to many early theorists of democracy such as Rousseau and Madison, much of this research remains quite celebratory or at least hopeful about the contributions that associations can make to democratic governance (Fung, 2003) although they have observed certain sign of decline in association participation. Putnam's (2000) statistics show a steady decline in membership in bowling leagues, bridge clubs, and community and church groups since the 1950s. His "Bowling Alone" (1995) depicts the demise of social capital in the U.S. which may lead to potentially less civil engagement. This work is an example of the collective, group-level approach adopted by some network analysts.

This approach has a long tradition. As stated at the beginning of the paper, it could be traced back to Tocqueville. In *Democracy in America,* Tocqueville argued, "Nothing is more deserving of our attention than the intellectual and moral associations of Americans… Americans of all ages, all conditions, and all dispositions constantly form associations." Tocqueville's associational topography is extensive and nuanced. He distinguished four principal forms of association: permanent, voluntary political, voluntary civic, and small private. Tocqueville believed these associations solidified American civil society.

However, associations and bowling clubs may have the potential to flourish into political involvement, they do not perform well in predicting political participation in our analysis. We conducted regression of political participation on over 20 association and club membership and none of them was a significant indicator of political engagement. The reason may be that many members are not active members. They may simply pay dues and benefit from their membership without being involved in activities the associations organize.

The insignificant relationship between association membership and political participation demonstrates the importance of approaching social reality from an interactive perspective. Of course, we cannot deny that Tocqueville and Putnam and other authors were mainly referring to active member involvement when they illustrated how these organizations helped the construction of civil society. To show the active membership and political participation relationship, we may need to combine some variables with membership variables. This will be the task of a future research.

7. Conclusions

By examining the role of ego network density in political participation, this research extends the literature on civil engagement. The external influence on ego from alters in ego's immediate network is important for individual political behavior. Political participation is related with ego network density positively. The more members in an ego network interact, the more likely they will participate in politics and hence democracy. The mechanism through which members participate in politics in an ego network could be further clarified with more data on the political outlook of ego network members.

What we do know now from the analysis above, especially the sensitivity analysis, is that the more the members of a network interact or the more cohesive the ego network is, the more likely the ego will participate in politics. On the other hand, the more sparse the network is, the less likely the ego will participate in politics.

Given that a cohesive network is constraining on its members through norms, it is safe to say that members of such networks are more homogeneous in political orientation. And homogeneous political orientation is more conducive to political participation.

References

Anderson, D., & Davidson P. E. (1943). *Ballots and the Democratic Class Struggle.* Stanford, CA: Stanford University Press.

Blau, P., & Slaughter, E. (1971). Institutional conditions and student demonstrations. *Sociological Problems*. 18,475-87

Borgatti, S. P., Jones, C., & Everett, M. G. (1998). Network Measures of Social Capital. *Connections (INSNA).* 21(2),29-36.

Borgatti, S.P. (2006). E-NET Software for the Analysis of Ego-Network Data. Needham, MA: Analytic Technologies.

Cohen, J., & Rogers J. (1995). *Associations and Democracy*. London: Verso.

Coleman, J. S. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Cornfield, D. B. (1991). The US Labor Movement: Its Development and Impact on Social Inequality and Politics. *Annual Review of Sociology.* 17,27-49

Esping-Anderson, G. (1994). *The Eclipse of the Democratic Class Struggle? European Structures at fin de siecle.* Unpublished manuscript.

Eisinger, P. K. (1973). Conditions of Protest Behavior in American Cities. *American Political Science Review.* 67, 1 1-28

Fung, A. (2003). Associations and Democracy: Between Theories, Hopes and Realities. *Annual Review of Sociology*. 29,515-540.

Korpi, W. (1983). *The Democratic Class Struggle.* London: Routledge.

Lieberson, S . , & Silverman, A. R. (1965). "The Precipitants and Underlying Conditions of Race Riots. *American Sociological Review*. 30,887-98.

Lipset, S. M. (1981 [1960]). *Political Man* (Rev. ed.). Baltimore: Johns Hopkins University Press.

Manza, J., Hout, M., & and Brooks, C. (1995). Class Voting in Capitalist Democracies since World War II: Dealignment, Realignment, or Trendless Fluctuation? *Annual Review of Sociology*. 21,137-62.

McPhail, C., & Wohlstein, R. T. (1983). Individual and Collective Behaviors within Gatherings, Demonstrations, and Riots. *Annual Review of Sociology*. 9,579-600.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology.* 27,415-44.

Morris, A. (1981). Black Southern Sit-in Movement: An Analysis of Internal Organization. *American Sociological Review*. 46,744-67

Musick, M. A., & Wilson, J. (2008). *Volunteers: A Social Profile.* Bloomington and Indianapolis: Indiana University Press.

Polletta, F., & Jasper, J. M. 2001. Collective Identity and Social Movements. *Annual Review of Sociology*. 27:283-305.

Putnam, R. (2000). *Bowling Alone: The Collapse and Revival of American Community.* New York: Simon & Schuster.

Putnam, R. D. (1995). Bowling alone: America's declining social capital. *Journal of Democracy, 6*, 65-78.

Skocpol, T. (1999). Advocates without members: The recent transformation of American civic life. In T. Skocpol, & M. P. Fiorina (ed.) *Civic Engagement in American Democracy* (pp. 461-509). Washington, DC: Brookings Inst.

Scott, J. (2000). *Social Network Analysis: A Handbook*. Los Angeles, London, New Delhi and Singapore: Sage Publications.

Snyder, D. (1979). Collective Violence Processes: Implications for Disaggregated Theory and Research. In L. Kriesberg (Ed.), *Research in Social Movements, Conflicts, and Change*. (2:35-61). Greenwich, CT: JAI Press

Spilerman, S. (1970). The Causes of Racial Disturbances: A Comparison of Alternative Explanations. *American Sociological Review*. 35,627-49

Tocqueville, A. de. (1835-1840). H. Reeve (trans.), *Democracy in America.* New York: George Dearborn & Co.

Whyte, W. F. (1943). *Street Corner Society*. Chicago: Univ. Chicago Press

Whyte, W. H. (1980). *The Social Life of Small Urban Spaces*. Washington DC: The Conservation Foundation.

Wilson, J. (2000). Volunteering. *Annual Review of Sociology*. 26,215-40.

CONNECTIONS

Cynthia Baiqing Zhang is a doctoral candidate in sociology department at the University of Kentucky. Her specialty areas and interests include: social network analysis, complex organizations and work, and culture and education.

Patricia Ahmed is an assistant professor in sociology department at the University of Kentucky. Her research interests include: culture, methods, and globalization studies.