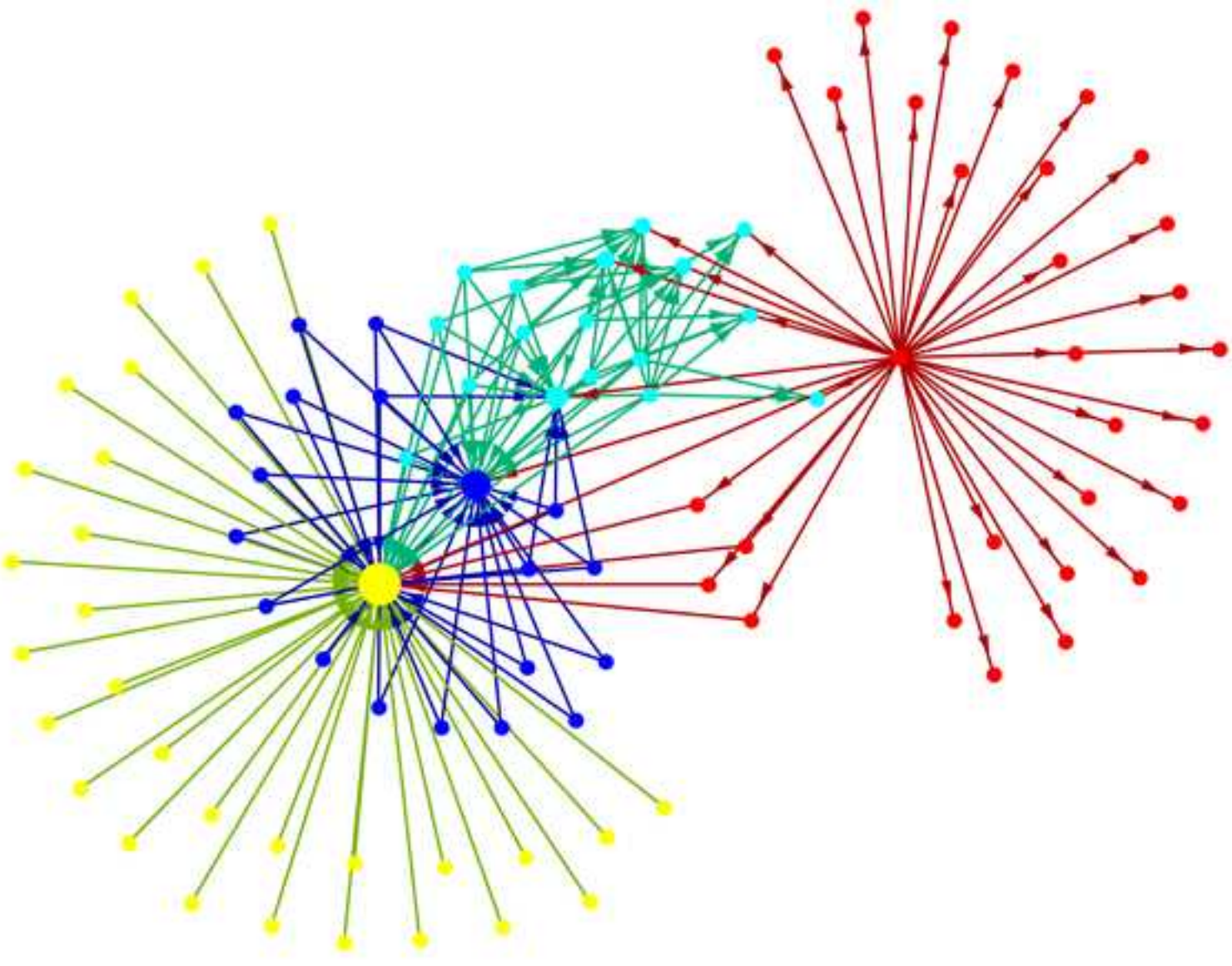
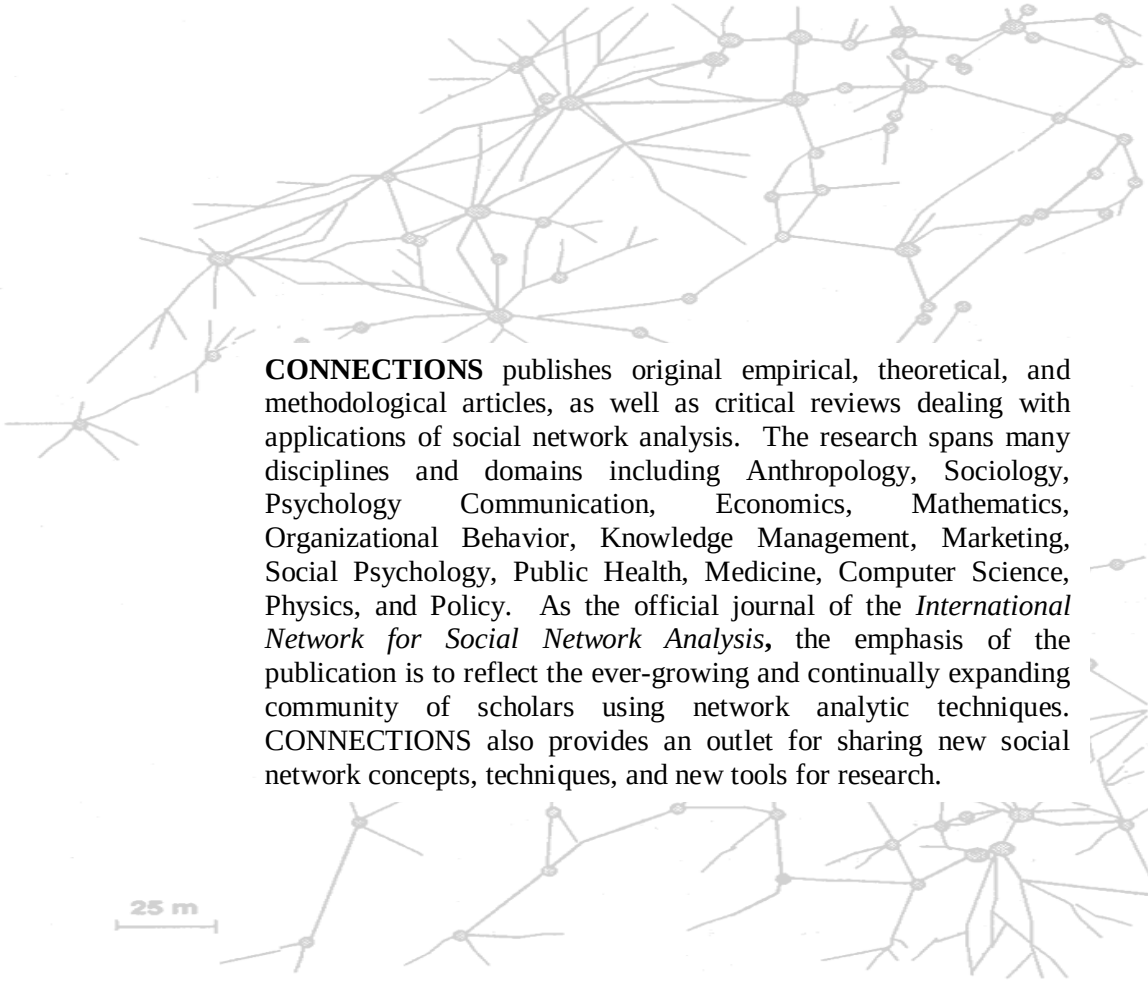




CONNECTIONS publishes original empirical, theoretical, and methodological articles, as well as critical reviews dealing with applications of social network analysis. The research spans many disciplines and domains including Anthropology, Sociology, Psychology Communication, Economics, Mathematics, Organizational 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: *Image is from enclosed article titled "Are We Treating Networks Seriously? The Growth of Network Research in Public Administration & Public Policy" by Sungsoo Hwang and Il-Chul Moon. This is a visualization of a citation network of social network research in Public Administration and Public Policy. Nodes are published articles in SSCI, and are connected with other nodes (articles) they have cited. Colors represent subgroups with different research foci.*

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 *tvalente@usc.edu* or *kate.coronges@usma.edu* for questions or change in address. (ISSN 0226-1776).

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 reached just over one thousand members. Subscription to CONNECTIONS can be obtained by registering for INSNA membership through the website*: www.insna.org*. Standard 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 XXX will be held in Riva del Garda, Italy from June 29-July 4, 2010.

Letters to the Editor
As a new feature of CONNECTIONS, we welcome letters to the Editor on all issues of academic interest. Letters should be brief and will be subject to editing and condensation. In addition, book reviews, network images or review articles will be considered for publication.

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, Thomas Valente at *tvalente@usc.edu*. 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 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. Every issue, we select an image from an accepted article to appear on the front cover of the journal.

November, 2009 **Volume 29 ● Issue 2**

ARTICLES

Are We Treating Networks Seriously?
The Growth of Network Research in Public Administration & Public Policy

Sungsoo Hwang, Ph.D.
Department of Public Administration
Yeungnam University, Korea

Il-Chul Moon, Ph.D.
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology (KAIST)

The purpose of this research is to explore how the term 'network' is used in public administration and public policy. Since O'Toole (1997) first called for scholars of public administration and policy to "[treat] networks seriously," a growing number of researchers use the term network as if it is a rising fashion trend. A recent article by Berry et al (2004) in *Public Administration Review* "Three Traditions of Network Research", illustrates this trend. This article empirically examines the influence of a few prominent scholars on network research over the last decade. Subgroups of network research articles and authors in the citation network are also identified to illustrate the subtopics in network research and to probe what the term network means in these studies. The goal of this research is, in part, to answer Rethemeyer's (2005) call for an empirical examination of network management. Secondly, this article aims to advance the understanding and use of methodology in the public administration discipline by showcasing the use of citation network analysis.

Acknowledgements: *Earlier versions of this work were presented at the 2008 International Sunbelt Social Network Conference and the 2008 Harvard Networks in Political Science Conference. We thank the panel and audience for their feedback, especially Phil Murphy, Chris Weare and Joerg Raab.*

Sungsoo Hwang thanks Jana Diesner for her help with Automap. Il-Chul Moon thanks KAIST as his work was supported by Brain Korea 21 Project, the School of Information Technology, KAIST.

Correspondence: *Contact Sungsoo Hwang at sungsoohwang@ynu.ac.kr.*

INTRODUCTION

What does "network" mean in public administration and policy? What is network analysis in public management studies? We pose these questions as we are encountering an increasing use of the term 'network' in recent public administration and policy scholarship. In public administration and policy, some terms associated with network are particularly noticeable. One example is increased references to 'policy networks.' Recently, there have been a number of similar terms that have also been gaining more popularity, including 'networked governance', 'collaborative network', 'inter-organizational network', and 'social capital/social network.' It is clear that the use of the term network is decidedly on the increase both in general social science and public administration.

Arguably, the most evident scholarly work that made us turn our attention to networks is O'Toole's paper published in 1997 calling for scholars of public administration and policy to "[treat] networks seriously." Since then, a growing number of researchers use the term network as if it were a rising fashion trend. A recent article in *Public Administration Review*, "Three Traditions of Network Research" by Berry et al (2004) illustrates this trend well. Rethemeyer (2005) stated that the theoretical approach of 'network management' has matured and is now subject to empirical examination. In fact, the entire social science discipline is experiencing a rapid increase in the number of studies employing the term 'network' in one way or another (Borgatti and Foster, 2003). This article will empirically examine influential authors in the discipline and patterns of research streams with the expectation of finding clusters or subgroups of authors and research under network research.

This article examines the trends and development of network research in public administration and policy literature by employing citation network analysis. Although citation analysis (e.g., citation index, citation network, co-citation network) can reproduce the history of the field, it is no substitute for extensive reading and in-depth content analysis. In that regard, we also supplement content analyses of abstracts and semantic network analysis for coded keywords as a complement to citation network analysis.

Background O'Toole's seminal paper in 1997 called for treating networks seriously in public administration. He stated:

> Networks are increasingly becoming important contexts for public administration and that networked settings are different in respects that matter for the conduct of administration. Public administration should attend to several types of network-focused research efforts. Some suggestions are: 1) Undertake systematic research to explore the descriptive questions on the network agenda. How much of managers' time, effort, and contingencies lie in or are devoted to network contexts? 2) Shift units and/or levels of analysis to the network. 3) Address both conceptual and theoretical agendas by identifying dimensions of network structure that may help to explain and mediate program and service delivery results… (1997, p. 50).

O'Toole expressed the idea that treating networks seriously had not been ignored so much as it had simply not been a priority in the world of public administration. O'Toole added that both administrators and researchers have begun to devote efforts to understand and study this theme. His work has stimulated further work in the area.

Upon our review of the literature, we found that a few scholars in public administration posed the very same set of questions a decade ago. Bogason and Toonen (1988) discuss the meaning of the concept of 'network' in relation to other conceptual developments in public

administration such as neo-institutionalism and neo-managerial analysis. They state that it is easy to predict that networks (their interdependency patterns, ways of non-hierarchical governance and conflict resolution) would be more important in the future with the trend of decentralization and devolution of governments. They contend that there is more need to link networks to other theories such as game theory, resource dependence theory, and communicative/discourse theory. Toonen (1998) then presented a meta-theoretical framework to encompass networks, management and institutions in public administration. He contends that the network concept is useful but does not present a sound basis for re-founding the study of public administration. He asks us to deal with the challenge of integrating institutional, managerial, and network concepts in the study of public administration.

Milward and Provan (1998) stated that the majority of network studies in public administration had been used as conceptual schemes or metaphors. They called for advancing measurements to clarify these concepts, using rigorous analytic measures (Provan & Milward, 2001; Provan, Veazie, Staten, & Teufel-Shone, 2005). Borgatti (2006) argues that many theories are rooted in relational thinking and point us in the direction that some theories share common roots. Similarly, Wellman (1998) argued that a network is a perspective or worldview to perceive social problems and research questions, rather than just an analytic tools or metaphor.

Berry et al (2004, p. 549) identified three parallel streams of literature about network theory and research: social network analysis, policy change/political science networks, and public management networks. In so doing, they offered recommendations for advancing current scholarship on public management regarding network research, including providing a social network analysis tradition for those who focus on public management networks. They state that there has been an abundance of network research since O'Toole's seminal article in 1997 and called for "a variety of methods for studying public management networks and cultivating discussion among those who employ different methods or whose work is guided by different theoretical orientations, including the value added by social network analysis. (p.549)"

At the advent of the ten-year anniversary of O'Toole's work, Robinson (2006, p. 589) claimed that the literature was clearly treating networks seriously. He stated that we are now past the need for demonstrations of the prevalence of networks and that it is now time to examine the origins, effects, and diversity of networks in public policy implementation and network governance. He suggests future research should investigate the diversity of networks, the relationships between the different types of collaborations and goes on to call for methodological pluralism and innovation to pursue this future research direction.

Rethemeyer and Hatmaker (2008) submit that there are four perspectives and two process models on networks and network management studies, but that there is no integration across them. They start with the common theme that "network management is unlike the management of hierarchies because it occurs outside the usual rational-legal basis for authority (p. 630)"

Their description of the four perspectives is:

1) Interest intermediation: in this school of thought the task of network management is reaching goal consensus, which restricts network management to the realm of policy networks.
2) Tools of government: to view network management as primarily a tool of implementation and collaboration while leaving aside the question of goal formation.
3) Focus on the information processing and knowledge management capabilities of networks.
4) Governance: taking seriously the idea that decision and implementation are not neatly divided (p.631).

They explain that the challenges of network research partly stem from the dual nature of any network: networks are both cause and effect. Therefore, some scholars focus on how managers can change 'action in network'; others focus on 'networks of action.' The two process models they describe are:

1) The Games-network approach.
2) POSDCORB (Planning, Organizing, Staffing, Directing, Coordinating, Reporting, and Budgeting) of the network era - four network management processes: activation/deactivation, synthesizing, framing, and mobilizing (Rethemeyer and Hatmaker 2008 p.632).

Rethemeyer (2005) also stated researchers should employ network measures and methods to theorize various network studies. Dowding (1995) critiqued policy network studies being metaphorical rather than theoretical. He argued:

Whilst we have learned much about the policy process by cataloguing the policy world into different types of network, the approach will not, alone, take us much further. Policy network analysis began as a metaphor, and may only become a theory by developing along the lines of sociological network analysis. In order to produce a network theory where the properties of the network rather than the properties of its members drives explanation, political science must utilize the sociological network tradition, borrowing and modifying its algebraic methods (pp. 136-137).

Hwang (2009) argues that the term 'network' means different things to different disciplines, this being related to the stages of network research in each discipline in terms of its maturity. He contends that network research grows idiosyncratically from metaphor to method, theory, and paradigm.

Hummon and Carley (1993) used citation network study for network research growth, particularly studying the citation pattern of SNA to gauge its advancement toward Kuhn's sense of normal science. They examined the patterns in the citation network and found a high density of multiple citations, both to articles within a given journal and to key articles outside the journal, and many authors who have published more than a single article in the journal. They concluded that the overall citation pattern is consistent with a pattern of scientific development labeled by Kuhn as normal science. Basically, they looked at the evolution of citation networks over time and claimed that the field was moving to Kunh's sense of normal science as citation network gets dense.

It is not a coincidence that different themes and theories in public administration, such as inter-organizational relations, neo-institutionalism, collaborative management and governance, share a common thread. Indeed, Bogason (2006) traces network analysis as developed in policy network literature in the 1970's, and discusses the status of network analysis in relation to the themes of public administration. He shows indirectly that groups of scholars over time have developed themes of network analysis in public administration under different names like inter-organizational relations, institutionalism, and governance.

The primary goal of this article is to continue this line of work by adding an empirical examination of network research studies through citation analysis. In doing so, we intend to identify or confirm influential scholars, their research and provide a visualization of the impact of scholarly research in this important and growing field. Our intention is not to function as a substitution for extensive reading and fine-grained content analysis such as Berry et al.'s (2004) work, but rather to confirm influential authors and demonstrate their impact on the citation network.

The premise of our study is in line with existing scholarship of citation studies, in which citations serve as a measure of the impact of that work (Garfield, 1992) and co-citation is used to map the intellectual structure of scientific disciplines (Bayer, Smart, & McLaughlin, 1990). Citation

analysis showcases highly cited scholarly manuscripts in order to measure their impact and track any emerging trends. The science citation index was proposed over fifty years ago and citation impact factors have recently begun to be treated as a proxy evaluation system for published articles (Garfield, 2006, 2007).

Whether citation impact factor accurately estimates citation frequencies and importance is controversial. Scholars in the information science domain have worked extensively in analyzing and visualizing citation index data (Chen, 2006; Garfield, 1992; Rousseau & Zuccala, 2004; Schwartz & Fang, 2007; Small, 1999; White & Mccain, 1998). Scholars in the network analysis domain have also contributed heavily to analyzing citation patterns and co-author collaboration (Batagelj & Mrvar, 2008; Lazer, Mergel, & Friedman, 2009; Leydesdorff, 2007; White, Wellman, & Nazer, 2004). It has become abundantly apparent that examining citation data can greatly augment our understanding of how a given study domain has progressed. As a professional norm, scholars pay tribute to the existing body of knowledge by citing them. We believe studying citation patterns using citation network analysis is a good measure of the impact of scholarly research

Certainly, citation analysis and citation network analysis have limitations and biases. Yet, we contend it is worthwhile introducing them to the readers of public administration studies because all of the public administration journals in the Social Science Citation Index (SSCI) advertise journals' impact factors as journals' authority-building and marketing. Also, peer reviewed articles are considered a very important part of scholarly achievement for tenure review, at least in the U.S. Thus, we believe citation analysis and citation network analysis have some values. We also believe citation network analysis has not been utilized in public administration scholarship so far, which makes this study useful.

Research Questions

In undertaking this research, we set out to examine three related research questions:

RQ1: What is the impact of O'Toole's work?

RQ2: What is the current status of network research in public administration and policy?

RQ3: What are the sub-topics in network research in public administration and policy?

METHODS

Data

Data for this research were acquired from the Social Science Citation Index (SSCI), ISI Web of Science. Publications were drawn from all journals in the SSCI that were published within the subject category of public administration. Publication dates were limited to the past decade. Such data allows for an empirical assessment of how the term network has been employed in research articles within the fields of public administration and public policy. There is no Public Policy classification but the Public Administration subject does include journals of Public Policy, including JPAM (Journal of Policy Analysis and Management). There were 26 journals in this category In February 2007, we did a search for the term 'network' in 26 public administration journals on SSCI, in the fields of 'abstract', 'keyword', and 'title'. This search returned 257 articles. Web interface is not particularly useful in data mining. We saved the data as xml and endnote format files but ultimately transformed into xml, with which we were able to extract information by computer programming so that the data is suitable for network analysis. Additionally, it should be noted that this citation network is limited to the 257 articles that were identified through the SSCI. We recognize that it is possible that articles that may be related to the subject, but do not have 'network' as a keyword may have been

excluded. That is, we may have inadvertently excluded other journals.

As briefly stated before, using SSCI data has limitations and biases. It reflects much of the scholarship in the U.S. but not scholarship globally. The chosen language is English; thus the data excludes other foreign language publications. Moreover, we know there is a tradition that many European scholars, unlike scholars in the U.S., publish their research in books, reports, and journals that are not in SSCI. Thus, this data does not present us with a whole picture. However, we argue that it is still meaningful because this data shows us the picture of scholarship in the U.S. and will provide a good starting point to expand in the future.

Co-Citation Networks

Social Network Analysis (SNA, network analysis) was employed using the software package, ORA (2009). Network analysis provides a description of citation patterns. The structure of citation patterns (citing, and cited) among scholarly works that involve network research were examined in this way to identify structures in social systems on the basis of the relations among the system's components rather than the attributes of individual cases (Wasserman & Faust, 1994). Using the 257 articles that were identified using the SSCI, we constructed a list of authors and a corresponding list of journals where their articles were published. Next, the citation relationships between authors and journals were determined. Resulting article-to-article node-sets display their own citation networks. For example, if *Article A* cites *Article B*, then the authors of *Article A* are citing those of *Article B*. In this procedure, each author of *Article A* has citation links to all of the authors of *Article B*. To discover the journal-to-journal citation relationships, we created a link from *Journal A* to *Journal B* if *Article A* was published in *Journal A* and *Article B* in *Journal B*. The direction of the citation links was determined as originating from a cited article, author, and journal and directed towards citing entities. While article-to-article citations offer no intuitive method for weighting links, author-to-author citations and journal-to-journal citations can be weighted according to citation frequency. This is because an author, or a journal, may cite a number of articles written by the same author, or journal.

Entities were clustered in accordance with the Newman-Girvan (Newman 2004) grouping algorithm which identifies clusters by disconnecting high edge-betweenness links and creating components from the disconnections. Lastly, we added a semantic network analysis (using Automap) and a complementary qualitative analysis. We coded keywords from titles, keywords, and abstracts of the 257 articles to use the frequency as a corroborating technique to the citation network analysis (using Atlas Ti). Semantic network displays the distance and grouping of the keywords in addition to their frequencies.

RESULTS

As discussed earlier, scholars in network science have already documented the rapid increase of network research in social science and beyond (Borgatti & Foster, 2003; Freeman, 2004, 2008). Over the past years, we have witnessed an increase in the use of network analysis in scholarly research in the field of public administration and policy, but the volume of network research in other fields is similarly expanding and well worth noting. When our search for the term network was conducted in SSCI, the management subject returned 314 articles in 2007, up from 115 in 1992; information science increased to 134 from 62; sociology increased to 98 from 33; economics increased to 167 from 45, and political science increased to 62 from 13.

A search of the SSCI from 1992-2007 on the frequency of published articles by author in SSCI shows that (Table 1) Klijin published the most articles (9), followed by Provan (6).

Table 1. Frequency of Publications by Author During 1992-2007 (Minimum Count Of 2)

No/Rank	Author	Frequency
1	KLIJN, EH	8
2	PROVAN, KG	6
3	CONSIDINE, M	4
3	MEIER, KJ	4
3	O'TOOLE, LJ	4
6	HARMAN, R	3
6	LEWIS, JM	3
6	SKELCHER, C	3
7	14 other authors	2

Table 2. Top 10 Central Authors: Centrality of authors in the citation network (OutDegree: Cited by Others)

No/Rank	Node Title (Author)	Centrality / Author-To-Author
1	O'Toole	0.2205
2	Klijn	0.0816
2	Meier	0.0816
4	Milward	0.0574
5	Provan	0.0544
6	McGuire	0.0453
7	Agranoff	0.0393
8	Borzel	0.0272
9	Bogason	0.0211
9	Toonen	0.0211

An initial scan of the data shows O'Toole's 1997 article as the most cited work and O'Toole as an influential scholar in this network of policy and administrative network research (see table 2 & 3).

Table 3. Top 17 Central Papers: Centrality of Papers in the Citation Network (OutDegree: Cited by Others)

No/ Rank	Node Title (Paper)	Centrality / Paper-to-Paper
1	O'Toole, 1997(b)	0.1484
2	Meier, O'Toole, 2001	0.0352
3	Klijn, 1996	0.0352
4	Provan, Milward, 2001	0.0313
5	Agranoff, McGuire, 2001(a)	0.0273
6	Meier, O'Toole, 2003	0.0234
7	Borzel, 1998	0.0234
8	Bogason, Toonen, 1998	0.0195
9	Lowndes, Skelcher, 1998	0.0156
10	Blom-Hansen, 1997	0.0156

Figure 1. Article-to-Article Citation Network 1992-2007

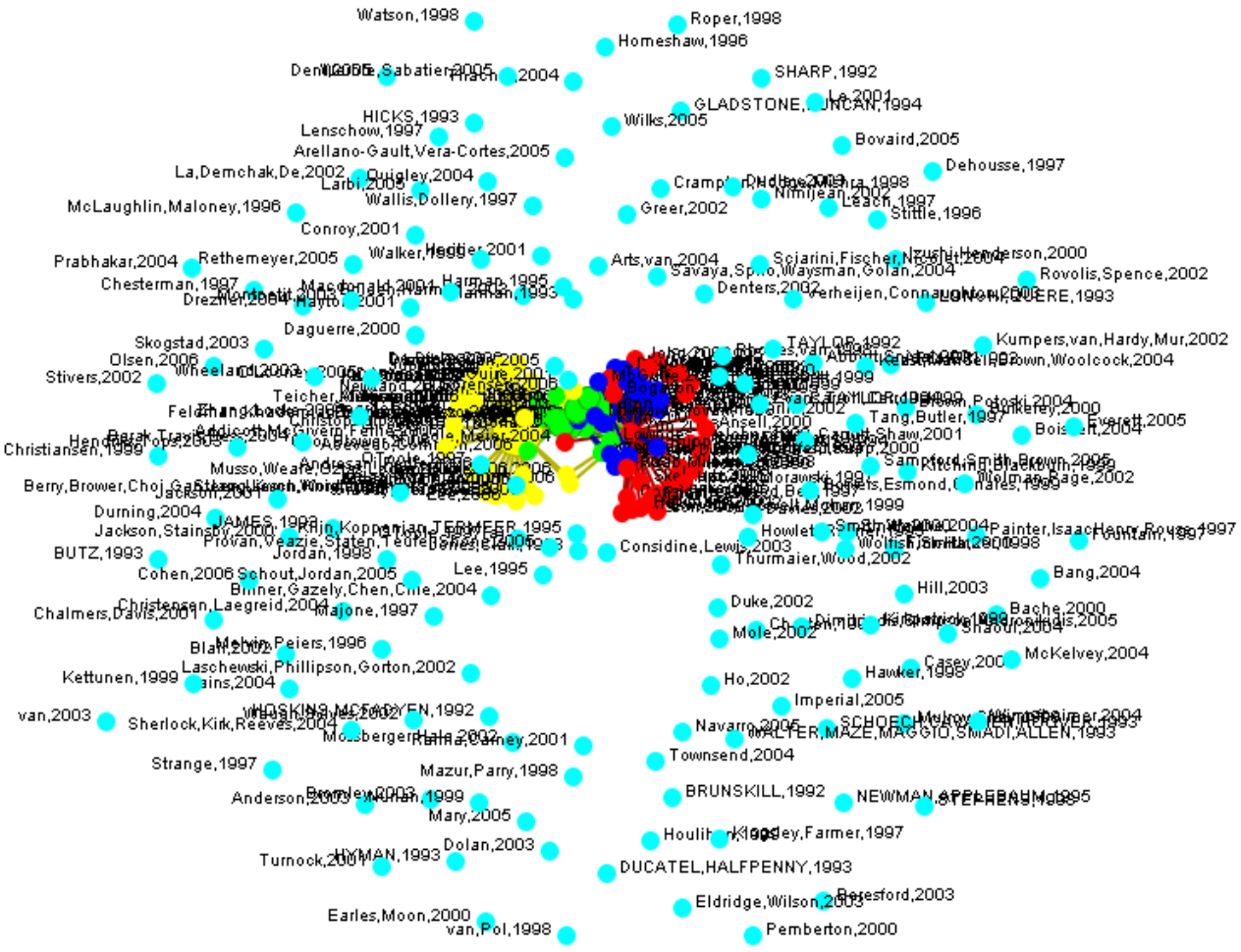


Figure 1 shows the metamatrix of article-to-articles citation networks.

Figure 1 and 2 display the entire network of article-to-article citations. A sizable number of isolates are apparent in addition to a densely connected citation network consisting of four subgroups. This pattern suggests that we may be experiencing a paradigm shift (Kuhn, 1970) within the field. According to Hummon and Carley (1993)'s reasoning, a dense citation network represents a matured domain or school of thought, which can be interpreted as a normal science paradigm.

Figure 2. Paper-to-Paper Citation Network 1992-2007

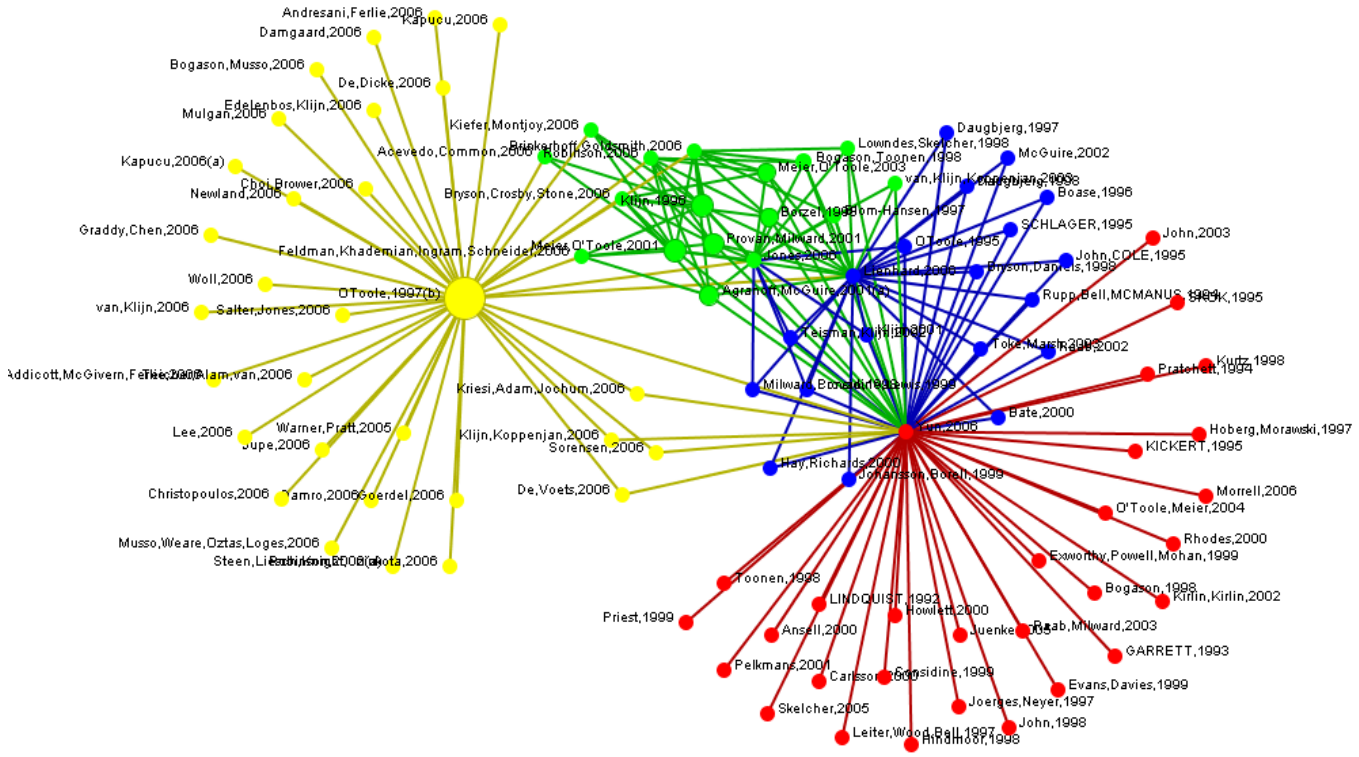


Figure 2 shows the metamatrix of citation network after excluding isolates.

Figure 2 displays the article-to-article citation network after excluding isolates. With Newman-Girvan grouping, there are four subgroups. One group is clustered around O'Toole's 1997 work. Another group is clustered around the works of Provan, Milward, Agranoff and McGuire. The other two subgroups do not show a node with strong centrality and we cannot point out major works in those subgroups. Although we would need to thoroughly examine these articles in order to formally categorize them, these findings appear to corroborate Rethemeyer and Hatmaker's suggestion that there are policy networks and collaborative networks. The presence of two groups of collaborative networks/management also seems to agree with their two process models.

Klijin's works are best known for studying policy networks. This partially explains the gap between frequency and centrality analyzed here. As we stated, therefore, we decided to look at the keywords to further our inquiry. As a complementary approach, qualitative coding to extract keywords and semantic network analysis was conducted. We read titles, keywords, and abstracts of 217 articles to investigate development trends and key topics.

Table 4 displays the results of key word frequencies that were generated through qualitative coding analysis. Not surprisingly, 'policy network' appeared most. The term 'governance' appeared equally often, reflecting the shift to the governance paradigm in public administration literature. 'Partnership' and 'collaboration' were next. 'Emergency management' appeared frequently as well. Policy network has a long history in the policy study domain. As discussed earlier, many studies use the term 'network' metaphorically. We believe this contributed to the high frequency in this table. Scholars studying governance and collaboration see the importance and merit of using networks in studying a new governance paradigm where public administrators work across the sectors and judicial boundaries.

Figure 3 illustrates the semantic network that was generated using Automap. Management, performance, governance, and collaboration were the keywords with the highest centrality[1]. The figure shows a group of keywords with governance, partnership, and collaboration. This, we believe, represents a group of scholars studying collaborative governance, inter-sectoral collaboration, and inter-organizational partnership. Another grouping was management, performance, local government, contracting, network-management, and network-structure. This group, we believe, represents scholars interested in looking at the performance and effectiveness of management through the lens of network. Another grouping was policy-network, policy-change, policy-making, and trust. This group represents long-standing policy network related studies. The proximity of themes of studies to each other is clear. Together, Table 4 and Figure 3 illustrate subgroups of scholarly works within the network research discussed above.

Table 4. Qualitative Coding Analysis: Keywords Frequencies

CODES	Frequency
Policy network	31
Governance	31
Partnership	12
Collaboration	9
Innovation	8
Emergency management	8
Institutions	7
Policy implementation	7
Health policy	7
Social welfare	7
Leadership	5
Contracting	5
Policy change	5
Homeland security/terror	5
Democratization/democracy	5
Stakeholder analysis	5
Organizational learn	4
Decision-making proc	4
Information technology	4
Managerial networking	3
Knowledge diffusion	3
Public sector reform	3
Social capital	2

[1] Centrality is a concept and measure in social network analysis. Simply put, it is a measure of an importance of a node in a network.

Figure 3. Consolidated Semantic Network

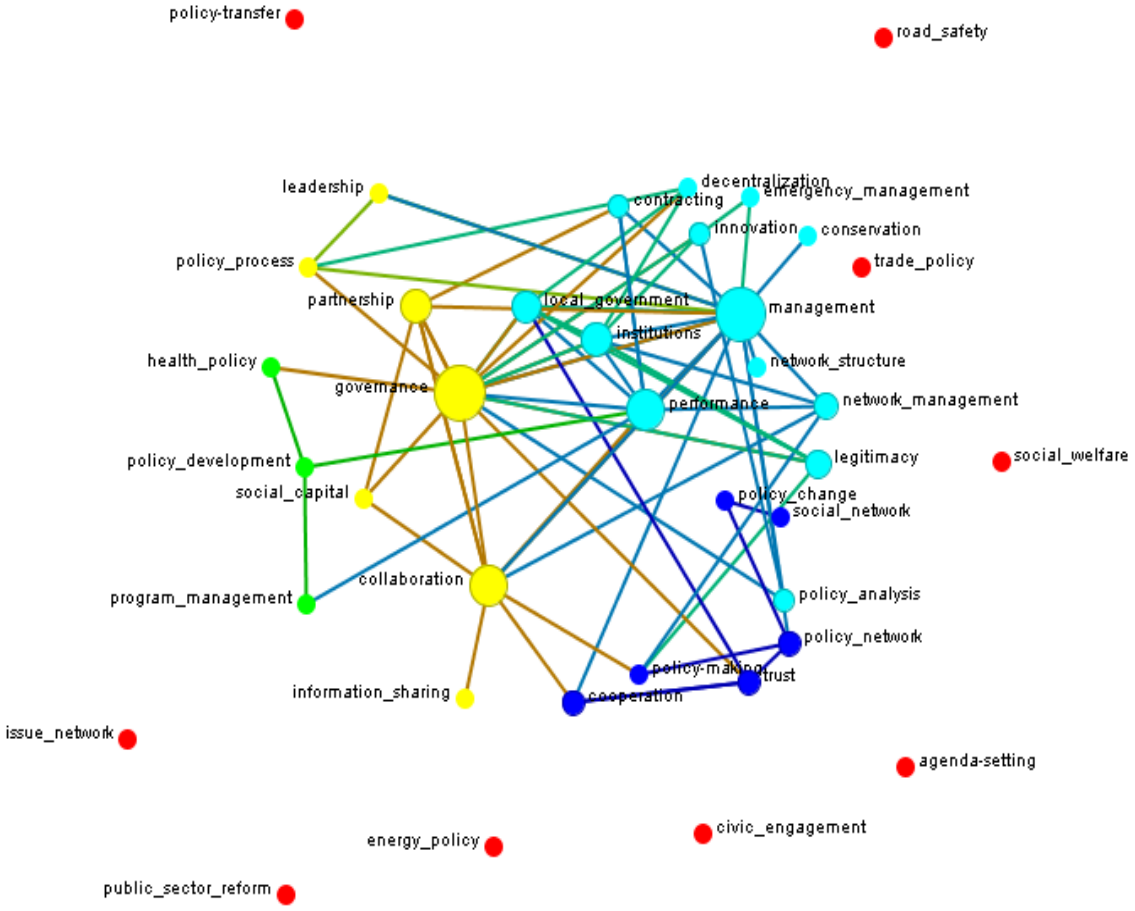


Figure 3 shows the semantic network of keywords that appear in the articles.

DISCUSSION & CONCLUSIONS

We believe that the above analyses offer an empirical confirmation of O'Toole's impact on network research in public administration and policy. Increased cross-citation over time (citation network maps over time) and a variety of keywords show that network research is growing rapidly. Subgroups of citation networks visually and empirically illustrate the work of Berry et al. (2004) and Rethemeyer and Hatmaker (2008) in demonstrating that there are different perspectives among network research in public administration and policy. Overall, the analysis of network results captured some of the many different perspectives and traditions now incorporated as part of network research. We recommend further qualitative analyses to investigate the nature of those differences and investigate how each subgroup is using network measures.

Citation patterns for the entire network do not seem to be dense enough to constitute Kuhn's sense of normal science according to Hummon and Carley's (1993) definition, but the network appears to be progressing in that direction. It seems that the public administration discipline is at the beginning of a growth stage in network research. It is therefore imperative that we re-visit this issue in the future.

In conclusion, we echo Rethemeyer's call for an empirical examination of the network approach in public management, particularly given the increase in the number of works discussing collaborative public management, collaborative governance, networked governance, and other similar themes. We also concur with Dowding in his critique of the contemporary use of the network approach as being too metaphorical. So far, the majority of network research seen in public administration is metaphorical or conceptual. Our claim is that until now, it has been appropriate to examine networks conceptually. We believe conceptual studies have contributed to the field in advancing the call for more empirical works. However, we argue that we are at a critical juncture as we need to move beyond the conceptual or metaphorical use of "network", particularly after a decade of network research growth. In other words, we believe that metaphorical usage could advance to development of methods and theories.

What is network research? Are we seeing the development of integrated network theory in public administration? A single definitive answer may not be possible or even desirable. But we believe that there is value and utility in searching for it, perhaps generating multiple sets of well-defined theories particularly suitable for various sub-domains such as policy networks or collaborative management. In doing so, we would improve sets of analytic tools, measures, and data collection instruments that are suited to public administration and policy.

So, are we treating networks seriously? We would like to provide two provisional answers here. Yes, there has been a great increase of network research answering O'Toole's call to treat network seriously. And no, there is not enough serious network research and, in that it is time for us to move beyond the metaphorical use of network in public administration research, we need more. There is a need for more empirical works on network research, particularly employing rigorous network analysis approaches. We should advance the measures of network research, both adopting and developing dyadic measures and analytic tools such as blockmodeling and logistic network regressions (e.g.: Quadratic Assignment Procedure). We also need to establish a solid body of network data collection instruments (survey instruments, interview protocols, etc). With few exceptions, including Provan and Milward's works, tested and reliable data collection instruments embedded in public administration and policy do not exist at this stage. With these advances, we can move toward building an integrated and comprehensive network theory in public administration. One notable exception in addressing methods, including challenges to measures (and collect) data, is the recent book edited by (Bogason & Zølner, 2007). They illustrate why we need to discuss methodology in studying network governance and discuss challenges and approaches in collecting data and developing empirical methods.

The data we extracted from SSCI has limitations. We know that the data on which we have focused may not have precisely captured influential works by particular scholars, such as Provan and Milward, in both the public administration field and the management science domain. This investigation should also be expanded into other journals and subjects in order to investigate the interdisciplinary citation patterns to detect influential works from outside of public administration to the network research in public administration. The data is also somewhat U.S. scholarship-oriented, as European scholars publish beyond these journals in SSCI more than U.S. scholars. Yet, we believe it captures public administration in the U.S. well as well as some influential scholars from Europe, and it sets a stage for future

research. We call for expanding this analysis, possibly incorporating Google scholar and other databases to collect data that we were unable to get with SSCI.

REFERENCES

Batagelj, V., & Mrvar, A. (2008). *Networks from Web of Science: Social Networks*. Paper presented at the Sunbelt Social Networks Conference.

Bayer, A. E., Smart, J. C., & McLaughlin, G. W. (1990). Mapping intellectual structure of a scientific subfield through author cocitations. *Journal of the American Society for Information Science, 41*(6), 444-452.

Berry, F. S., Brower, R. S., Choi, S. O., Goa, W. X., Jang, H., Kwon, M., et al. (2004). Three Traditions of Network Research: What the Public Management Research Agenda Can Learn from Other Research Communities. *Public Administration Review, 64*(5), 539-552.

Bogason, P. (2006). Networks and bargaining in policy analysis. In B. G. Peters & J. Pierre (Eds.), *Handbook of public policy* (pp. 97-113). London ; Thousand Oaks, CA: Sage Publications.

Bogason, P., & Toonen, T. A. J. (1988). Introduction: Networks in public administration. *Public Administration, 76*(2), 205-227.

Bogason, P., & Zølner, M. (2007). *Methods in democratic network governance*. Houndmills, Basingtoke, Hampshire ; New York: Palgrave Macmillan.

Borgatti, S. P. (2006). *What is network theory?* Paper presented at the International Sunbelt Social Network Conference, INSNA

Borgatti, S. P., & Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. *Journal of Management, 29*(6), 991-1013.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359-377.

Dowding, K. (1995). Model or Metaphor? A Critical Review of the Policy Network Approach. *Political Studies, XLIII*, 136-158.

Freeman, L. C. (2004). *The Development of Social Network Analysis*. Vancouver: Empirical Press

Freeman, L. C. (2008). Going the Wrong Way on a One-Way Street: Centrality in Physics and Biology. *Journal of Social Structure, 9*(2).

Garfield, E. (1992). Citation Data: Their use as quantitative indicators for science and technology evaluation and policy making. *Science and Public Policy, 19*(5), 321-327.

Garfield, E. (2006). Commentary: Fifty Years of Citation Indexing. *International Journal of Epidemiology, 35*(5), 1127-1128.

Garfield, E. (2007). The Evolution of the Science Citation Index [Electronic Version]. *Perspectives: International Microbiology*, 10, 65-69.

Hummon, N. P., & Carley, K. (1993). Social networks as normal science. *Social Networks, 15*(1), 71-106.

Hwang, S. (2009). Past, Present, and Future of Social Network Analysis: Network as a Metaphor, Method, Theory, or Paradigm? *International Journal of Interdisciplinary Social Sciences 4 (4)*

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*: University of Chicago Press.

Lazer, D., Mergel, I., & Friedman, A. (2009). Co-citation of prominent social network articles in sociology journals: The evolving canon. *Connections, 29*(1), 43-64.

Leydesdorff, L. (2007). Visualization of the citation impact environments of scientific journals: An online mapping exercise. *Journal of the American Society for Information Science and Technology, 58*(1), 25-38.

Milward, H. B., & Provan, K. G. (1998). Measuring network structure. *Public Administration 76*(2), 387-407.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E, 69*(6), 066133.

O'Toole, L. J., Jr. (1997). Treating Networks Seriously: Practical and Research-Based Agendas in Public Administration. *Public Administration Review, 57*(1), 45-52.

Provan, K. G., & Milward, H. B. (2001). Do Networks Really Work? A Framework for Evaluating Public-Sector Organizational Networks. *Public Administration Review, 61*(4), 414-423.

Provan, K. G., Veazie, M. A., Staten, L. K., & Teufel-Shone, N. I. (2005). The Use of Network Analysis to Strengthen Community Partnerships. *Public Administration Review, 65*(5), 603-613.

Rethemeyer, R. K. (2005). Conceptualizing and Measuring Collaborative Networks. *Public Administration Review, 65*(1), 117-121.

Rethemeyer, R. K., & Hatmaker, D. M. (2008). Network Management Reconsidered: An Inquiry into Management of Network Structures in Public Sector Service Provision. *Journal of Public Administration Research and Theory, 18*(4), 617-646.

Robinson, S. E. (2006). A Decade of Treating Networks Seriously. *Policy Studies Journal, 34*(4), 589-598.

Rousseau, R., & Zuccala, A. (2004). A classification of author co-citations: definitions and search strategies. *Journal of the American Society for Information Science and Technology, 55*(6), 513-529.

Schwartz, F., & Fang, Y. C. (2007). Citation Data Analysis on Hydrogeology. *Journal of the American Society for Information Science and Technology, 58*(4), 518-525.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science and Technology, 50*(9), 799-813.

Toonen, T. A. J. (1998). Networks, management and institutions: Public administration as 'normal science'. *Public Administration, 76*(2), 229-252.

Wasserman, S., & Faust, K. (1994). *Social network analysis : methods and applications*. Cambridge ; New York: Cambridge University Press.

Wellman, B. (1998). Structural analysis: From method and metaphor to theory and substance. In B. Wellman & S. D. Berkowitz (Eds.), *Social Structures: A Network Approach* (pp. 19-61). Cambridge: Cambridge University Press.

White, H. D., & Mccain, K. W. (1998). Visualizing a Discipline: An Author Co-Citation Analysis of Information Science *Journal of the American Society for Information Science and Technology, 49*(4), 327-355.

White, H. D., Wellman, B., & Nazer, N. (2004). Does citation reflect social structure?: longitudinal evidence from the "Globenet" interdisciplinary research group. *Journal of the American Society for Information Science and Technology, 55*(2), 111-126.

The Structure of Undergraduate Association Networks: A Quantitative Ethnography

Sean A.P. Clouston
Department of Sociology, McGill University/Mailman School of Public Health, Columbia University

Ashton Verdery
Department of Sociology, University of North Carolina

Sara Amin
Department of Sociology, McGill University/Georgetown University

G. Robin Gauthier
Department of Sociology, Duke University

The challenge of collecting complete associational networks has restricted network studies to small datasets. To deal with larger processes, two general procedures have been developed: the use of indicators such as citation structures or the diffusion of innovations to model human interactions, and limiting the sample of associates' names. A body of theoretical and empirical work has identified several problems with these methods. We examine a unique solution to these problems—measuring online social networks of college students. In this paper we present an original network dataset of undergraduate Facebook users and demonstrate the feasibility and acceptability of this form of measurement. We conclude with a preliminary exploration of Network Homophily and Multiplexity on Facebook.

Authors: *Sean Clouston is a Ph.D. candidate in Sociology at McGill University in Montreal and a 2008-09 Canada/US Fulbright scholar at the Center for the Study of Social Inequalities in Health at the Mailman School of Public Health in New York. His research focuses on the Social Inequalities in Global Health, Social Networks, Life Course Analysis, Family Dynamics and Public Policy.*

Ashton Verdery is a Ph.D. student in Sociology at the University of North Carolina where he works on understanding the importance of social networks to immigration, both at the place of arrival and of origin. Ashton is affiliated with the Carolina Population Center, and his areas of interest are in Social Networks, Social Demography, Migration, Quantitative Methods, and Human Geography.

Sara Nuzhat Amin is a Ph.D. candidate in Sociology at McGill University, a 2007-08 Canada/US Fulbright scholar at Georgetown University, and teaches at the Asian University for Women in Chittagong, Bangladesh. She is currently completing her dissertation on the conversations and debates in the Canadian and American Muslim leadership and faith. She specializes in Political Sociology, Social Networks, and Quantitative Methods.

G. Robin Gauthier is a Ph.D. student in Sociology at Duke University. Her research interests focus on Social Networks, Methodology, and Social Structure. She is the Graduate Associate Editor at the Journal on Social Structure.

Acknowledgments: *The authors would like to acknowledge Jack Sandberg and Steven Rytina, without whom this paper would never have been designed. We would also like to thank our supervisors and institutions for making this project possible.*

Correspondence: *Contact Sean Clouston,, Department of Sociology, McGill University, Room 713, Leacock building, 855 Sherbrooke St. West, Montréal, QC, Canada, H3A 2T7; email: sean.clouston@mcgill.ca*

INTRODUCTION

The collection of all associates of a given ego requires some form of name generator. The creation of name generators give rise to two main issues, one methodological, the other conceptual, which must be considered in any attempt to gather network data. The conceptual issue requires one to specify the "cognitive principle" (Degenne and Forsée, 1999) that underpins the study: e.g., family ties, spatial proximity, high school friends, or previous contacts. The main methodological issue relates to operationalizing the cognitive principle into a list that is both accurate and comprehensive in a meaningful and useful way.

Studies that have explored the structural qualities of associational ties of individuals in mass societies have often had to rely on resource intensive procedures to gather complete associational networks, which are often subject to errors of memory and measurement (Degenne and Forsé, 1999). Other kinds of studies have relied on gathering meaningful data by restricting the number of associates that can be named by certain criterion, ranging from arbitrary to hierarchical ordering and exclusion. These methods too are subject to problems of respondent error. Moreover, since network measures are sensitive to changes in their numbers of nodes/edges (Wasserman and Faust, 1997), it is hard to describe to what extent associational network studies limited to only 5 alters are able to reflect actual structures of relations.

Diaries are probably the best name generators, yet they require a great deal of time investment from respondents (Degenne and Forsée, 1999; Marsden, 1990). Moreover, they still suffer from the problem of having to either gather information about associates through multiplicative interviews or rely on perceptions of egos of their contacts' relations (Marsden, 1990). We suggest that social networking sites such as Facebook be considered as active diaries, while solving both problems of diaries: they record an ego's alters as symmetric ties (since both ego and alter have to accept each other as "friends") and they allow us to gather information about both alter and the alter's ties (the alter's "Friends") from their profiles in a way similar to the original sample ("egos"). We note that, unlike traditional diaries, Facebook lists do not allow us to identify the specific role of a tie: friend, girlfriend, sister, classmate and so on. "Friends" in Facebook could occupy any of these roles. While Facebook users have the option to identify the basis of a given friendship (high school, work, class, family, relationship), this tool is problematic for data collection because: (1) users frequently do not use it and (2) when users do use it, they often forge stories for reasons of hilarity, impression management, and so on. Therefore, the cognitive principle underlying datasets created from Facebook lists is limited to a broad one: "friends" on an actor's Facebook profile.

Finally, the recording of activity between alter and ego is possible as a measure of how frequently contact is made between contacts (through records of "Wall posts" and other forms of dyadic activity possible in Facebook), without the required effort and potential recordkeeping errors of a respondent's diary. While researchers have explored time use of users as various social indicators (strength of tie, use of social capital, investment of and in relations), for our purposes we are more interested in using the Facebook lists and information as a name generator indicating ties and the attributes of those connected.

The main problem with lists generated off social network sites is in wondering what kinds of ties are actually being captured: or more specifically, do these ties have any correspondence to offline ties, and if so to what extent and how? Moreover, since online sites are visible representations of networks, is there a visualization effect on lists? An emerging body of studies provides some insight into these challenges of utilizing Facebook as a name generator (Hogan, 2008).

Research on Facebook

Social network sites (SNSs)[1] such as MySpace, Facebook, Cyworld, and Bebo, are populated by millions of users, a large number of whom have incorporated SNSs into their daily practices (Boyd and Ellison, 2007). The growth of sites and users have attracted scholars from a wide range of backgrounds researching user practices and engagement, the consequences and ramifications of SNS growth and structure, and the development of cultures and sub-cultures in SNSs.

Specifically, we can identify four overarching trends in Facebook studies. Firstly, Facebook friendships are articulated on "latent ties" (Haythornwaite, 2005) sharing offline connection *prior to* online meetings. While in certain social network sites, participants engage in 'networking' to meet new people, users on Facebook utilize Facebook to maintain offline friendships (Boyd and Ellison, 2007). The most common uses of Facebook are to maintain previous high school relationships and to gain information about offline contacts (Boyd and Ellison, 2007; Lampe *et al.,* 2007). However, it was also found that students identifying as ethnic minorities showed that those who were non-white were significantly less able to make high school or strong bonds than were whites (Ellison *et al.,* 2006).

Secondly, studies comparing social capital and integration of users and non-users have found that non-Facebook users have fewer offline contacts than Facebook users both in intensity of the relationships and in frequency on face-to-face contact. Moreover, there is a positive relationship between a student's perception of integration into their university community and both the intensity of Facebook use and the number of their Facebook friends. Facebook may therefore have the capacity to convert latent ties between users into weak ties (Ellison *et al.,* 2007).

Thirdly, while SNSs are designed on the premise of wide accessibility, researchers have found that groups often use sites in ways which manifest segmentation by nationality, age, educational level, or other stratification axes common in society (Hargittai, 2007). In particular, Facebook researchers have found that networks on the site show network homophily. Facebook networks exhibit ethnic homophily, especially among students identifying as white, while ethnic minorities have more heterogeneous friendship networks (Ellison *et al.,* 2006).

Finally, the degree of information disclosure and relative openness of the network has raised concerns of user privacy (Gross and Acqusti, 2005). In Facebook, users who are part of a common network may view each others' complete profiles, unless the user has specifically chosen to deny permission.[2] In any

[1] We follow the definition of SNSs provided in the review of Boyd and Ellison (2007). They define social network sites "as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. While we use the term "social network site" to describe this phenomenon, the term "social networking sites" also appears in public discourse, and the two terms are often used interchangeably (Section on Social Network Sites: A Definition)." They choose not to use "networking" since it implicates a functionality that varies between and within sites, and across users, which is not inevitable or prescribed.

[2] Privacy settings in Facebook are unique relative to other SNS's. Originally Facebook was created only for college students who were required to have a valid institutional e-mail address to become members of the college network. Between September 2005 and September 2006, Facebook expanded to include professionals in corporate networks, high school students, and finally everyone. While regional networks (Montreal, Chile, etc) impose no rules about membership, access to closed networks remain relatively restricted: administrator approval is

given analysis the researcher has to decide carefully what the relationship may be between non-respondents (those with high privacy settings) and outcome variables. In Facebook, privacy patterns are themselves interesting outcomes, and as we will see in our analyses are an important source of bias in producing representative (offline) network maps. One classical solution would be to over-sample public profiles on those sharing the given attribute that is underrepresented (visible minority, gender, etc); however, any analyses with such methods, as in classical studies, must remain wary of possible qualitative differences between those who respond (share a public profile) and those who do not. We explored patterns of privacy through an ordered logit analysis of the privacy settings of our original sample at two different time points.[3]

In short, prior research on Facebook has focused on the way that ties on Facebook complement, compete or substitute for offline ties. The main finding has been that Facebook is utilized for 'social searching' (keeping in touch with those already known or searching for people already sharing an actual connection offline) rather than 'social browsing' (to meet new unknown people, such as sexual partners, online to create friends offline). We propose that Facebook ties strongly reflect offline ties and therefore could be used as a name generator of associational data of users.

In this paper we put our proposition to the test by creating an original associational network dataset from Facebook users of an undergraduate university.

required to gain access to high school networks; the appropriate '.com' address is required for access to corporate networks. Moreover, unlike other SNS's there is no way for users to make their profiles public to all users (Boyd and Ellison 2007).

[3] Results not shown here.

METHODS

Data Collection

We follow the well trodden paths of network analysts utilizing random samples to observe relations among sample subjects and drawing inferences about the population (Frank 1981; Granovetter, 1976). Using the random search tool in Facebook, in January 2007, we sampled McGill University's undergraduate Facebook population at random and stratified by faculty/school for representativeness into overarching faculties: Arts, Sciences, Engineering, Management, Education, and Other. We continued sampling until a quota was filled for each faculty/school before stopping. This yielded an original sample of 257 undergraduate users of Facebook. Of these, 37 had profiles that we were unable to view due to security restrictions made by the users themselves. We find this analytically equivalent to non-response in traditional survey methods. Other non-Response was due to misclassification (10) and non-reliable accounts (21), separately by three different coders and which were removed before data collection began; thus the response rate was 73.54 %, yielding a final sample of 189 users. We also note that the total population of Facebook users in this university is between 16,000 and 20,000 and the average individual network size is 175.8, with 77.1 links within the university network itself.

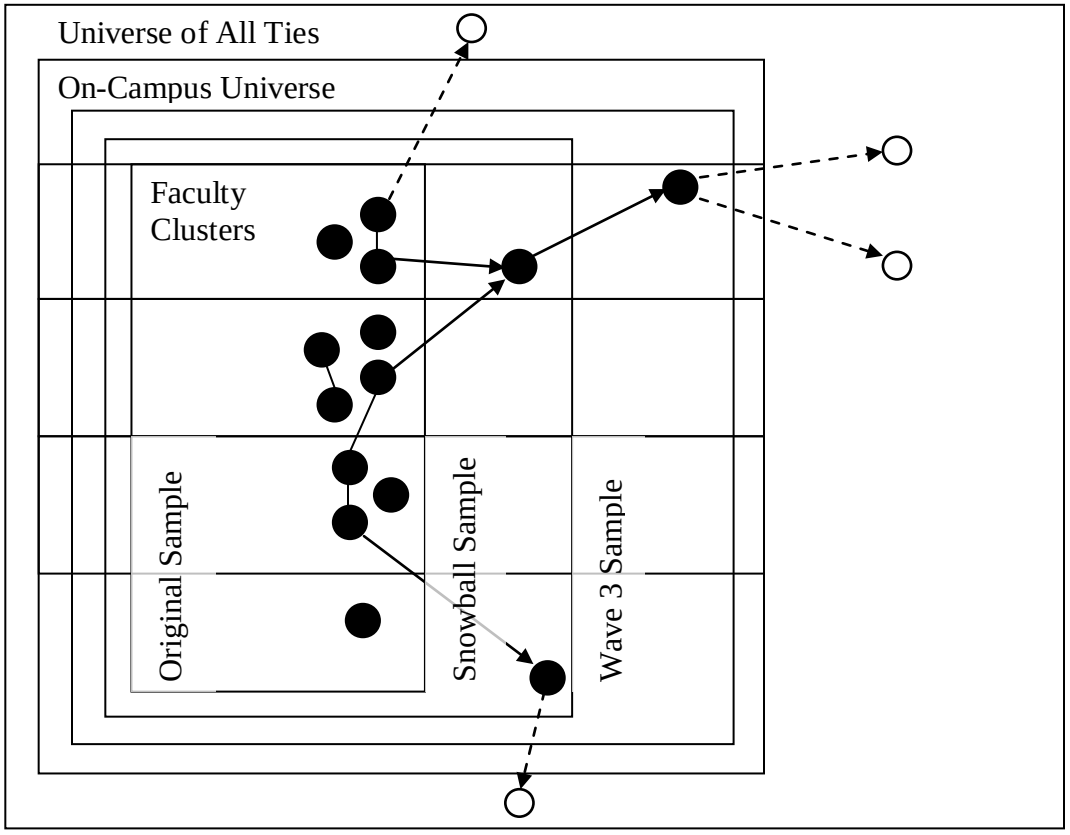
We began collecting data one month later on these 189 users to catch non-reliable users and to give new Facebook users a chance to fill in their networks. We then saved the pages for each person in the original sample and captured their complete social networks, as privacy settings permitted. Data include information about organizational affiliations (schools, workplaces, and regions) to which each friend belongs. This "snowball sample" (Goodman, 1961; Snidjers, 1992) made up a total of 33,191 'overlapping' people from the Facebook network. We pared this sample down to individuals who were members of the McGill network. We then

collected the same Facebook data for each member of this much smaller sample (N = 8,152).

We have thus constructed an 8,152-actor sociomatrix containing direct ties to campus peers, who were subsequently linked to another ~14,000 on-campus users (McGill network users). Our data thus contain ~22,000 on-campus Facebook users, as well as a potentially valuable enumeration (n=~50,000) of direct and indirect off-campus (non-McGill network) ties. Thus, we have three waves of data collection with webpages and either direct or indirect ties for their members: Wave 1 is the original sample; Wave 2, the snowball sample, for whom we also have direct ties (which are indirect ties to the original sample); the webpages of these indirect ties constitute wave 3. Finally, we recollected, in January 2008, data on the original random sample to identify changes in privacy settings of our primary actors. The sampling frame is represented in Figure 1 below. Here, black spots represent individual data that we measured, while white spots represented individuals that we did not gather full data about.

At each level of data collection, on-campus ties were coded on several attribute variables: gender, ethnicity, faculty/school, country of citizenship, affiliations to other college, regional, or employment networks, and graduating year. All attribute variables are coded based on respondents profile information, except for minority status and faculty/school. Minority status was coded using the profile picture of the respondent: to ensure reliability of coding, each actor in the original sample was coded and

Figure 1. Faculty-Clustered Snowball Sampling Frame



minority or white; when profile pictures were not provided or profile pictures did not include the actor, ethnicity was coded as indeterminate. Faculty/school was coded based on profile information on "Major" of actor and classifying majors by faculty/school given the disciplinary structure at the university. Some problems arose in majors such as Psychology, Geography and Mathematics, which are considered to be under both the Faculty/school of Arts and the Faculty/school of Science. In these cases, minors were used when they placed the actor clearly into a Faculty/school such as Arts, were coded as Arts for their faculty/school, while those who did not report a minor were coded as Science for their faculty/school. This is because at the university under study, Arts requires a minor whereas Science does not.

Measurement Issues

Prior to describing the undergraduate Facebook user network and our results, there are two specific measurement issues that must be discussed: the problems posed by the small-world phenomena, and the challenges of missing data.

Network datasets using random samples have to tackle the small-world problem/phenomenon. The small world phenomenon is grounded in the findings of Milgram (1967) and subsequent researchers (Lin *et al.*, 1978; Watts and Strogatz, 1998; Killworth and Bernard, 1978) showing that two randomly chosen people (strangers) can reach each other through a finite and very small number of alters, usually estimated as 6 affiliates or less (Watts and Strogatz, 1998). Thus, in an institutional population, irrespective of size it is to be expected that a randomly chosen sample will not be strangers to any high degree (Shotland, 1976; Lundberg, 1975). In fact, in our random sample of Facebook users, the average path length is 1.08. We do not however find this problematic for our purposes of examining the utility of Facebook as a name generator. It is unlikely, given that each *starting individual* was chosen randomly that their connections were unusually dense or sparse.[4] Our data collection has led us to 15-18000 unique users from a random sample of fewer than 200 actors, where the entire population of the university's undergraduate Facebook network is approximately 20,000 users. What is fairly obvious through the friendship ties that we observe *after* data collection is that most people in this network are no more than 2 degrees away from any other person in the Facebook network. Given this density it would have been surprising if our random sample was not interconnected.[5]

Missing data in network analysis have been considered to be more problematic than in other methodologies. Recent studies suggest however, that results may remain robust in the context of both tie and node level missing data, albeit much less so with the latter (Costenbader & Valente, 2003; Borgatti, Carley and Krackhardt, 2006; Kossinets, 2006). As discussed above, node level missing data occurred in our original sample by those who did not have public profiles. We removed them from our analysis

[4] In the collection of associational data, there is always the question of at what degree of acquaintance do we stop collecting data? The problem needs to be addressed by the research goals of the study as well as the practical issues in collecting information. While the latter suggests that a strict limit is set up at what level we stop collecting alters of alters, the former suggests what level this maximally should be. We collected alters of our random sample egos as well as alters. However, in this paper we present the analyses of only the egos' and alters' networks.

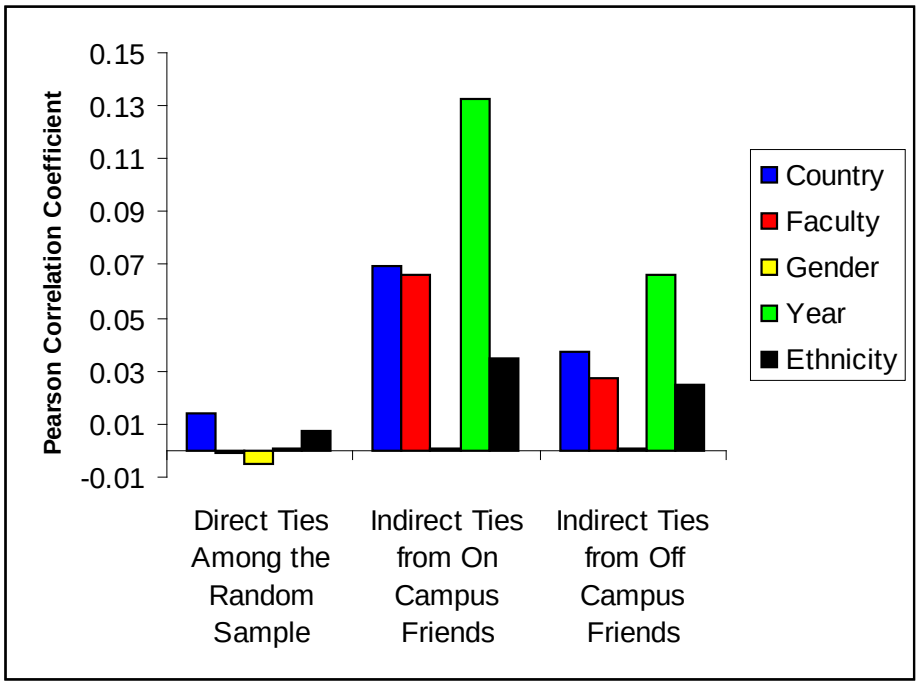
[5] If we had wanted to start with an unconnected set of actors, we would have been required to take a purposive sample, and skew our results in order to find a set of people whose first friends were also not in our sample. If we assumed non-cliquing, then we'd need a McGill viable sample size of 10000 just to get a set of 100 people who were not interconnected. While this method would have allowed us to maximize our coverage, we would not have found a sample that accurately represented the group, but rather one that necessarily over-sampled people with smaller social networks, and under-sampled those whose social networks are large, an unnecessary and confounding bias.

and do not consider it problematic given that it falls well within the accepted range of non-response evident in other social research. Given our research design, we do not have tie-level missing data. However, we do have missing data on attributes of actors in our original sample. As Table 1 shows, we have 20.6% missing on Faculty/school, 15.9% on country of citizenship, 19.6% on year of graduation, and 10.1% missing on ethnicity. There are two ways of accounting for this: examine the profiles of those missing and decide whether we can impute values of the missing attributes to the user; or let clustering and heterogeneity in the network relations help in imputing missing data. We have elected to take the second approach, as friends clique together on some important social information.

Analytic Procedures

Network maps were produced by UCINET 6.0 (Borgatti, Everett, and Freeman, 2002) and Statistics were run using Stata 10/SE. All network maps are presented with equal node repulsion and edge-length bias, so that nodes that share more ties are closer to each other, as well as to physically centralized nodes that are more connected. Alters and egos are mapped into one network through the affiliations procedure in UCINET.

Figure 2. Correlation Coefficients between Networks and Attributes



To test a range of hypotheses regarding how gender, race/ethnicity, class year, and citizenship pattern ties of friendship at McGill University, we ran autocorrelations as well as the QAP procedure (Borgatti, Everett and Freeman, 2002). For both procedures we developed a series of square attribute matrices indicating whether or not two individuals in the random sample shared the same categorical qualities. We then ran a series of independent correlations to test the extent to which these attributes correlated with the matrices of friendship. Figure 2 shows these correlations.

These tests were run thrice: one for the direct ties on campus, and once each for the indirect ties generated by co-occurring friendships from on- and off-campus alters. In essence, the

correlations in Figure 2 show homophily among the members of the random sample. *However*, it should be noted that the measures of homophily created through indirect ties are limited to indirect homophily among the random sample; that is, they do not account for the attributes of the direct friends who have generated the relations. In addition, correlations contain individuals' missing information. Individuals whose attributes were unknown were not assigned values, leaving their missing/uncodable attributes as a valid category for which homophily was possible.

RESULTS

Comparing University & Facebook Populations

The average number of friends an actor has in our sample is: 175.84 (163.03). This is comparable to previous network studies, which suggest that respondents have on average 100 to 200 'immediate contacts' s/he can link up with in an attempt to reach a target stranger (Degenne and Forsée, 1999). However, the distribution is skewed right, with a maximum of more than 750 **f**riends per person in our original sample and a large group of individuals reporting having no friends (N=37), an unlikely reality. Removing people with no friends increases the average while decreasing the standard error to 217.80 and (154.17) respectively.

Tables 1 to 3 compare the distributions of McGill undergraduate students and the Facebook sample of McGill undergraduates by gender, faculty/school, class levels and country of origin. We find that the sample under-represents Education students and slightly over-represents women (Table 1), as well as students from the province of Québec, considered here to be 'regional students' (Table 2). Our sample also shows an overrepresentation of female Science students (Table 1).

Table 1. Distribution of Undergraduate Students at McGill University and in the Facebook Sample by Gender and Faculty, 2006-2007

	Female Proportion of Faculty	Faculty Proportion Female	Male Proportion of Faculty	Faculty Proportion Male	Total	Total Proportion in Faculty
At McGill						
Arts	36.7%	67.2%	23.1%	32.8%	7,446	30.8%
Science	19.8%	53.2%	22.5%	46.8%	5,077	21.0%
Engineering	6.5%	25.9%	24.0%	74.1%	3,421	14.1%
Education	14.9%	78.9%	5.1%	21.1%	2,571	10.6%
Management	11.7%	48.3%	16.1%	51.7%	3,295	13.6%
Other	10.4%	59.5%	9.2%	40.5%	2,394	9.9%
Total	13,630	56.3%	10,574	43.7%	24,204	
In Sample						
Unknown	20.7%	59.0%	20.5%	41.0%	39	20.6%
Arts	27.9%	63.3%	23.1%	36.7%	49	25.9%
Science	27.0%	71.4%	15.4%	28.6%	42	22.2%
Engineering	3.6%	16.7%	25.6%	83.3%	24	12.7%
Education	4.5%	100.0%	0.0%	0.0%	5	2.6%
Management	12.6%	60.9%	11.5%	39.1%	23	12.2%
Other	3.6%	57.1%	3.8%	42.9%	7	3.7%
Total	111	58.7%	78	41.3%	189	

Table 2. Distribution of Undergraduate Students at McGill University and in the Facebook Sample by Citizenship Countries, 2006-2007

	Proportion at McGill	Proportion in Sample
QC	56.2%	24.3%
Rest of Canada	26.04%	33.3%
USA	7.9%	10.1%
Other	9.9%	16.4%
Unknown		15.9%
Total	24,463	189

Table 3. Distribution of Undergraduate Students by Class Level at McGill University and on Facebook, 2006-2007

	Proportion at McGill	Proportion in Sample
First Year	10.6%	15.3%
Second Year	27.3%	19.6%
Third Year	25.9%	22.2%
Fourth Year	32.9%	23.3%
Fifth Year	3.2%	1.1%
Unknown		19.6%
Total	20,347	189

While our sample seems to be fairly representative of undergraduates by status at university (First Year, Second Year, etc.), the distribution of our sample is skewed towards first year students and under represents fourth year students (Table 3). Our results therefore suggest that using Facebook as a name generator for offline ties may require us to pay attention to known distributions in the study population to either (1) oversample those who may be otherwise missed; (2) create weights for analysis; or (3) explore theoretically the reasons for why we may be systematically missing certain parts of the study population.

We take the latter route here. Firstly, it is not surprising that both the education and the regional students are underrepresented in our sample, given that these categories tend to overlap at the undergraduate level. This is a consequence of the structure of the program and its goals at the study university. Secondly, previous research has shown that Facebook is under-utilized (proportionally speaking) by non-English speakers; given that a large proportion of regional students come from a non-English speaking background, it is not surprising to find that these regional students have been missed in our random sample. This is an interesting finding that should be explored to examine to what extent there is segmentation within the undergraduate community of friendships between regional, national, and international students.

In this sample there appears to be support for the hypotheses that regional students tend to have fewer on-campus ties compared to their national and international counterparts. From anecdotal evidence, it appears that these regional students tend to identify themselves primarily with the regional network rather than the university network, though alumni make up a large proportion of the people in the university network.

This is further confirmed if we compare raw data on university populations and Facebook populations of the university: we find that among all the major non-English speaking universities in the region, only one has a high rate (60%) of identification by institutional network, while all others show rates lower than 30% (Table 2). This latter finding could be explored in future research by more rigorously testing for integration of non-English speaking students in Anglophone institutions (Table 5). Non-English speakers are in a minority in the current sample. Social networking, along with many other forms of computer use, is considered to be a measure of social status. This finding suggests that large forms of categorical social inequality (Tilly, 1998) may indeed cross over to both the formation of networks, the maintenance of these networks, and for the simple use of social networking aids such as Facebook. Thus, we can consider this to be a form of selection bias based on larger social structures.

Separating the sample, so as best to consider gender differences, results in some fairly interesting results. Table 4 gives cross-tabulations of the numbers of indirect ties and direct ties on and off campus, separated by gender. This shows that there are significantly more indirect ties on-campus for men than for women while indirect ties off campus show no significant gender differences. This, coupled with the lack of meaningful gender differences in the total number of friends, hints at the possibility of gender-based differences in cliquing tendency. Specifically, men may be more integrated into on-campus network and more likely to be embedded in transitive triads.

Table 4. Summary Statistics of Direct and Indirect Friendships in a Random Sample of McGill Undergraduate Facebook Users, by Gender

		Direct Ties	Indirect Ties On Campus	Indirect Ties Off Campus
Female	Mean	1.081	77.090	150.126
	Number of Cases	111	111	111
	Standard Error	0.131	5.019	10.166
Male	Mean	1.081	91.333	137.436
	Number of Cases	78	78	78
	Standard Error	0.147	8.638	12.233

Table 5. Francophone and Anglophone Networks in Montreal

University	Actual student population (2007)	Facebook population	% University Students with Facebook profiles
Université de Montréal	35000	8034	22.95%
Université de Sherbrooke	19000	5662	29.80%
Université Laval	35000	6084	17.38%
HEC Montréal	10000	6016	60.16%
McGill University	33000	32775	99.32%
Concordia University	31000	10936	35.28%
Université de Quebec á Montréal	40000	0	0

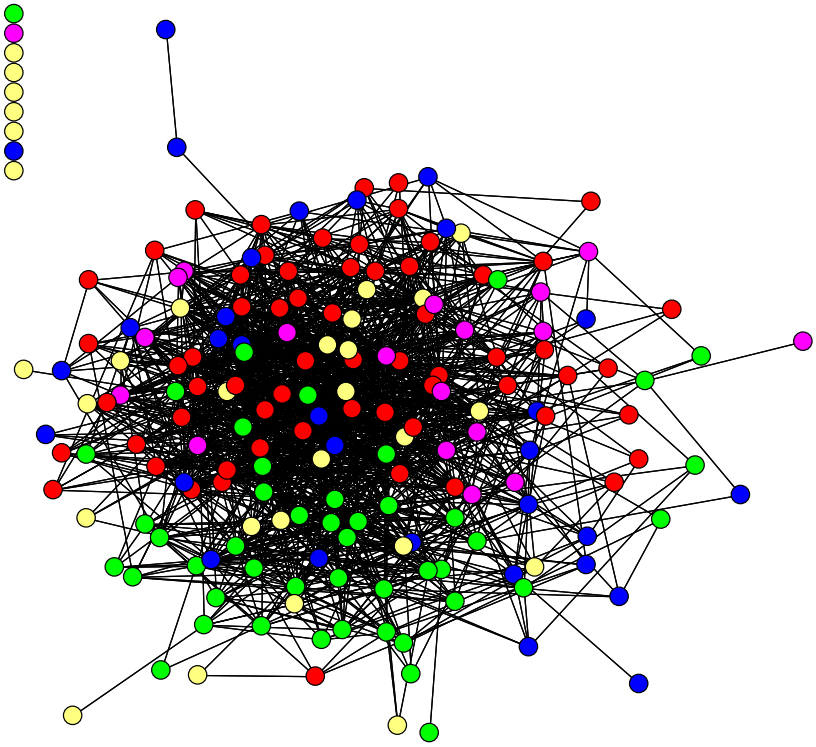
Source: Association of Universities and Colleges of Canada

Structural Properties of Ties by Attributes

If we consider the spring-embedded layout in Figure 3 below, we can see that there is very little distance between most of the nodes due to nationality. This was fairly representative, as many of our nodes showed very little distance due to country, faculty and even less to ethnicity. As was shown in Figure 2 above, the year of expected graduation was the strongest correlate of being connected, while gender was the weakest. These relationships hold through both indirect ties from on-campus friends as well as indirect ties from off-campus friends. We suggest that these results can be interpreted to mean that ties display the greatest homophily by class graduating year and the least homophily by gender as is evident in Figure 2. While

homophily normally refers to ties being generated by nodes sharing similar attributes (a possibility that is not explored here, but will be in future analyses), here we are also able to interpret homophily as the degree of shared ties by those who have similar attributes: those with the same class year for example will have more common friends (on campus and off campus) than those with differing class years. This, we label *Network Homophily*. Not surprisingly, gender does not predict network homophily, as there tends to be a lot of interaction between genders. The importance of network homophily by year does imply an ingrained importance of cohort and may also suggest a mechanism for the creation of a culturally homogeneous cohort – all share similar friends and thus diffusion of information is higher within than between cohorts.

Figure 3. Spring Embedding Layout of Indirect Ties Among the Random Sample Generated from Off-Campus Friends colored by Country



Notes: Red Nodes are from the Rest of Canada. Blue Nodes are from Other Countries. Green Nodes are from Quebec. Pink Nodes are from the USA. Yellow Nodes are indeterminate.

Multiplexity refers to the existence of two or more types of relations linking actors (Fischer *et al.*, 1977), and can be thought of as "the degree to which relations between participants include overlapping institutional spheres. For instance, individuals who are work associates may also be linked by family ties, political affiliations, or club memberships" (Portes, 1995). Conventional multiplexity refers to variation in the number of ties (e.g., friendship vs. business ties).[6] In our

[6] Multiplexity normally refers to kinds of ties. However, there is also what is called nodal multiplexity, which refers to variation in relational experiences within pluralistic actors (e.g., teams, organizations, collectivities). Nodal multiplexity is

analysis, we examined multiplexity in terms of direct and indirect ties, such that: individuals A and B may be tied to each other both directly and indirectly, with different associations capturing a meaningful picture of variations in relations. Thus, if actors A and B are tied to each other directly without any common friends, while actors C and D are tied to each other directly (through their Friends' lists) but also through actors E and F, i.e., C and D share friend E and friend F, then we suggest that relation between A and B is qualitatively different than the relation between C and D.

Multiplexity can be seen in the different natures of friendship. Sample Multiplexity Score (SMS) can be considered the aggregate measure of this type of difference in types of network connection and edge sharing. Here, the sample has a multiplexity score of 0.240, as calculated using Equation 1 below:

$$SMS = \sum_{i=1}^{k} \frac{noct_i}{nct_i} / k \quad (1)$$

Here, the number of off-campus ties (NOCT) is related to the number of on-campus ties (NCT) for each i^{th} number of shared edges and indicates that people have around 24% as many indirect ties off campus as compared to on. People who shared no edges were left out of the analysis for two reasons: 1) semantic differences in the meaning of multiplexity for this group, 2) numerical suppression in sample sizes due to uncontrolled limiting factors inherent to campus life.

Future Possibilities

Representations on Facebook are good representations of offline relationships. Social Networking Sites can potentially provide considerable network data in a cost-effective and efficient manner. Further, the data provide information on both attributes and networks in a minimally-biased manner. Social network studies have largely faced the problem of small numbers versus overwhelming data collection and verification procedures. Here, we show that Facebook provides an easy way to gain some insight into the ways that friends cluster, and the ways that clusters intertwine for individuals who use the site. Finally, while it is still the case that use of Facebook is not universal, it is growing and has been accepted by a vast number of individuals in a way that allows researchers a stable way to measure the interconnections of individuals, even if they are not stable, as is the case with college undergraduates.

Further data are also available on political, religious, educational, employment, and regional network affiliations that allow us to understand how people clique, with whom they clique, as well as the geographical placement of social resources. Data are further identifiable to the researchers by name and can thus be linked to data from the University itself on their academic achievements as well as some characteristics of their family of origin. With the proliferation of Facebook, data can only become richer. We look forward to gathering meaningful data on waves of individuals longitudinally to assess occupational outcomes, track changes in political affiliation over time, as well as following network maintenance through the process of maturing.

DISCUSSION

Facebook has provided us with a number of possible insights, as well as a few ideas for new theoretical constructs. The insights into the ways that gender differences play out in overall network shape and ability has important consequences. Men were more connected than women on-campus, though no differences were seen off-campus. Moreover, the gender of the node had little effect on whether ties were direct or indirect, suggesting that while the genders might use ties differently, a proposition that was not tested here, they do not really make them at any different rate. Cohort effects were obvious

---

relevant when assessing how individuals' and organizations' prior exchange experiences influence subsequent inter-organizational exchange behavior.

and significant, though within cohorts there were no gender differences. Differences in use by language suggest that there are some categorical inequalities that are playing out in the overall sample. This is also evident in the slight over-representation of females in our sample, suggesting that differences between women and men in our sample may be artificially reduced due to selection into the Facebook community.

In this study, it was also necessary to make a number of new theoretical constructs, and to nuance others, in order to understand the SNS format. Homophily, a well-used construct in the literature, was retooled here to describe the evident clustering of 'like' individuals, rather than the propensity for people to find other 'like' individuals. Interestingly, some of the strongest results here exhibited temporal rather than categorical clustering, suggesting that people make lots of friends when situated in a class with them. We have also had to look at multiplexity as being related to the differences between individuals in their propensity to have friends off- versus on-campus. This form of multiplexity, intra versus extra-institutional ties, allows us a nuanced view of how multiplexity might be created as each friendship in each institution starts in early life to overlap with others.

The greatest contribution this paper makes is in considering the application of rigorous descriptive methodologies to the gathering of social networks data. We have called this study a Quantitative Ethnography. This is mostly due to our focus on describing sociologically the interactions of individuals in a virtual society. This has allowed us a little freedom to actually focus on the meaning of ties in this way, and we hope that others will follow our example. Using this type of data also allows us to make inferences from the clustering of virtual representations of actual people to discuss a much greater variety of social topics than could be fully addressed before. We therefore believe that we have presented a much less biased solution to acquaintance or friendship network studies than can be given by traditional name generation methods. Rather than understanding how individuals cluster by the ways that information can pass, the ways that academic citations work, the manners by which managers serve on boards, or by the use of small numbers of mostly familial ties, we have shown that we can measure the actual extant ties between individuals. Moreover, we have shown that we can measure ties that are both acknowledged by the individual and small-world ties of which they may remain unawares. We have also shown that with a fairly small budget and extremely limited resources, that large and fairly complete social network data can be gathered that includes a variety of important social, educational, economic, and geographical data that attach to large and easily accessed social networks data.

Previous research has suggested that online networks reflect offline networks in important ways. This paper started with the proposition that Facebook data could overcome some of the disadvantages posed by the survey methodology of collecting network data. Following collection of data and network effects of certain socio-demographic variables, we have shown that online networks appear to mimic offline trends of social ties based on gender and age-groups. Thus, given the rich source of network data that Facebook offers and the relative ease with and cost-effective means by which it can be collected, future research could uncover important mechanisms of the maintenance of social ties that are not restricted to dynamics of online forums but rather of offline communities.

REFERENCES

Borgatti, SP, MG Everett, LC Freeman. 2002. "Ucinet for windows: Software for social network analysis." *Harvard: Analytic Technologies.*

Borgatti, Stephen P., Kathleen M. Carley and David Krackhardt. 2006. "On the robustness of centrality measures under conditions of imperfect data." *Social Networks*. 28(2): 124-136.

Boyd, Danah and Nicole Ellison. 2007 “Social network sites: Definition, history, and scholarship." *Journal of Computer-Mediated Communication* 13(1): http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

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

Degenne, A, M Forsé. 1999. *Introducing social networks*. Sage Publications Inc.

Ellison, N, C Steinfield, C Lampe. 2006. “Spatially bounded online social networks and social capital” *International Communication Association*. 1-37.

Fischer, Claude S., Robert M. Jackson, C. Ann Stueve, Kathleen Gerson, Lynne M. Jones and Mark baldassare. 1977. *Networks and Places.* New York: Free Press.

Frank, O. 1981. “A Survey of Statistical Methods for Graph Analysis.” In: S. Leinhardt (ed.) *Sociological Methodology*. Jossey-Bass, San Francisco, 110-155.

Goodman, Leo A. 1961. “Snowball Sampling.” *The Annals of Mathematical Statistics*. 32(1): 148-170

Granovetter, M. 1976. “Network Sampling: Some First Steps.” *The American Journal of Sociology*, 81(6), 1287-1303.

Gross, Ben and Alessandro Acqusti (2003), “Balances of Power on eBay: Peers or Unequals?” 1-5. http://www2.sims.berkeley.edu/research/conferences/p2pecon/papers/s2-gross.pdf

Hargittai, Eszter. 2007. “Whose Space? Difference Among User and Non-Users of Social Network Sites.” *Journal of Computer-Mediated Communication* 13(1): http://jcmc.indiana.edu/vol13/issue1/hargittai.html

Haythornthwaite, C. 2005. “Social networks and Internet connectivity effects.” *Information, Communication, & Society* 8(2): 125-147.

Hogan, B. 2008. “Analyzing social networks via the Internet.” *The Sage handbook of online research methods*, 141-154.

Killworth, PD, and HR Bernard. 1978. "The reversal small-world experiment." *Social Networks* 1:159–192.

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

Lampe, Cliff, Nicole Ellison and Charles Steinfeld. 2007. “A Familiar Face(book): Profile Elements as Signals in an Online Social Network.” CHI 2007 Proceedings on Online Representation of Self, San Jose: April 28-May 3.

Lampe, C., Ellison, N., and Steinfield, C., 2006. “A Face(book) in the crowd: Social searching vs. social browsing.” *Proceedings of CSCW-2006* (pp. 167-170). New York: ACM Press.

Lampe, C., Ellison, N., and Steinfeld, C. 2007. “A familiar Face(book): Profile elements as signals in an online social network.” *Proceedings of Conference on Human Factors in Computing Systems* (pp. 435-444). New York: ACM Press.

Lin, Nan, Paul W. Dayton and Peter Greenwald. 1978. “Analyzing the instrumental use of relations in the context of social structure.” *Sociological Methods Research* 7:149.

Lundberg, A. 1975. “Control of spinal mechanisms from the brain.” In: Tower, DB (Ed.) *The Nervous System. Vol. l.* New York: Raven Press.

Marsden , Peter V. 1990. “Network Data and Measurement.”*Annual Review of Sociology*. 16: 435-463.

Milgram, Stanley. 1967. *Psychology Today*. 2:61-67.

Portes, Alejandro. 1995. “Children of immigrants: Segmented assimilation and its determinants.” In *The economic sociology of immigration: Essays on networks, ethnicity, and entrepreneurship*, ed. A. Portes, 248-80. New York: Russell Sage Foundation.

Shotland, RL. 1976. *University communication networks: The small world method.* John Wiley & Sons Publishing.

Snijders, TAB. 1992. “Estimation on the basis of snowball samples: How to weight.” *Bulletin de méthodologie sociologique*.

Tilly, Charles. 1998. *Durable inequality*. Berkeley, Calif. ; London: University of California Press.

Wasserman, Stanley, Katherine Faust Social. 1994. *Network Analysis: Methods and Applications.* Cambridge University Press.

Watts, Duncan J.; Strogatz, Steven H. 1998. “Collective dynamics of `small-world' networks” *Nature*, 393(6684): 440-442.

Productivity and Performance in Academic Networks: Applications of Liaison Communication to Simmelian Ties, Structural Holes, and Degree Centrality

Devan Rosen
Department of Speech, University of Hawaii

In 1968, Donald Schwartz completed what is now seen as the first network analysis performed in the field of communication (Rogers, 1994). The results found in this paper confirm the significance of Schwartz' (1968) original research and extend his research to findings associated with the performance and productivity of academic researchers. The ability to retest the data collected by Schwarz in 1968 is a testament to his methods and processes, while new processes, such as a unique measure of Simmelian ties, were developed and utilized in this study. The similarity of the perceived and demographic data across three dimensions, Simmelian tie, structural holes, and degree centrality, not only support the original research but also provide insights into the effects of structure and position on performance and perception of academic networks. Findings related to categorical demographic data, rank and gender, offer a view into the nature of academic organizational networks and help tell their story. Structural holes (constraints) were found to decrease as tenure increased in an educational context, contrary to Burt's (1992b) findings and in support of Susskind et al.'s (1998) findings. This finding is explained as a combination of the level of seniority of the respondents and general organizational structure. The current research highlights the ability of network analysis to reveal organizational structure via communication linkages.

Authors*: Devan Rosen, Ph.D. (Cornell University, 2007) is an Assistant Professor at the University of Hawaii at Manoa. He has published on topics including decentralized social networks, self-organizing systems, flock theory, and computer-mediated communication. He has also developed network analytic methods for the structural and content analysis of online communities and virtual worlds.*

Acknowledgments: *An earlier version of this manuscript was presented to the 2001 International Communication Association Conference, Washington, D.C. The author would like to thank Dean Krikorian for editorial comments on earlier drafts, Mirit Shoham and Dean Krikorian for data coding, Sung-Choon Kang and Hichang Cho for statistical help, and Donald Schwartz for his pioneering research, and for sharing his original data set.*

Correspondence: *Contact Devan Rosen, Department of Speech, University of Hawaii, George 326, 2560 Campus Rd, Honolulu, HI 96822, rosend@hawaii.edu, 808-956-8911.*

INTRODUCTION

Schwartz' study, "Liaison Communication Roles in Formal Organizations" (Schwartz, 1968), has become a classic in the field of organizational communication. His research, then a doctoral dissertation, is now seen as the first communication network analysis performed in the field of communication (Rogers, 1994). Schwartz based his study largely on the work of Jacobson and Seashore (1951) and Weiss and Jacobson (1955), which postulated that interpersonal and group work-related communication patterns could help conceptualize and describe the structure of an organization. Seven years later, Schwartz revisited these notions in a study that reported on the liaison role in complex organizations and how they represented an elaboration of the descriptive analysis of complex organizations (Schwartz & Jacobson, 1975).

The current research revisits and reanalyzes Schwartz' original data using more recent theories and methodologies, and also extends the original research with contemporary concepts and methods. The first section describes the social network approach and the three network theoretic approaches used; Simmelian ties, structural holes, and degree centrality. Following a brief description of the main research question and the relevant sources of data, Simmelian ties (Krackhardt, 1996), structural holes (constraints) (Burt, 1992a), and degree centrality (Freeman, 1979) are operationalized in the context of the current study. Finally, results are presented, followed by discussion, theoretical and practical applications, and conclusions.

Social Networks

Social network perspectives focus on the structure of social systems and how the elements of a social system come together. Individual characteristics are only part of the story; people influence each other, and ideas and materials flow throughout the network. From the network perspective, the social environment can be expressed as patterns or regularities in relationships among interacting units. These patterns are often called structure. The current section elaborates some of the network concepts and terminology used in the subsequent methods of analysis.

The form of network that will be utilized is a communication network, defined as the patterns of contact that are created by the flow of messages among communicators through time and space (see Monge & Contractor, 2003; Rogers & Kinkaid, 1981). Communication network analysis identifies the communication structure, or communication flow. Relation ties (linkages) between actors are channels for either the transfer (flow) of material or nonmaterial resources, or for an association between actors, such as friendship ties. The ties that exist between the nodes can vary along several elements, including direction, reciprocity, and strength.

Links between actors can be measured as either directional or non-directional. Links that are directional indicate the movement from one point to another, such as the number of phone calls one person makes to another, or the degree of liking one person has for another. Additionally, these links can also be symmetrical or asymmetrical. If the link is directional but without the same value of relation the link is asymmetrical and lacks reciprocity. Non-directional links simply indicate an association of two actors in a shared partnership, such as two students being part of the same class. Several measures of how connected individual nodes are, as well as how connected the entire network is, are discussed below.

Simmelian Ties

The notion of Simmelian ties largely stems form the work of David Krackhardt (see Krackhardt 1992, 1996), as a combination of Granovetter's (1982) notions of strong ties and Simmel's (1950) contributions concerning triads as the fundamental unit of analysis. Strong ties within a network provide a greater motivation to assist and are typically more available than weak ties (Granovetter, 1982). Strong ties can be comprised of four main elements (Granovetter 1973):

1) Amount of time interacting
2) Emotional intensity during the interaction
3) Extent of mutual confiding in the relationship
4) The degree of reciprocal services enacted

Similarly, Simmel focused on the relationships that form between actors as being integral to the understanding of behavior. Simmel (1950) visited the notion that social triads are fundamentally different from dyads and should be studied accordingly. Simmel's model visited three main ways that triads could be distinguished from dyads in the way that the participants interact.

First, triads preserve a smaller amount of individuality that dyads. In a group of three or more a majority can be derived and an individual can thus be outvoted, resulting in the likely suppression of individual interests; that regardless of an individuals strength of preference, majority still wins.

Second, actors have less bargaining power in a triad that a dyad. A dyadic group can be destroyed if the demands of an individual are not accommodated, whereas in a triad the demanding actor can leave and thus has the most to lose; departing individuals would be isolating themselves while the group still remains intact. The remaining members are still able to make decisions without the defector, albeit no longer able to take advantage of the social benefits of the triadic structure.

Third, triads are more able to deal with conflict than dyads. The presence of the third party allows for hardened positions to be moderated and reformulated. This action is not necessarily intentional, the simple presence of the third party can alleviate tensions. Simmel (1950) states, "The appearance of the third party indicates transition, conciliation, and abandonment of absolute contrast. Such mediations need not occur in words: a gesture, a way of listening, the quality of feeling which proceeds from a person, suffice to give this dissent between two others a direction toward consensus."

Krackhardt (1996) derives a consequence of Simmel's approach, "One would expect that individuals who are part of a three person (or more) informal group are less free, less independent, and more constrained than a person who is only part of a strong dyadic relationship." Based largely on the work of Simmel, Krackhardt (1996) defines a Simmelian tie thusly, "two people are 'Simmelian Tied' to one another if they are reciprocally and strongly tied to each other and if they are each reciprocally and strongly tied to at least one third party in common." He goes on to point out that Simmelian ties are best thought of as "super strong" ties that add durability and power above that found in strong dyads, thus making Simmelian ties longer lasting.

Applying a Simmelian tie to the Schwartz dataset allows for a level of analysis utilizing the richness of strong tie information. Ties can be modeled as Simmelian and indicative of the strength of triadic relationships.

Structural Holes

Structural holes have been studied widely but are primarily based on the work of Ronald Burt (1992a, 1992b). Burt (1992a) states that, "a structural hole is a relationship of non-redundancy between two contacts. The hole is a buffer, like an insulator in an electric circuit. As a result of the hole between them, the two contacts provide network benefits that are in some degree additive rather that overlapping." He also highlights that these holes can have different effects for individuals with different attributes as well as for organizations of different kinds.

Burt (1992b) performed a longitudinal study looking at the rate of promotions for managers in a large diversified company. He found that managers in conditions of higher structural holes were promoted more quickly, suggesting that structural holes lead to increased recourses and network influence. It was also noted that structural holes might not be equally advantageous for all (such as women or the elderly). Burt's (1992a, 1992b) research focused

on social networks in business organizations, which may represent different social and structural phenomena than an academic organization.

In looking at organizational down-sizing, Susskind, Miller, and Johnson (1998) summarize structural holes as existing when two members not directly connected to each other lack a common network contact. Likewise, structural holes make a network more constrained or sparse as individuals have less opportunity to access novel information and resources. Structural holes can also lead to inequality between network members and power opportunities.

Constraint, as presented by Susskind et al. (1998), represents the distribution of relationships across a member's network or the extent to which an actor's network is dependant on a limited number of network members. "Constraint is positively related to the formation of structural holes, as high constraint indicates more structural holes for an employee." (Susskind et al, 1998). The measure used to indicate structural holes in the current study is the Constraint measure. Constraint is the most applicable structural hole indicator for the Schwartz dataset because of the nature of the academic organizational context; educational ties are generally distributed to limited actors within the overall network. For example, departmental units in an educational context would tend to cluster around functional linkages (Shoham, Lee, & Jones, 2001). By its very nature, pockets of clusters would seemingly reflect intra-departmental communication within departmental cliques (Stefanone, Moyersoen, & Krikorian, 2001). Constraint, as dependency among actors, can be viewed as a negative characteristic in oligarchical, or diffused organizational structures.

Degree Centrality

The degree measure of centrality is calculated by counting the number of adjacent links to or from an actor in a network (Brass & Burkhardt, 1992). Freeman (1979) conceptualized this measure as an indicator of individual activity, yet it does not capture system-wide properties of the network. It does, however, represent the number of alternatives available to an individual in the network. This in turn makes it a viable centrality to use in conjuncture with structural holes.

Degree centrality may also be appropriate for capturing those power-enhancing behaviors that happen via direct interaction, such as integration and reciprocation. Degree centrality can also indicate other direct interactions such as coalitions or the avoidance of relying on mediating actors for indirect access to resources (Brass & Burkhardt, 1992).

While a relatively straightforward measure, degree centrality provides insight into individual contributions to the interconnectedness of the overall network (Rogers & Kincaid, 1981). In the Schwartz dataset degree centrality can be used in comparison with structural holes and Simmelian tie measures. In this manner, the goal of the current research is not only to apply contemporary measures of network analysis, but also to contrast the results of these three different measures.

The research question addressed in this paper is: To what extent are Simmelian ties, structural holes, degree centrality, and individual characteristics reflected in the Schwartz dataset as related to productivity and performance?

METHODS

Data

The data used in this study was obtained from Donald Schwartz with permission for use. Schwartz initially used the data in 1968, and then revisited these notions in a study that reported on the liaison role in complex organizations and how they represented an elaboration of the descriptive analysis of complex organizations (Schwartz & Jacobson, 1977). The data provides for a multi-level evaluation of the productivity and effectiveness of professors and researchers in a research-based environment. Particularly relevant to the current analysis are demographic questions, liaison actor

identification, and perceived characteristics of liaison ties.

The population from which samples were drawn consists of the professional faculty and staff of a single college, situated in a single building on a university campus, with a sample size of 142. The questionnaires incorporated a contact checklist for network data, perceived characteristics of the personal contacts, and demographic data such as the number of publications and academic rank.

The data from the Schwartz study is of particular interest because it was collected specifically to investigate the relationship of network structure, communication, and performance. Whereas the initial study was specifically looking at liaison roles, which act as "gatekeepers" between otherwise unconnected parts of the network, the data that was collected lends itself extremely well to investigating structural holes, degree centrality, research performance, productivity (publications), and organizational performance (academic rank).

The methods described below explicate the techniques used to reconfigure the Schwartz matrix into Simmelian ties, structural holes, and degree centrality.

Simmelian Ties

The method to build a Simmelian tie matrix is a five-step process:

Step1: Symmetrize the original strength tie matrix with average values of tie strength with two actors (e.g., if actor 1 → 2 & tie strength = 3; actor 2 → 1 & tie strength = 5, the symmetrized strength of the tie between actors 1 and 2 is (3+5)/2 = 4).

Step2: Construct a strong tie matrix by dichotomizing the ties into strong and weak ties

Step3: Perform a clique analysis to identify a co-clique member; with Ucinet 5, Tool: Clique overlap; (see Borgatti, Everett, & Freeman, 1998).

Step4: Dichotomize the ties if A and B have a same clique membership; if members of the same clique, then Simmelian tie value = 1, else = 0.

Step5: In above co-clique matrix, diagonal elements reflect the total number of individual membership cliques; diagonal elements are the sum of Simmelian ties of each actor.

The process used to generate the Simmelian tie matrix is unique to the methodology described herein. Further, the process can be converted into a working algorithm (i.e., allowing for automatic calculations), and relies upon tie-strength data.

Liaison Data

For the liaison data the difference between liaison and non-liaison actors is derived in terms of ties measured from the demographic questionnaire in Schwartz (1968). See Appendix A for details of items from the original survey. A value of one is assigned as liaison (n=22), values of two, three, and four are assigned as non-liaison (n=95), and values of five are isolates.

Structural Holes

The constraint measure in Ucinet V (Borgatti, Everett, & Freeman, 1998) was used to identify structural holes. As noted earlier, this particular measure seems best suited to the context of academic departments within a collocated geographical building.

Degree Centrality

Measures of degree centrality were likewise derived from Ucinet V (Borgatti, Everett, & Freeman, 1998). The centralities were then normalized by dividing the simple degree (or number of links) of an actor by the maximum degree possible, n*(n-1) and dividing this by 2 (bi-directionally) (see Wasserman & Faust, 1994).

The results of these various tests are then run through SPSS to produce correlation and MANOVA to indicate the relationships of Simmelian ties, structural holes, and degree centrality with demographic variables, liaison tests and perceptual variables in the original dataset.

RESULTS

Demographic Data

The results of the demographic tests can be found in Table 1, and revealed a Pearson Correlation of 0.464 ($p<0.01$) between the normalized degree centrality and Simmelian ties and a -0.687 ($p<0.01$) correlation between constraints (structural holes) and normalized degree centrality A non-significant correlation of –0.170 is revealed between constraint and Simmelian ties.

The specific demographic questions have generally low or insignificant correlation to the three variables tested, with the exception of committee work and degree centrality. There is a general trend of negative correlations with constraints and positive correlations with centrality.

Categorical Demographic Data

The two categorical questions tested accordingly are academic rank and gender (see Table 1 and 2). As academic rank increased, centrality also increased ($F= 4.096$, $p<0.001$), and structural holes decreased ($F= 5.584$, $p<0.001$). For gender, males had 38% lower structural hole scores than women ($F= 13.548$, $p<0.001$) and were 33% more central than women ($F= 4.987$, $p<0.001$).

Table 1. Demographic Variables

	Simmelian Ties	**Normalized Degree Centrality**	**Constraints**
Simmelian Ties	1.000	.464**	-.170
Normalized Degree Centrality	.464**	1.000	-.687**
Constraints	-.170	-.687**	1.000
Age	-.015	.107	-.278**
Rank	-.066	-.244**	.182*
Research	-.147	-.264**	.183*
Consulting	.002	.052	-.054
Committee Work	.108	.252**	-.175*
Admin. Duties	.105	.367**	-.275**
Year First at University	.104	.125	-.173*
Dept. Level Committees	.272**	.231**	-.152
College Level Committees	.097	.508**	-.270**
University Level Committees	.171	.542**	-.346**
Total # Committees	.232**	.462**	-.301**
Committees Meetings/Month	.286**	.383**	-.213*
# of Articles	.044	.123	-.213*
# of Books	.005	.109	-.173
Hours/Week Work	-.191*	-.223**	.247**

*Note: * =p≤.05, ** = p≤.01*

Table 2. Categorical Demographic Results, Rank

	Simmelian Ties	Normalized Degree Centrality	Constraints
Lecturer (n=2)	2.00	10.85	.53
Instructor (n=17)	1.00	13.82	.42
Assistant Professor (n=29)	1.12	19.99	.36
Associate Professor (n=29)	1.04	23.84	.24
Professor (n=49)	1.80	29.38	.22
	F=1.056	F=4.096***	F=5.584***

*Note: * =p≤.05, ** = p≤.01, ***= p≤.001*

Table 3. Categorical Demographic Results, Gender

	Simmelian Ties	Normalized Degree Centrality	Constraints
Male (n=112)	1.23	24.47	.23
Female (n=17)	1.76	16.55	.43
	F=1.132	F=4.987*	F=13.548***

*Note: * =p≤.05, ** = p≤.01, ***= p≤.001*

Table 4. Liaison Test

	Simmelian Ties	Normalized Degree Centrality	Constraints
Non-Liaison (n=95)	1.11	20.82	.32
Liaison (n=22)	3.14	39.71	.17
	F= 29.68***	F= 45.51***	F=14.68***

*Note: * =p≤.05, ** = p≤.01, ***= p≤.001*

Table 5. Perceived Characteristics of Liaison and Non-Liaison Actors

	Structural Diversity	Number of Contacts	First Source of Information	Importance of Secondary Contact	Perceived Power	Dyadic Opinion Leader
Non-Liaison (n=159)	11.64	8.92	7.18	14.55	23.72	20.53
Liaison (n=22)	12.82	11.55	9.45	18.60	28.08	22.48
	F=14.41 ***	F=35.17 ***	F=28.34 ***	F=54.88 ***	F=22.14 ***	F=5.39 *

*Note: * =p≤.05, ** = p≤.01, ***= p≤.001*

Liaison Tests

Liaisons have more Simmelian Ties (F= 29.68, p<0.001) and higher Degree Centrality (F= 45.51, p<0.001) than non-liaison actors, and less constraint and lower structural hole scores (F=14.68 p<0.001).

Perceived Characteristics of Liaison Ties

Tests of the perceived characteristics of liaison and non-liaison actors revealed that liaison actors have greater structural diversity (F=14.41, p<0.001), larger numbers of contacts (F=35.17, p<0.001), are more likely to be the first source of information (F=28.34, p<0.001), greater importance of secondary contact (F=54.88, p<0.001), and higher perceived power than non-liaison actors (F=22.14, p<0.001).

Results from the tests of perceived characteristics of liaison ties, as based on the personal contact questionnaire, show that Simmelian ties have a correlation of 0.476 (p<0.01) with normalized degree centrality, and –0.212 (p<0.01) with constraints. Normalized degree centrality has a correlation of -0.763 with constraints (p<0.01).

Table 6. Network Measure Correlations from Perceived Characteristics

	Simmelian Ties	Normalized Degree Centrality	Constraints
Simmelian Ties	1.00 (n=223)		
Norm. Degree Centrality	.476**	1.00 (n=223)	
Constraints	-.212**	-.736**	1.00 (n=223)

*Note: * =p≤.05, ** = p≤.01*

DISCUSSION

In general, the findings in this paper support the original research with some notable extensions to the original findings. Concerning the results of the demographic data, it can be inferred that since constraint and Simmelian ties are not significantly correlated, degree centrality can be seen as a predictor variable. Degree centrality, having a positive correlation of 0.464 with Simmelian ties and a negative correlation of –0.687 with constraint, sets centrality as a good predictor of both. This finding alludes to the notion that as an actor is more central in a network, the more likely they are to have really strong ties with other actors, as well as fewer structural holes around them. In an academic organization this is an intuitive finding. The more prominent faculty members and central administrators will be likely to have fewer roles not filled around them, including committee work, which is reflected in the data (See Schwartz & Jacobson, 1975). It is also important to note that the measures of productivity (the number of articles and books published) have insignificant correlations accept for number of articles and constraints with a correlation of –0.213 ($p<0.05$). While this correlation is narrowly significant, one cannot help but pose a refinement to the old adage by saying that in academe one must "publish or become a structural hole."

The generally low correlation of the non-categorical demographic variables with the Simmelian tie matrix shows that strongly associated cliques do not necessarily affect their performance and do not relate to their tenure. However, the generally positive correlations of the demographic areas with normalized degree centrality show that as an actor is more central they will sit on more committees more frequently and have more administrative duties. While the greater the amount of structural holes an actor experiences the fewer committees they will sit on and the less administrative duties they will have. This result provides more evidence of the nature of constraints as indicative of disproportionate power relations. Similar to Susskind et al's (1998) findings, structural holes in this study can lead to the "down side" of power relations; that structural holes can be used to indicate disadvantaged or advantaged individuals. The similarity between the current research and Susskind et al.'s (1998) findings regarding the down side of power relations could be that both studies examined the structural holes of employees at mostly lower levels of the organization. However, Burt (1992b) examined the structural holes of managers. Given a one-up one-down relationship (Rogers & Millar, 1979; Watzlawick, Bavelas, & Jackson, 1967) structural holes can be indicators of one-up and one-down power relations in organizations.

Analyses of the categorical demographic data showed that academic rank is positively related to degree centrality ($F= 4.096$, $p<0.001$) and negatively related to constraint ($F=5.584$, $p<0.001$). The constraint values went from a mean of 0.53 for a lecturer down to a 0.22 for a professor. While lecturers had a mean centrality of 10.85 going up to a 29.38 for professors, again with increasing through instructor (13.81), assistant professor (19.99), and associate professor (23.84). This finding provides further evidence of structural holes (constraint) as negative indicators of advancement in the context of this study.

Gender was significantly correlated with constraint ($F=13.55$, $p<0.001$) and degree centrality ($F= 4.987$, $p<0.05$). There are 112 males in the sample and only 17 females, which strengthens the constraint findings as a result of the high significance. The finding that males are more central and have fewer constraints leads to a conclusion that there is indeed a disparity in the way females acted in this network as opposed to the males. However, the disparity between the two categories was not as large concerning constraints as it was concerning centrality. Thus although the males were far more central, they only experienced 33% less structural holes.

The current study found the opposite relation of promotion and structural holes than Burt (1992b), as individuals increased in tenure their structural holes decreased. Burt (1992b) also indicates that structural holes can be disadvantageous for women and the elderly--this

study adds full professors to the list by indicating a negative relationship between career advancement and structural holes.

The results of the liaison tests echo Schwartz' (1968) results. Non-liaison actors were found to have fewer Simmelian ties ($F= 29.67$, $p<0.001$) and thus fewer "super" strong ties and lower degree centrality ($F= 45.51$, $p<0.001$), as well as more structural holes ($F= 14.68$, $p<0.001$) than liaison actors. Liaison actors thus play a more central role, have stronger ties, and have fewer constraints than non-liaison actors.

Incorporating perceived aspects of the data in liaison tests also replicates previous findings. Liaison actors have greater structural diversity, which supports the finding in the previous liaison test. Liaisons are also more likely to be the first source of information, have a greater importance of secondary contact, and have a higher perceived power.

The results from the tests of the perceived characteristics of liaison actors from the personal contact questionnaire show a very similar correlation to the results of the aforementioned demographic tests (see table 5). The demographic correlation of Simmelian tie to degree centrality is 0.464 ($p<0.01$), while the liaison correlation is 0.476 ($p<0.01$). Likewise, the demographic correlation of constraints to degree centrality is -0.687 ($p<0.01$), while the liaison correlation is –0.763 ($p<0.01$). These similarities further support the original research and call for further investigation into the similarity of these findings. These findings point to the accuracy of Schwartz' (1968) data by linking perceived characteristics with demographic information—as related to network variables.

Implications and Applications

Theoretical implications of this research examine the relationship between degree centrality, constraints, and Simmelian ties. The mediating role of degree centrality has interesting applications. For example, one who communicates as a hub, with many indegrees and outdegrees, would seemingly be able to perform more liaison functions because they already have the most number of ties. The relation between constraint and Simmelian ties is mediated by the number of ties. Also, differences in constraints for women in this study echoes findings from Ibarra (1997) who found structural differences between women and men regarding network homophily and contact range. Ibarra's research (1992, 1993), along with other studies that have measured structural differences between the genders (Brass, 1985, Burt, 1992b), have focused on traditional organizations for analysis. As such, further research into similar network measures (e.g. structural holes, homophily) in academic or research organizations may produce interesting differences.

The current research introduced a new approach to the development of Simmelian ties measures and how they may be implemented. The five-step process can be used as a type of measure for Simmelian ties. It would be interesting to know how such Simmelian ties operate in other organizational conditions. This measure can be used as an indicator of cluster strength and could potentially be used to detect covert operations (e.g., cabals) at early stages. As another example, the growth of an online community can be seen as clusters of activity around message topics or threads (See Krikorian & Kiyomiya, 2001). In this manner, the strength of message ties can be based on frequency and duration of message threads.

Schwartz (1968) found that more linkages were evident in higher positions in the network. This study echoed this finding in a different manner: the higher the position, the less structural holes. This finding was perhaps most interesting as it opposed Burt's (1992b) finding of more structural holes for advancing managers. Arguments supported both the organizational level (e.g., seniority) and the structure of the organization (e.g., oligarchy) as potential explanations for the appositional relations between tenure and structural holes. Practically, one should pay attention to the inherent structure of an organization before analyzing network ties. Also, the height and width of the organizational structure can affect whether structural holes or

Simmelian ties are more favorable in advancement processes.

It is interesting to see the similarity of demographic and attitudinal data. More research is called for in this comparison. If there is a mediating effect of networks between demographic and attitudinal variables, then this could have implications in the use of network data by providing insight into the mechanisms of interpersonal and group communication networks.

CONCLUSIONS

The results found in this paper confirm the significance of Schwartz' (1968) original research and also highlight the methodological foundations for supporting his original results. The ability to retest the data collected by Schwarz in 1968 is a testament to his methods and processes, while new processes, such as a unique measure of Simmelian ties were developed and used in this study. The similarity of the perceived and demographic data across three dimensions; Simmelian tie, structural holes, and degree centrality, not only support the original research but also provide unique insight into the data by extending these findings to include new measures and methods of analysis, as well as new theoretical implications. Likewise, the findings of the categorical demographic data, rank and gender, offer a view into the nature of organizational networks and help tell their story. Structural holes (constraints) were found to decrease as tenure increased in an educational context, contrary to Burt's (1992b) findings and in support of Susskind et al.'s (1998) findings. This finding is explained as a combination of the level of seniority of the respondents and general organizational structure. The current research highlights the power of network analysis to reveal organizational structure via communication linkages. It is hoped that this paper has helped open a window into the organization of organizations, and how this organization affects all of the actors within.

REFERENCES

Borgatti S., Everett M., & Freeman L. (1998). *Ucinet 5 For Windows*. Harvard, MA: Analytic Technologies.

Brass, D. (1985). Men's and women's networks: A study of interaction pattern's and influence in an organization. *Acadamy of Management Journal, 28*, 327-343.

Brass, D. J., & Burkhardt, M. E. (1992). Centrality and power in organizations. In N. Nohria & R. G. Eccles (Eds.), *Networks and organizations: Structure, form, and action* (pp. 191-215). Boston: Harvard Business School Press.

Burt, R. S. (1992a). The social structure of competition. In N. Nohria & R. G. Eccles (Eds.) *Networks and organizations: Structure, form, and action* (pp. 57-91). Boston: Harvard Business School Press.

Burt , R. S. (1992b). *Structural holes, the social structure of competition*. Cambridge, MA: Harvard University Press.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification.*Social Networks, 2*, 215 – 239.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal ofSociology, 78*, 1360-1380.

Granovetter, M. S. (1982). The strength of weak ties: A network theory revisited.In P. V. Marsden & N. Lin (Eds.), *Social structure and networks analysis* (pp. 105-130). Beverly Hills, CA: Sage.

Ibarra, H. (1992). Homophily and differential returns: Sex differences in network structure and access in an advertising firm. *Administrative Science Quarterly, 37*, 422-447.

Ibarra, H. (1993). Personal networks of women and monorities in management: A conceptual framework. *Academy of Management Review, 18*(1), 56-87.

Ibarra, H. (1997). Paving an alternate route: Gender differences in network strategies for career development. *Social Psychology Quarterly, 60*(1), 91-102.

Jacobson, E., & Seashore, S.E. (1951). Communication practices in complex organizations. *Journal of Social Issues, 7*, 28-40.

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In N. Nohria & R. G. Eccles (Eds.), *Networks and Organizations: Structure, form, and action* (pp. 216-239). Boston: Harvard Business School Press.

Krackhardt, D. (1996). *Groups, Roles, and Simmelian Ties in Organizations*. Working Paper Series: Heinz III School of Public Policy and Management. Pittsburgh, PA: Carnegie Mellon University.

Krikorian, D. H., & Kiyomiya, T. (2001). Bonafide groups as self-organizing systems: Applications to elecrtronic newgroups. In L. Frey (Ed.), *Groups in context: Bona fide groups*. New York: Houghton-Mifflin.

Monge, P. R. & Contractor, N. (2003). *Theories of communication networks*. New York: Oxford University Press.Rogers, E. M. (1994*). A history of communication study: A biographical approach.* New

York: Free Press.

Rogers, E. M., & Kincaid, L. D. (1981) *Communication networks: Toward a new paradigm for research*. New York: Free Press.

Rogers, L. E., & Millar, F. E. (1979). Domineeringness and dominance: A transactional view. *Human Communication Research, 5*, 238-246.

Schwartz, D.F. (1968). *Liaison communication roles in a formal organization* (Doctoral Dissertation, Michigan State University, 1968). University Microfilms No. 69-11, 162.

Schwartz, D. F., & Jacobson, E. (1977) Organizational communication network analysis: The liaison communication role. *Organizational Behavior, 18*, pp.158-74.

Simmel, G. (1950). Individual and Society, In Wolf, K.H. (Ed.), *The Sociology of Georg Simmel*. New York: Free Press.

Shoham, M., Lee, J., & Jones M. (2001). The microanalysis of liaison communication. Paper presented to the 2001 *International Communication Association Conference*, Washington, D.C.

Stefanone, M.A., Moyersoen, J., & Krikorian, D. (2001). The microanalysis of liaison communication. Paper presented to the 2001 *International Communication Association Conference*, Washington, D.C.

Susskind, A.M., Miller, V.D., and Johnson, J.D. (1998). Downsizing and Structural Holes. *Communication Research, 25*, 1, 30-65.

Weiss, R. S., & Jacobson, E. (1955). A method for the analysis of the structure of complex organizations. *American Sociological Review. 20*, 661-668.

Watzlawick, P., Bavelas, J. B., & Jackson D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. NY: W. W. Norton.

Wasserman, S., & Faust, K. (1994). *Social Network Analysis: Methods And Applications*. New York: Cambridge University Press.

Appendix A – Survey Items from Schwartz (1968)

Academic Rank
What is your academic rank?
1. Instructor, 2. Assistant Professor, 3. Associate Professor, 4. Professor
5. Other (Please specify)_____

Committee Service
How many faculty or administrative committees do you belong to including both standing and ad hoc committees?
___1. Departmental (or Institute) level committees
___2. College level committees
___3. University level committees

Publications
How many professional journal articles have you published (or had accepted for publication) and how many papers have you presented at professional meetings since 1965? ____ (combined total)

Personal contact checklist
Now go back over the past two or three months and think of the professional people in the College of (--) with whom you worked most closely. We would like to have you list below the names of the people in the college with whom you work most closely. By "work with most closely" we mean the professional people with whom you usually have at least one contact per week on matter related to programs or activities of the College, or in teaching, research, or consulting in which you or the other person is engaged. You need to only list people who are officed in (---) Hall. By "professional people" we mean faculty with academic rank of instructor or higher and/or administrators.

For each of the individuals you list below, check how frequently in an "average" week you have contact with (talk to in person or on the phone, write) each of them. Name as many or as few people as accurately describe your usual contacts

(A) List the name of each person in the College with whom you work most closely.
(B) For each person listed, check the appropriate frequency column.

Name_____
Frequency of contact:
1) Several times daily __
2) About once per day __
3) 2 or 3 times per week __
4) About once per week __

Identifying Organizational Influentials: Methods and Application using Social Network Data

Russell Cole, PhD
Mathematica Policy Research, Inc., Princeton, New Jersey

Michael Weiss, PhD
MDRC, New York, New York

Uncovering the most influential individuals in an organization may be of great use for researchers and practitioners. As central hubs in the organization, these individuals can be key co-creators or co-adapters for the diffusion of organizational reform. In this paper we examine the question "Who are the most influential individuals in an organization?" Using social network data, we assess organizational members' levels of influence in four different advice-seeking networks, as well as in one "friendship" network through a measure of peer-endorsement. We investigate four methods for the classification of individuals as "influentials". These methods are compared and contrasted according to their performance in handling problems of differing network sizes, densities, non-response rates, researcher decisions and parsimony. A nonparametric random permutation method is shown to be a consistent and objective process for the identification of influential individuals in a sample of High school staff members in nine schools.

Acknowledgements: *The research on which this article is based was supported by the U.S. Department of Education's Institute of Education Sciences (Grant #R308A960003). The authors are especially grateful for the ideas and criticism of Jonathan Supovitz, Elliot Weinbaum and Kelley Borradaile.*

Correspondence: *Contact Russell Cole, Mathematica Policy Research, Inc., P.O. Box 2393, Princeton, NJ 08543, or by E-mail at rcole@mathematica-mpr.com.*

INTRODUCTION

Researchers interested in educational reform have become increasingly aware of the role of the social context of schools (Schneider, 2005). Schools are complex social organizations, where informal communication can aid or hinder the implementation of innovative reforms (Fowler, 2004; McLaughlin, 1990). Interpersonal relationships among teachers in schools have been demonstrated to play a key role in influencing attitude and behavior changes regarding reform (Cole & Weinbaum, 2007; Frank, Zhao, & Borman, 2004). Teachers both influence and are influenced by their peers, and as such, the topic of influence in schools is of import for research into reform.

The broad topic of influence in an organization has often been framed as an aspect of leadership. Leadership in an organization requires the exercise of influence over the beliefs, actions, and values of others in an organization (Hart, 1995). However, influence over others can exist outside of the framework of formal leadership (Spillane, 2006). Social influence occurs when one individual adapts their own attitude, behavior, or belief to that of others in the organization (Leenders, 2002). It is possible, even probable, that individuals are influenced more by informal conversations with peers than by communication with formal leaders (Ibarra & Andrews, 1993; Ibarra, 1993). As such, it is necessary to investigate influence in schools without restricting its focus to that of "leader" and "follower." Leadership is not always synonymous with influence, and titular leaders are not always the most influential members of a school.

Uncovering the most influential individuals in an organization may be of great use for researchers or practitioners. As central hubs in the organization, these individuals can be key co-creators or co-adapters for the dissemination of a new program or reform (Kempe, Kleinberg, & Tardos, 2005). Through their extraordinary influence, the extent of the diffusion of an innovation (or the rate at which individuals adopt the innovation) may be increased. These "influentials" may be existing formal leaders or potential candidates for future leadership roles in their organizations. Identifying the key influential individuals in schools can play a key role in the introduction, longevity, and fidelity of program implementation (Riggan & Supovitz, 2008; Valente & Pumpuang, 2007).

Our primary research question asks, "Who are the most influential individuals in an organization?" Through social network analysis, we investigate methods for answering this question and illustrate the utility of these methods for future research. In our study, we ask school staff members to indicate individuals who influence them professionally (through advice) as well as those who have helped them with respect to social support or friendship. For the purposes of this paper, we define influentials to be those members of the organization who are frequently mentioned by their peers as highly influential in a given communicative context.

In the context of education research, identifying individuals as the organization's key influentials using social network methods has recently become an area of interest. Riggan and Supovitz (2008) identified influentials for the purpose of targeted follow-up interviews in their study of school leadership using one of the methods to be described in this paper. Spillane et al. (2006) examined whether being a central influential in a school is associated with holding a position of formal leadership. We will extend the work of these researchers to develop improved methods for identifying influentials.

Previous research using social network analysis to investigate influence in organizations is briefly described later. While there has been extensive research in this area, the particular idea of identifying a key influential subset of an organization's staff has not yet been sufficiently examined. In the next section, four methods will be developed for the identification of influential individuals in organizations. This will be followed by an examination of merits and shortcomings of each method. These methods

are then briefly employed in an empirical examination of influence as it exists in the staff of nine high schools. In the final Methods section presents limitations of these methods with implications for further research.

Social Network Analysis

Social network analysis assumes that individuals are interdependent and that the communication between individuals defines this interdependence (Wasserman & Faust, 1994). Social network data for an organization are often collected through surveys, where individuals indicate others with whom they communicate. These data, therefore, define the structure of communication, or the relationships between the actors (individuals) within the organization.

The social structure of a school can be represented as a sociogram (Wasserman & Faust, 1994). Previous research has used sociograms in an effort to identify key individuals in an organization (Birk, 2006). Often, individuals who are named by their peers are considered to be sources of influence and are located centrally in these sociograms. However, two researchers might view these sociograms differently and, therefore, would come to different conclusions about whom they perceive to be the most influential. In order to combat the subjective interpretation of sociograms, statistical models have been employed in an effort to understand the networks of influence.

We continue our work in this vein by considering the endorsement of one's influence by peers to be an indicator of influence in an organization. We build on the work of previous researchers by attempting to identify those individuals whose influence is extraordinary. While models of contagion (Marsden & Friedkin, 1993) or selection (Frank, 1998) have been demonstrated to enhance our understanding of influence in organizations, their purpose is not to identify the most influential individuals. We now focus on our primary research question "Who are the most influential individuals in an organization?" through the lens of prestige measured by social networks.

METHODS

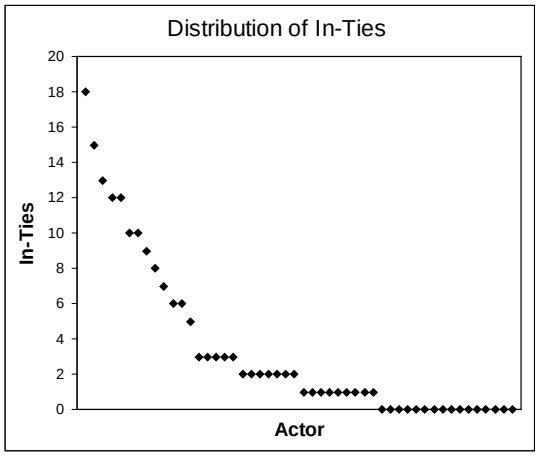
Methodological Considerations in the Identification of Influential Individuals

As stated earlier, social network surveys are used to obtain information about the communication patterns among individuals in an organization. Given a set of directed communication relations among the actors in an organization, it is possible to create a measure of prestige centrality for each individual. We use Freeman's (1979) in-degree centrality (also known as in-ties) as a measure of the total communication directed at each individual. For each actor, the in-tie measure indicates the number of people who say that they are connected to him/her. Individuals who receive a great number of peer nominations are considered the most prestigious actors, with high influence in their organizations (Moreno, 1934).

In order to visualize the distribution of the in-degree measure within an organization, the in-degree scores for all individuals can be sorted in descending order and then graphed. The result is a "scree" plot, as shown in Figure 1. This technique allows a researcher to get a sense of the distribution of in-ties for all the actors in the network. In strongly centralized networks, the scree plot will contain a few relatively high scoring individuals and will quickly drop off and plateau close to the x-axis. Alternatively, in an organization where influence is more "distributed," there will not be such a precipitous drop off in the in-tie scores when viewed from left to right, rather there will be a gentler decline in the slope.

One objective of this paper is to establish a defensible cut-point for identifying the most influential individuals in an organization.

Figure 1. Scree Plot



Given the plot in Figure 1, one could attempt to identify a point of inflection and label those individuals whose scores lie above the inflection point as influential (Cattell, 1966). However, such a subjective method of analyzing the scree plot can produce different results based on an individual's viewing of the plot, one's goal in identifying people as influential, and on a host of other contextual factors. In addition, a visual identification of the scree plot's cut-point can be particularly difficult to obtain when there are multiple points of inflection, or if there is a smooth curve. In these cases, a researcher may be tempted to set the cut-score at the number of influentials warranted by his or her research agenda, leading to potential researcher bias.

As an alternative, we investigate four reproducible methods for categorizing influential individuals in an organization. In reviewing these four methods, we hope to provide researchers with a useful set of tools that can be utilized to analyze social network data in the identification of influential individuals in an organization. A brief description of each method is provided in Table 1 below, and a more detailed explanation of these methods follows.

Method 1 - Absolute Cut Score

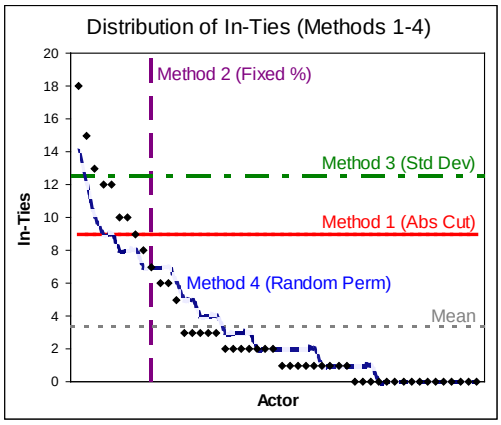
The simplest and most intuitive method for determining a cut score is to set a predetermined absolute criterion above which individuals are deemed influential and below which they are not. For example, if anyone in a school has an in-degree score greater than nine (i.e. at least nine teachers have turned to them for advice), then they are considered "influential."

Table 1. Method Names and Descriptions

Method	Name	Definition
1	Absolute Cut Score	Individuals identified as influential if in-degree score greater than a priori specified level
2	Fixed Percentage of Population	Individuals identified as influential if in highest percentage of population (specified a priori)
3	Standard Deviation	Individuals identified as influential if in-degree score is sufficiently greater than the average in-degree of the rest of the individuals in the network
4	Random Permutation	Individuals identified as influential if in-degree score is significantly greater than what would occur under a random distribution of communication

Graphically, this can be accomplished by superimposing a horizontal line over the in-degree scree plot, as indicated by Method 1 in Figure 2. Those individuals whose influence scores are above the red horizontal line are then categorized as influentials.

Figure 2. Method Comparisons



The Absolute Cut Score is based entirely on a single point and is determined independent of variation in the distribution of in-ties. As such, this is the only method where every individual (or no individual) in a network can potentially be deemed influential since the criterion is absolute, not relative.

Method 2 - Fixed Percentage of Population

An alternate method of identifying influentials is to select a fixed percentage of the population as influential (Spillane et al., 2006; Valente & Pumpuang, 2007). Unlike the previous method where a horizontal line was superimposed over the scree plot, in this method a vertical line is used. If the top 20% of individuals in an organization are to be categorized as influentials, this is equivalent to selecting the leftmost 20% of the individuals in the graph. Those individuals to the left of the violet vertical line in Figure 2 represent the top 20% of individuals.

As with the Absolute Cut Score, this method identifies individuals as influential independent of the variation of in-ties. Method 2 ensures that a given percentage of individuals in the organization are identified as influential and their identification is based upon their performance relative to the performance of other individuals in the organization.

Method 3 - Standard Deviation

Unlike the first two approaches, the Standard Deviation method focuses on the variation in the distribution of ties. This procedure requires users first to calculate the mean and standard deviation of the number of in-ties. Then the researcher creates a horizontal line *x* standard deviations above the mean, which can be superimposed over the scree plot. This horizontal line approach is similar to the Absolute Cut Score (Method 1), however, the Standard Deviation method does not choose a cut score *a priori*, instead it utilizes the observed data in determining where to set the cut point. In Figure 2, the grey dashed line represents the mean, and the green dashed line is two standard deviations above the mean. Under this method, those individuals whose in-degree scores are two standard deviations above the mean are titled "influential." This method is useful in comparing the in-degree scores of each teacher against the "average" level (grey dashed line) of influence in the organization in standard deviation units (Spillane et al., 2006).

Method 4 - Random Permutation Method

Through the use of random permutations, Method 4 produces results which identify those individuals who received significantly more in-ties than would have occurred by chance alone. This method capitalizes on the creation of a sampling distribution of potential networks that could have occurred, conditional on the fixed row marginals.[1]

[1] See Snijders (1991) for a detailed exploration of a simulation method for (0,1) matrices with fixed marginals.

In order to obtain a sampling distribution of influence for the network, the survey respondents' out-ties are randomly reassigned to individuals in the network. Once all of the ties in the network are randomly reassigned, individual influence scores are recalculated according to the in-degree measure described earlier. Both influence distributions (actual and random) are sorted and each individual's actual score in the original dataset is compared with the influence score of the individual of the same rank in the random dataset. Visually, this is represented by the blue line in Figure 2.

Borrowing from parallel analysis (a method used to identify the number of components to extract from a correlation matrix in exploratory factor analysis), one-thousand permutations of random out-tie allocations are performed in order to create a sampling distribution of influence under conditional independence (Buja & Eyuboglu, 1992). The ties are not completely independent, as we restrict their new random locations to only emanate from their original sources in the actual data (i.e. the row marginals are fixed). However, in forcing this restriction, we are able to create a sampling distribution of influence that is comparable to our actual data. The result is the distribution that would arise by random chance, given the set of survey responses, and therefore can be used to identify those individuals whose influence is statistically greater than random chance.

We are interested in measuring how the observed influence scree plot line compares with the sampling distribution of simulated scree plot lines. In doing so, we are able to create a metric of statistical significance. If an individual's actual influence score is higher than his/her ranked counterpart for 95% of the random iterations, then the individual is labeled a "significant influential" at the $p < .05$ level ($\alpha = .05$). We also require that an individual can only be labeled as influential if all other actors in the network with higher in-degree scores have also been identified. This condition ensures that only the individuals with the highest in-degree scores are identified.

This method builds on the work of Bonacich, Oliver, and Snijders (1998) through the use of random permutations and fixed marginals. In their work, both the row and column marginals were held fixed for the random permutation process, and eigenvector centrality scores were calculated for both the observed and simulated data. However, in their work, they compared each actor to him/herself in the random process, whereas we compare each actor with their randomly ranked counterpart, the appropriate comparison for this case of simulating with fixed row marginals only. Furthermore, we use this comparison process to identify individuals with extraordinary in-degree scores and not solely for the purpose of normalizing centrality.

Considerations for Method Selection

Each of the four methods discussed has both merits and drawbacks. The selection of a method for identifying influential individuals is a highly contextualized decision. It will generally depend upon the audience to whom the results will be presented, the theoretical construct which is being measured, and the purpose of the categorization. Another consideration for method selection is whether or not the researcher is interested in making comparisons across multiple organizations, comparisons across multiple networks within a single organization, or simply identifying influentials in a single context. Since networks differ both within organizations (with respect to density, described below) and across organizations (with respect to network size, density, etc.), it becomes important to consider alternate influential identification methods that are appropriate for different research questions.
We now turn our attention to a series of considerations in selecting an appropriate method. We first outline the general considerations when selecting a method and then we review the strengths and weaknesses of each method with respect to these general considerations.

Parsimony

There is a natural desire on the part of the public and the educational establishment for simple methods, understood by many, to be used in educational research. The methods described above vary in terms of their complexity and how easily they can be understood. If the results of the influential categorization are to be shared beyond the audience of educational theoreticians, then it is essential that the cut-point method be both easily understood and appear to be reasonable (Anastasi, 1988). In contrast, there are many instances when the intended audience is researchers only, in which case ease of explanation may be of lesser concern.

Theoretical Merit

While all four methods are at times theoretically defensible, selecting an appropriate method depends upon the application and the literature surrounding that application. Does the researcher prefer a cutoff criteria that is based on how influential an individual is relative to his/her peers (Methods 2-4), or how influential s/he is compared to an absolute criterion (Methods 1). Does the literature support a definition of influence as being sought out for communication a specific number of times? Does the organizational literature in this field suggest that all organizations of a particular type have *x* percent influential individuals? Is the researcher interested in whether or not the most influential people are being sought after more frequently than chance would suggest? These types of questions help to determine the theoretical merit of each method, given a particular usage.

Network Size

Thesize of the networks in consideration (i.e. the number of individuals in the network) becomes particularly important when conducting inter-organizational analyses. Differing network sizes will impact the number of identified influentials in an organization, by making it relatively easier or harder to be deemed influential. An individual has more opportunity to receive nominations in a larger organization than in a smaller organization, based on the size of the network alone. In a small organization, the ability to influence very few individuals may make you a relatively influential individual. As a result of these nuances, making meaningful comparisons regarding the number of influentials across networks becomes increasingly challenging as inter-organizational variation in network size increases. Some researchers may believe that in order to be influential, an individual must exceed a certain fixed threshold of influence, regardless of the size of the organization. Other researchers may believe that the threshold for being an influential member of an organization changes depending upon the size of the organization. Generally, network size will be positively related to the number of influentials identified according to any of the four methods, so researchers must be careful in making cross organizational comparisons regarding the number or percentage of influential individuals.

Network Density

One concern which relates to the relative or absolute nature of the cut-score is how each method accounts for the central tendency, or density, of the data. The density of a social network is a simple proportion of the amount of communication that occurs to the maximum amount of communication that can potentially occur. Density is often related to both network size and response rate. Generally, the larger the network's size the lower the density, and the lower the response rate the lower the density. Each method's strategy regarding density (and therefore network size and non-response) is important to consider, because these factors impact the number and proportion of individuals who are identified as influential. Increased density can only increase the number of influentials identified under Method 1. The relationship between density and the number of influentials identified should be less strong according to Methods 2-4, as these tend to focus

on an individual's influence compared to the influence of other individuals in their network.

Stringency of Researcher Cut Point

Each of these methods requires the researcher to determine a stringency level for a cut point. The researcher must specify the number of ties required for the Absolute Cut Score, the percentage of individuals to identify for the Fixed Percent method, the number of standard deviations above the mean in the Standard Deviation method, and the α level for the Random Permutation method. While the selection of a more stringent cut point will always lead to the identification of equal or fewer influential individuals, the significance of this decision may depend upon the method.

Comparing the 4 Methods

One commonality among the 4 methods is that they all utilize relational social network data alone to identify influential individuals and are thus blind to the positions and titles (attributes) of the individuals in the network. As such, the results obtained through these methods allow researchers to make decisions that reflect school communication measured independent of the formal organizational structure. This is a strength of all four of the methods. However, the methods differ according to the ways that they handle the other considerations mentioned above. Strengths and weaknesses associated with each method are briefly explored here.

Method 1 - Absolute Cut Score

Strengths:
This method is easy to explain, and is therefore optimal for presentation to a non-technical audience. If used to make comparisons regarding the number of influentials across networks or organizations, it uses the same cut point consistently to make that decision. As such, unlike the other 3 methods, a researcher will never have to explain why a person with an in-degree score of 8 was deemed influential in one network, while a person with in-degree score of 10 was deemed not influential in a different network.

Weaknesses:
The worth of this method depends upon the availability of a clear and theoretically justifiable cut-point. However, use of this method will likely involve an arbitrarily selected cut score, so the soundness of the results will rely on the success of defending a subjective criterion. In addition, the results yielded when using an absolute cut score are heavily impacted by the size and density of the network. As mentioned previously, these factors will affect inter-organizational comparisons. In situations where an individual is sought out for advice relatively often but due to small network size or low density does not meet the a-priori cut-point for in-ties, s/he will not be recognized even though his/her influence is relatively large. If one chooses to use the *a-priori* determined cut score for inter-organizational comparisons, one must consider the impact of varying response rates, network densities, and network sizes.

Method 2 - Fixed Percentage of Population

Strengths:
Like Method 1, this method is extremely easy to explain and therefore scores highly in terms of parsimony. This method could be useful for making comparisons among networks or among organizations if a researcher is interested in comparing the characteristics of the top *x* percent of influentials in network *a* vs. network *b*, or organization *a* vs. organization *b*.

Weaknesses:
This procedure essentially pre-specifies the number of influential individuals in a network as a percentage of the population. The theoretical validity of this method is low, since it is difficult to justify why a given percentage of the members of any network are influential by definition. In spite of this drawback, this method allows researchers to set the number of individuals who will be identified as influential, regardless of the composition of the data. It is important to recognize that this method assumes

that there must be some influentials in an organization, as the given percentage is always recognized. The Fixed Percentage method differs from the Absolute Cut Score in a major way: unlike the Absolute Cut Score, the Fixed Percentage ignores the central tendency of the data. If a network doubles in density, the Fixed Percentage of the Population method will yield the exact same number of influentials.

Method 3 - Standard Deviation

Strengths:
This method is also relatively easy to explain. Unlike the first two methods, it considers the variation of in-degree in its identification procedure by using the observed network data. Also, it may be appealing to describe influential individuals as "those whose peer-endorsed influence is sufficiently higher than the average".

Weaknesses:
Use of the mean and standard deviation might seem tempting, but the distribution of in-ties tends to be positively skewed, with the modal value typically 0 in-ties. Readers unfamiliar with this underlying distribution might wrongfully expect to see 2.5% of individuals in an organization identified as influential if their in-tie scores are approximately 2 standard deviations above the mean, as would be expected under a normal distribution. As such, it is imperative to caution against any comparisons with a typical normal curve. Furthermore, the number of influentials identified by this method is unaffected by the magnitude of the in-degree scale. That is, if all network members increase their in-degree by 10, the same number of individuals will be identified as influential.

Method 4 - Random Permutation

Strengths:
By capitalizing on the concept of sampling distributions of ties conditional on row marginals, this method affords the researcher the term "statistical significance." For the research community, this may be the most appropriate method for determining a cut point for intra or inter-organizational comparison. This technique holds the underlying characteristics of the network constant and establishes a baseline against which one can compare the observed influence distribution to see if it is truly different from what one would expect to see by chance. As such, the results of this method, as compared with the other three methods, are less dependent on non-response, network size, and density.

Weaknesses:
This method is the most complicated of the four discussed, and therefore scores poorly with respect to parsimony. Furthermore, the "practical" significance of being identified as an influential under this method is unclear. Using the Absolute Cut Score method, all individuals were considered influential by virtue of having been the recipient of conversation a given number of times. With this method, the cut point for influential selection is different for each person; it depends on how each person's randomly ranked counterpart scores on their respective in-ties.

Methods Selection Conclusion

The four methods described above differ in terms of how they operationalize a cut point for influential identification. A separate measurable construct could be used for influential identification, such as eigenvector centrality (Bonacich, 1972), brokerage opportunities (Burt, 1995), or individuals who bridge unconnected alters (Granovetter, 1972). We chose to use in-degree centrality as it is a relatively simple construct and is based on the literature correlating prestige with influence. Those individuals who are centrally located in an organization's communication network have greater control over information and influence than those on the outskirts (Brass, 1992; Burt, 1982).

As all four methods operate on the same in-degree measure, there is significant overlap in many of the individuals identified across the

different methods. We explore the utility of the Random Permutation method as it pertains to uncovering the most influential individuals in a sample of high school staff members, as defined through peer-endorsement.

METHODS

In this section we empirically compare the four methods of identifying organizational influentials in a sample of nine high schools. Prior to these analyses we first provide a brief description of our research sites, data collection instrument, and the unique measure of peer endorsed influence used in the analyses.

Research Sites and Participants

The survey social network data used in this analysis comes from a Consortium for Policy Research in Education (CPRE) study of high school reform. The sample consists of a collection of nine high schools across the country that were working with five external assistance providers during the 2005-2006 school year. The external reform organizations, High Schools That Work, First Things First, RampUp, Penn Literacy Network, and SchoolNet were selected as representative of the types of external assistance found in high schools during previous research (Gross & Goertz, 2005). Each external reform organization supplied the names of two representative schools that had used their program for one or two years. Data were not collected from the school in its second year of implementing First Things First (FTF2) due to a natural disaster.[2]

At each of the nine sites, a survey was conducted with all teaching staff. The survey provides information on the communication networks that exist within schools. Our findings are based on 730 surveys returned, with individual school response rates ranging from 48 to 83 percent (Overall response rate = 67%). These response rates reflect the percentage of the instructional staff that completed a survey; surveys were given to teachers, administrators, instructional coaches, and any other staff member who spent the majority of his/her time in an instructional capacity with students.

Low response rates can be extremely problematic for studies interested in examining contagion, or large network structures (Burt, 1982). In this study, the moderate response rate is sufficient for our research question, as we are interested in examining the sources of influence captured through in-degree. In a simulation study, Costenbader & Valente (2003) demonstrated that in-degree centrality is extremely robust, even in cases of severe non-response (over 50% non-response), given that the data are missing at random. While we are confident that our response rates are sufficiently high to perform these analyses, we cannot be certain that our non-respondents are systematically different from the respondents, which may be a limitation in our results.

The nine schools surveyed vary in size from as small as 36 teachers (415 students) to 220 teachers (4,778 students). In addition, substantial variability exists among schools in both their percentages of students receiving free or reduced-price lunch (from 0% to 99%) and their racial/ethnic compositions.

Data Collection and Instrument

The survey instrument that was used to generate our social network data contains five network questions, three of which focus on professional practice, one on friendship, and one on communication about the external reform organizations (referred to as the "reform" network). For each of the network questions, survey respondents could list up to five individuals (without the aid of a roster) to whom

[2] For the purposes of anonymity, the school sites are named by their reform type and length of implementation in the school. Accordingly, FTF1 is a school that had just begun its partnership with the FTF reform program, HSTW2 indicates a school that was in its second year of work with HSTW at the time of survey administration.

they went to for help during the school year.[3] The five network questions are included in Table 2. Due to the purposeful advice nature of the questions, the survey implies a directionality of influence in all but the friendship network. As such, being designated as the recipient of requests for help serves as peer-endorsement of the influence of the recipient of the tie.

Table 2. Description of Five Networks

Network Label	Survey Question
Classroom Management	To whom, in your school, have you turned to for advice about classroom management during this school year?
Course Content and Planning	During this school year, to whom in your school have you gone for help in selecting and planning course content coverage and pacing?
Low Performing Students	During this school year, to whom in your school have you turned for advice on strategies to assist low performing students?
Reform	Please list the people inside or outside your school to whom you turned for advice in using [reform name] during this school year.
Friendship	During this school year, with whom among your colleagues at this school do you "hang out" and discuss family, home, and/or personal issues?

[3] Therefore, as in many network surveys, there is an artificial cap on the maximum density of our networks. Limiting the potential responses to five colleagues may have impacted the data collected. However, few of our respondents provided up to five names, suggesting that the limit did not constrain the potential information. Fewer than 13% of respondents provided either four or five names in the 2006 survey.

These expressive connections are often stronger ties than work related, or "instrumental" ties, and have been shown to influence the information obtained through instrumental links (Frank et al., 2004; Ibarra & Andrews, 1993; Uzzi, 1997).

By describing school communication with multiple instrumental and expressive relations, we feel that we have painted a fuller picture of the influence avenues permeating high schools than would have been possible by simply examining professional conversations.

Responses to the above prompts include the name of the advice-giver, the frequency with which advice was sought from this helper, and the influence of advice from this helper.
The responses to the frequency stem "How often have you sought guidance from this person?" were on a 4 point scale (recoded to represent an approximate number of days in a school year: "Daily or almost daily" = 150, "Once or twice a week" = 40, "Once or twice a month" = 20, "A few times a year" = 2). The responses to the influence stem of "How influential is the advice of this person …" were also on a 4 point scale (operationalized to represent the proportion of conversations that respondents reported having influence on their practice: "Highly Influential" = 100%, influential = 70%, "Slightly Influential" = 40%, "Not influential" = 10%).

Due to the fact that the survey respondents were able to associate a frequency and a level of influence with the requests for help, strength of the relationship was built into the survey, and thus, the ties between individuals could be assigned a weight. By including the influence component to the weighting procedure, we felt that the face validity of our methodology was enhanced: in attempting to answer the research question "Who are the most influential individuals in schools?," we actually asked the school staff members who they felt were influential to them (Brass, 1984).

Measures

In an attempt to maximize the information about the tie between the survey respondent and the influential, we combined the frequency and influence constructs by multiplying the respective recoded scores. By multiplying the frequency score (number of conversations in the school year) by the influence score (proportion of influential conversations), we created a metric of the number of highly influential conversations on the survey respondent's practice. This new metric allowed respondents to describe the nature of their communications with respect to how they influenced their respective practice. For example, this metric would only produce large scores for influence if the recipient was spoken to frequently and if the content of the communication was influential. Frequent but non-influential conversations and non-frequent, but highly influential conversations were given less weight due to this recoding.[4] Thus, the final weighted tie from sender to receiver is an indicator of the frequency and influence of communication.

To eliminate the positive skew of the product, we took the natural log of the result combined with a small constant to produce non-negative results (Frank et al., 2004). The final possible tie strengths took on 16 values approximately in the range (1,5), according to the formula:

$$freqinf_{ij} = ln(frequency_{ij} * influence_{ij} + k)$$

Individual *j*'s peer-endorsed influence score (or weighted in-degree) was calculated by summing the *freqinf* scores assigned by all other individuals to *j*, where a non-relation was considered to have *freqinf* = 0.

Therefore:

$$influence_j = \sum_{i \neq j} freqinf_{ij}$$

As a result, each individual in a school is associated with an influence score for each of the five networks that is a function of the responses from all responding teachers to that particular social network section of the survey. This total influence score is an indicator of the peer-endorsed influence of an individual.

Empirically Comparing the Influential Identification Methods

In order to compare the four different methods of leadership identification, we examined the five networks of communication in the nine study schools. The number of school staff identified as influential individuals in all five networks, nine schools, and under each of the four methods, are available in Appendix A. For each of the four methods, we selected two different criteria for the cut-points in order to illustrate the impact of such decisions in the identification process. We created these cut-points to illustrate generous and stringent requirements for the identification of influential individuals.

Under the Absolute Cut Score method we considered influential those individuals with an influence score (weighted in-degree) greater than 5 (low stringency) or 10 (high stringency). Using the Fixed Percentage method we selected as influential the top 10% of the population (low stringency) and the top 5% of the population (high stringency) with respect to their influence scores. For the Standard Deviation method we set the cut score at 1.0 and 1.5 standard deviations above the mean influence score for low and high stringency levels, respectively. Finally, using the Random Permutation method we selected α levels of .50 and .05 as low and high stringency respectively.

In this section we explore the relationship between the number of influentials identified

[4] Note that infrequent but highly influential conversations receive similar weight to frequent but low-influential conversations. While the weighting calculation used in the analyses included in this paper place slightly greater emphasis on frequency, in separate analyses other weighting calculations were used yielding similar results.

and several factors: network size, network density, and the stringency of the user defined cut point. To illustrate the ways that these variables relate to the number of influentials identified, correlations coefficients were calculated[5].

Number of Influentials and Network size

In general, under all four methods there was a positive relationship between network size (school size) and the number of individuals identified as influential. This relationship is a perfect correlation under method 2 due to its definition. In method 3, there is a strong correlation (average *r=.78*) between school size and number of influentials identified. Although still positively correlated, this relationship was less strong for method 1 (average *r=.38)*. In contrast, under method 4 there were less consistent results, with both positive and negative correlations found. The consistently positive correlations (under methods 1-3) imply that as network size grows, the number of influentials identified increases; however, the strength of this relationship is not uniform across all methods.

Number of Influentials and Density

The relationship between density and the number of influentials identified can be somewhat murky since density is negatively correlated with network size. We examine the correlation between density and number of influentials across the five networks *within each school* in order to hold network size constant. As expected, there is a very strong positive correlation (average *r* =.93) between the number of influentials identified and network density under method 1. Since under method 2 there is no variation in the number of influentials identified within a school across networks, these correlations are incalculable. Under method 3 we observed a moderate correlation between density and number of influentials (average *r=.53)*[6]. A moderate correlation between density and number of influentials was also observed under method 4 using the low stringency cut point (average *r=.56*); however, we found no consistent relationship between density and number of influentials using the more stringent cut point under method 4 (average *r=-.04)*. Similar to the results found for network size, increased density appears to be positively associated with the number of influentials identified in methods 1 and 3.

Number of Influentials and Stringency of Cut point

The selection of a user defined cut point also has an impact on the number of individuals identified as influential. As the cut point becomes more stringent the number of influential individuals identified decreases. The reduction in the number of influentials between the less stringent and more stringent cutoff can be observed by comparing adjacent columns under each method in the table in Appendix A. The significance of this relationship seems relatively straightforward under methods 1 and 2, where the researcher's decision regarding the cut score has a fairly large impact on the number of individuals identified. Under method 3 we find that the difference between the number of influentials identified under our 1 and 1.5 standard deviation cut points were not very large. This reflects the characteristics of the observed skewed distribution, where few observations lie in the area between 1 and 1.5 standard deviations from the mean. Had we selected cut scores at 0.5 and 2.0, then the differences between the low stringency and high stringency cuts would have been far greater.

The relationship between the researcher's cut point and the number of influentials identified under method 4 is less consistent. In the low performing students network and the reform network the researcher decision had little

[5] Due to the extremely small sample size (subsequent examples will contain as few as five observations), p-values are not reported. Tables of these correlation coefficients are available from the authors.

[6] Note: three out of the eighteen correlations were small and negative.

influence on the number of influentials identified, whereas in the three other networks the researcher's decision had a larger impact on drop-off. Generally, the researcher's selection of a cut point criterion impacts the number of influentials identified.

An Application Using the Random Permutation Method

In the beginning of this paper, we posed the question "Who are the influential individuals in high schools?" This research question was investigated using the Random Permutation method ($\alpha = .05$) to identify influential individuals in this sample of high school staff members. Detailed descriptions of these results are available in Supovitz (2007). However, these substantive results are not presented here for the sake of brevity.

DISCUSSION

In this manuscript, we applied four different perspectives on how to define influentials and found that different methods resulted in different numbers of individuals identified. In this concluding section, we inspect network factors that might have contributed to how the methods differed in terms of identifying individuals. We conclude with the limitations to this study suggesting avenues for future research throughout.

Network factors that Influence Method Results

All of the four methods described above are in some way impacted by network structural factors, including network size and density. We point out these factors to alert the research community to the difficulties in disentangling network structure when making inter (or intra) organizational comparisons of influentials. In our Results section, we examined how these variables related to the number of influentials identified, but did not examine how these structural variables co-vary as they relate to the outcome of interest.

In an exploratory analysis, we created regression models where network structural variables, researcher cut points, and cluster variables (school and network type) were used to predict the number of influential individuals according to each method. By including all of these variables in a single model, we were better able to partial out the impact of each variable on the number of influentials identified, after controlling for other highly correlated variables. The results we obtained from these regressions comport well with the simple correlational results, but we feel that more research in this area is necessary. The use of simulated data for such an investigation is advised, as it would afford the researcher control over the variables of interest, and would therefore allow for results that are not necessarily unique to a particular set of actual network data.

Limitations/Future Research

The results in this paper tended to focus on how to identify the most influential members of an organization and the characteristics of these individuals. We chose not to explore how the unique context of each school helps explain the distribution of influence across organizations. For example, we ignored the fact that each of our schools was working with an external reform agency. Future research might investigate the ways that different reforms impact communication patterns and therefore the distribution of influence in schools. While we have called for the use of simulated data to examine the relationships between influentials and network structural variables, an examination of the relationship between school contextual variables and influentials using actual data would produce a substantive contribution to the literature. In addition, our analysis focused on four advice seeking and a single friendship network. Had we incorporated alternate network questions into our survey, we would likely have uncovered different substantive findings regarding the nature and characteristics of influential members of schools. In future research, it may be of interest to apply these identification methods to different network

questions to paint an alternate picture of school influentials.

High schools are organized into academic departments, often celebrated as the foundation for professional communication networks (Siskin, 1994; Stodolsky & Grossman, 1995). In our analysis, we ignored the role of department as it relates to influence, as we wanted our methods to be blind to both the attributes of the advice seeker and advice giver. However, it has been demonstrated that this organizational structure creates "homophilous" conversation (McPherson, 1987), where members of the same department are more likely to interact than members across department (Weinbaum, Cole, Weiss, & Supovitz, 2008). Future researchers are encouraged to apply these methods within department (or small learning communities, focus teams, etc.) in order to find the most influential members within these organizational groupings.

Another avenue of further research in the identification of key individuals in an organization would be to find the most efficient way to disseminate information through a network, by targeting appropriate individuals. When looking to optimally spread information in an organization one must consider the extent of overlap among those individuals who are influenced by the influentials. If the two most influential people in an organization have a high level of redundancy with respect to whom they influence, it may be unnecessary to identify both individuals as influential. Much of the literature points to naturally occurring subgroups (Frank, 1995), or cliques, that exist in an organization as appropriate targets for dissemination. These tightly knit cliques are defined by having more within-group conversation than outside group conversation. As such, centrality in the entire organization might not be the most appropriate method for identifying individuals in the hopes of disseminating programs; rather, centrality within each clique might be the optimal strategy.

This paper is intended to make both a methodological and substantive contribution to organizational researchers and administrators concerned with identifying influence in organizations. We are hopeful that future researchers will consider these methods as mechanisms for identifying sources of influence, both in schools and other social organizations.

REFERENCES

Anastasi, A. (1988). *Psychological testing*. New York, NY: Macmillan.

Birk, S. M. (2006). Application of network analysis in evaluating knowledge capacity. *New Directions for Evaluation, 107*, 69-79.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology, 2*, 113-120.

Bonacich, P., Oliver, A., & Snijders, T. A. B. (1998). Controlling for size in centrality scores. *Social Networks, 20*(2), 135-141.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly, 29*(4), 518-539.

Brass, D. J. (1992). Power in organizations: A social network perspective. In G. Moore, & J. A. Whitt (Eds.), *Research in politics and society* (pp. 295-354). Greenwich, CT: JAI Press.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509-540.

Burt, R. S. (1982). *Toward a structural theory of action*. New York, NY: Academic Press.

Burt, R. S. (1995). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Cattell, R. B. (1966). Scree test for number of factors. *Multivariate Behavioral Research, 1*(2), 245-276.

Cole, R. P., & Weinbaum, E. H. (2007). Chain reaction: How teacher communication influences attitude. *Paper Prepared for Presentation at the Annual Meeting of the American Educational Research Association*, Chicago, IL.

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

Fowler, F. (2004). *Policy studies for educational leaders* (2nd Edition ed.). Upper Saddle River, NJ: Pearson.

Frank, K. A. (1995). Identifying cohesive subgroups. *Social Networks, 17*(1), 27-56.

Frank, K. A. (1998). Quantitative methods for studying social contexts in multilevels and through interpersonal relations. *Review of Research in Education, 23*, 171-216.

Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education, 77*(2), 148-171.

Freeman, L. C. (1979). Centrality in social networks I. conceptual clarification. *Social Networks, 1*(3), 215-239.

Granovetter, M. (1972). The strength of weak ties. *American Journal of Sociology, 78*(1), 1360-1380.

Gross, B., & Goertz, M. E. (2005). *Holding high hopes: How high schools respond to state accountability policies* No. CPRE Research Report No. RR-056). Philadelphia PA: University of Pennsylvania, Consortium for Policy Research in Education.

Hart, A. W. (1995). Reconceiving school leadership - emergent views. *Elementary School Journal, 96*(1), 9-28.

Ibarra, H. (1993). Network centrality, power, and innovation involvement: Determinants of technical and administrative roles. *Academy of Management Journal, 36*(3), 471-501.

Ibarra, H., & Andrews, S. B. (1993). Power, social-influence, and sense making - effects of network centrality and proximity on employee perceptions. *Administrative Science Quarterly, 38*(2), 277-303.

Kempe, D., Kleinberg, J., & Tardos, E. (2005). Influential nodes in a diffusion model for social networks. *Automata, Languages and Programming, Proceedings, 3580*, 1127-1138.

Leenders, R. T. A. J. (2002). Modeling social influence through network autocorrelation: Constructing the weight matrix. *Social Networks, 24*(1), 21-47.

Marsden, P. V., & Friedkin, N. E. (1993). Network studies of social-influence. *Sociological Methods & Research, 22*(1), 127-151.

McLaughlin, M. W. (1990). The rand change agent study revisited: Macro perspectives and micro realities. *Educational Researcher, 19*(9), 11-16.

McPherson, J. M. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American Sociological Review, 52*(1), 370-379.

Moreno, J. L. (1934). *Who shall survive? foundations of sociometry, group psychotherapy, and sociodrama*. Beacon, NY.: Beacon House, Inc.

Riggan, M., & Supovitz, J. A. (2008). Interpreting, supporting, and resisting change: The geography of leadership in reform settings. In J. A. Supovitz, & E. H. Weinbaum (Eds.), *The implementation gap: Understanding reform in high schools* (pp. 103-125). New York, NY: Teacher's College Press.

Schneider, B. (2005). The social organization of schools. In L. V. Hedges, & B. Schneider (Eds.), *The social organization of schooling* (1st ed., pp. 1-12). New York, NY: Russell Sage Foundation.

Siskin, L. (1994). *Realms of knowledge: Academic departments in secondary schools*. London: Falmer.

Snijders, T. A. B. (1991). Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika, 56*(3), 397-417.

Spillane, J. P. (2006). *Distributed leadership.* San Francisco, CA: Jossey-Bass.

Spillane, J. P., Camburn, E., Lewis, G., & Pareja, A. S. (2006). Taking a distributed perspective in studying school leadership and management: Epistemological and methodological trade-offs. *Paper Prepared for Presentation at the Annual Meeting of the American Educational Research Association*, San Francisco, CA.

Stodolsky, S. S., & Grossman, P. L. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal, 32*(2), 227-249.

Supovitz, J.A., (2007). Instructional Influence in American High Schools. In M.M. Mangin & S.R. Stoelinga (Eds.), *Effective Teacher Leadership: Using Research to Inform and Reform.* (pp. 144-162). New York, NY: Teachers College Press.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly, 42*(1), 35-67.

Valente, T. W., & Pumpuang, P. (2007). Identifying opinion leaders to promote behavior change. *Health Education & Behavior, 34*(6), 881-896.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Weinbaum, E. H., Cole, R. P., Weiss, M. J., & Supovitz, J. A. (2008). Going with the flow: Communication and reform in high schools. In J. A. Supovitz, & E. H. Weinbaum (Eds.), *The implementation gap: Understanding reform in high schools* (pp. 68-102). New York, NY: Teachers College Press.

Appendix A. Number of Influential Individuals by Method and Network

			Method 1		**Method 2**		**Method 3**		**Method 4**	
			Absolute Cut Score		*Fixed % of Population*		*Standard Deviation*		*Random Permutation*	
School	**n**	**Density*100**	>5	>10	*10%*	5%	*>1sd*	>1.5sd	*>50th*	>95th
					Classroom Management Network					
RU1	44	2.7	10	3	4	2	6	3	14	3
PLN2	66	1.9	12	4	6	3	8	7	7	0
SN1	80	1.5	20	7	8	4	11	5	20	2
HSTW1	102	1.4	31	4	10	5	12	5	4	3
RU2	109	1.2	24	10	10	5	10	7	20	12
HSTW2	120	1.5	43	17	12	6	15	9	24	17
PLN1	126	0.8	24	2	12	6	20	14	31	2
FTF1	173	1	61	24	17	8	26	19	36	31
SN2	265	0.3	32	7	26	13	24	17	24	20
					Course Content and Pacing Network					
RU1	44	2.2	8	1	4	2	8	4	14	0
PLN2	66	2.2	18	9	6	3	9	7	16	13
SN1	80	0.9	11	0	8	4	13	10	19	0
HSTW1	102	1	26	4	10	5	17	9	34	0
RU2	109	0.9	15	5	10	5	6	5	10	9
HSTW2	120	1	32	12	12	6	19	11	26	20
PLN1	126	0.7	19	2	12	6	19	12	13	0
FTF1	173	0.9	62	18	17	8	24	16	44	19
SN2	265	0.3	29	8	26	13	30	17	24	2
					Strategies for Assisting Low Performing Students Network					
RU1	44	2.9	10	3	4	2	7	3	10	10
PLN2	66	2.2	13	6	6	3	7	4	11	8
SN1	80	0.9	12	5	8	4	10	6	12	12
HSTW1	102	1.1	18	4	10	5	10	5	5	4
RU2	109	0.8	15	7	10	5	11	7	15	14
HSTW2	120	0.9	23	12	12	6	17	9	22	19
PLN1	126	0.6	15	5	12	6	14	9	16	15
FTF1	173	0.8	41	15	17	8	16	10	20	17
SN2	265	0.2	16	4	26	13	17	4	4	4
					Reform Network					
RU1	44	1.3	2	0	4	2	7	4	3	0
PLN2	66	0.4	3	0	6	3	5	5	5	5
SN1	80	0.6	3	2	8	4	3	3	3	3
HSTW1	102	1.2	25	2	10	5	6	2	2	1
RU2	109	0.6	8	3	10	5	3	2	8	6
HSTW2	120	0.6	17	7	12	6	14	12	17	16
PLN1	126	0.1	1	0	12	6	6	6	1	1
FTF1	173	0.5	24	12	17	8	12	7	22	20
SN2	265	0.1	4	1	26	13	18	11	9	9
					Social Network					
RU1	44	2.7	14	6	4	2	7	6	11	0
PLN2	66	3.4	30	18	6	3	10	6	12	6
SN1	80	2.7	39	23	8	4	13	5	28	5
HSTW1	102	2.2	55	23	10	5	17	12	19	0
RU2	109	1.7	48	20	10	5	15	8	15	6
HSTW2	120	1.6	63	30	12	6	24	13	31	0
PLN1	126	1	41	20	12	6	23	15	23	0
FTF1	173	1.3	90	44	17	8	21	18	29	21
SN2	265	0.3	56	8	26	13	37	27	4	1

Node Discovery Problem for a Social Network

Yoshiharu Maeno, Ph.D.
Social Design Group, Tokyo, Japan

A node discovery problem is defined as a problem in discovering a covert node within a social network. The covert node is a person who is not directly observable. The person transmits influence to neighbors and affects the resulting collaborative activities (e.g. meetings) within a social network, but does not appear in any information reported by the intelligence. Throughout this study, the information comes from data that record the participants of collaborative activities. Discovery of the covert node refers to the retrieval of the data and the corresponding collaborative activities that result from the influence of the covert node. The nodes that appear commonly in the retrieved data are likely to neighbor the covert node. Two methods are presented for detecting covert nodes within a social network. A novel statistical inference method is discussed and compared with a conventional heuristic method (data crystallization). The statistical inference method employs the maximal likelihood estimation and outlier detection techniques. The performance of the methods is demonstrated with test datasets that are generated from computationally synthesized networks and from a real organization.

Author: *Dr. Yoshiharu Maeno is a founder management consultant and scientist at Social Design Group. He has developed mathematical methods to reveal the topological structure and to profile the information diffusion in social networks.*

Correspondence: *Contact Yoshiharu Maeno at Sengoku 1-6-38F, Bunkyo-ku, Tokyo 112-0011, or email maeno.yoshiharu@socialdesigngroup.com.*

INTRODUCTION

A covert node refers to a person who transmits influence and affects the resulting collaborative activities among others in a social network, but does not appear in any information reported by the intelligence. Throughout this study, the information is a set of data which record the participants of the collaborative activities. The covert node is not observable directly. Which data (and corresponding collaborative activities) result from the influence of the covert node? This problem is called a node discovery problem for a social network. It predicts the existence of a covert node near those overt nodes whose presence and potential interaction with the covert node are known or suspected. Where do we encounter such a problem?

Globally networked clandestine organizations such as terrorists, criminals, or drug smugglers pose a great threat to civilized societies (Sageman, 2004). Terrorism attacks cause great economic, social and environmental damage. Governments make great efforts in managing the clean-up and recovery from an attack's aftermath. The short-term goal of the efforts is the arrest of the perpetrators. The long-term goal is identifying and dismantling the clandestine organizational foundation that raised, encouraged, and helped the perpetrators. The threat can be mitigated and eventually eliminated by discovering the covert wire-pullers and key conspirators in the clandestine organization. Difficulties arise with the limited capability of the intelligence. Information on the wire-pullers and key conspirators is intentionally hidden by the organization.

Let's take an example from the 9/11 terrorist attacks of 2001 (Krebs, 2002). Mustafa A. Al-Hisawi, whose alternate name was Mustafa Al-Hawsawi, is alleged to have been a wire-puller who had acted as a financial manager of Al Qaeda, had attempted to help terrorists enter the United States, and had provided the hijackers with financial support worth more than 300,000 dollars. Osama bin Laden is likewise suspected of having been a wire-puller behind Mustafa A. Al-Hisawi and the conspirators behind the hijackings. These persons were not recognized as wire-pullers at the time of the attack. They were the covert nodes waiting to be discovered from the information on the collaborative activities of the perpetrators and known conspirators.

In this study, two methods are presented to solve the node discovery problem. The first method is a conventional heuristic method (data crystallization) that was first proposed by (Ohsawa, 2005) to explore a latent structure with dummy variables, and then streamlined by (Maeno, 2009). (Maeno, 2009) demonstrates a simulation experiment of the node discovery problem for the social network of the 9/11 perpetrators. The second method is a novel statistical inference method that is discussed and compared with the first method in this study. The statistical inference method employs the maximal likelihood estimation and outlier detection techniques.

This study is organized as follows. Related works are reviewed briefly. After the node discovery problem is defined mathematically, the two methods are presented. Next, the test datasets are introduced. They are generated from computationally synthesized networks and from a real clandestine organization. The performance characteristics of the methods are then demonstrated by measuring precision, recall, and Van Rijsbergen's F measure (Korfhuge, 1997).

Related Work

Social network analysis is a study of social structures consisting of nodes linked by one or more specific types of relationships. Examples of the relationship are influence transmission in communication or the presence of trust in collaboration (Lavrac, 2007). Network topological characteristics of clandestine terrorist organizations (Krebs, 2002) and criminal organizations (Klerks, 2002) are studied. The trade-off between remaining secret and efficiently exercising coordination and

control is of particular interest (Morselli, 2007). The impact on the network topology of the trade-off is analyzed (Lindelauf, 2009).

Link discovery predicts the existence of an unknown link between two nodes from the information on the known attributes of the nodes and the known links (Clauset, 2008). Link discovery is one of the tasks of link mining (Getoor, 2005). Link discovery techniques are combined with application domain specific heuristics. Collaboration between scientists can be predicted from published co-authorship (Liben-Nowell, 2004). Friendship between people is inferred from the information available on their web pages (Adamic, 2003).

The Markov random network is a model of the joint probability distribution of random variables that can be applied to discover the dependency between links that share a node. It is an undirected graphical model similar to a Bayesian network, and one of dependence graphs (Frank, 1986). The Markov random network has been extended to various models. The models include hierarchical models (Lazega, 1999), models of multiple networks that treat different types of relationships (Pattison, 1999), models of valued networks that include nodal attributes (Robins, 1999), models for higher order dependency between links which share no nodes (Pattison, 2002), and models of 2-block chain graphs that associate one set of explanatory variables with another set of outcome variables (Robins, 2001). A family of such generalized and elaborated models is named an exponential random graph (Anderson, 1999).

In addition to link discovery, the related research topics are the exploration of an unknown network structure (Newman, 2007), the discovery of a community structure (Palla, 2005), the inference of a network topology (Rabbat, 2008), and the detection of an outlier in a network (Silva, 2009). Stochastic modeling to predict terrorism attacks (Singh, 2004) is relevant practically. Machine learning techniques (algorithms allowing computers to learn from databases or sensor data) to infer latent variables from other observable variables (Silva, 2006) are potentially applicable to discovering an unknown network structure.

Node Discovery Problem

Problem Definition

The node discovery problem is defined here mathematically. A node represents a person in a social network. A link represents a relationship that transmits influence between persons. The symbols n_j (j = 0,1,…) represent the nodes. Some nodes are overt (observable), but others are covert (unobservable). $\mathbf{O}$ denotes the set of the overt nodes $\{n_0, n_1, \ldots, n_{N-1}\}$. The number of overt nodes is $|\mathbf{O}| = N$. $\mathbf{C} = \overline{\mathbf{O}}$ denotes the set of covert nodes $\{n_N, n_{N+1}, \ldots, n_{M-1}\}$.

The number of covert nodes is $|\mathbf{C}| = M - N$. The total number of nodes is $|\mathbf{O} \cup \mathbf{C}| = M$. The unobservability of the covert nodes arise either from a technical defect of the intelligence or from an intentional cover-up operation.

The symbol δ_i represents a set of participants in a particular collaborative activity. The set is the co-occurrence pattern among the nodes (Rabbat, 2008) for the i-th collaborative activity. Any subsets of $\mathbf{O} \cup \mathbf{C}$ can form δ_i. For example, if the nodes $n_0, n_1, \ldots$ join in on a particular conference call, the collaborative activity pattern is $\delta = \{n_0, n_1, \ldots\}$. Note that the unobservability of the covert nodes do not affect the collaborative activity patterns themselves.

Data d_i records a set of the overt nodes in a collaborative activity pattern δ_i. Data d_i is a subset of overt nodes in $\mathbf{O}$. It is given by Equation (1).

$$d_i = \delta_i \cap \mathbf{O} = \delta_i \cap \overline{\mathbf{C}} \quad (0 \le i < D) \quad (1)$$

The dataset is represented by $\{d_i\}$. The amount of data is D. Note that neither an individual node nor a single link alone can be observed

directly, but nodes can be observed collectively as a collaborative activity pattern. The dataset $\{d_i\}$ can be expressed by a 2-dimensional $D \times N$ matrix of binary variables, $\mathbf{d}$. The presence or absence of the node n_j in the data set d_i is indicated by the elements in Eq. (2).

$$\mathbf{d}_{ij} = \begin{cases} 1 & (n_j \in d_i) \\ 0 & (n_j \notin d_i) \end{cases} \quad (0 \le i < D, 0 \le j < N) \quad (2)$$

Solving the node discovery problem is defined as retrieving the data which result from the influence of the covert node. The covert node is a member of the set that represents a collaborative activity pattern δ_i when the pattern results from the influence of the covert node. But the covert node is not a member of the set that represents the corresponding data d_i because of the relationship in Eq. (1). The essence of the problem is retrieving all i's for which $d_i \neq \delta_i$ holds.

Influence Transmission

Possible collaborative activity patterns are governed by how influence is transmitted in a social network. The nodes that appear in a collaborative activity pattern are those receiving the transmitted influence. For example, the calling party n_0 wishes to have more than one party $n_1,\ldots$ listen to the telephone call. The influence of n_0 to $n_1,\ldots$ establishes the setup for the conference call, and the resulting collaborative activity among the parties is $\delta_i = \{n_0, n_1,\ldots\}$. Influence transmission in a social network is a stochastic process that is described by probability distribution functions. Influence transmission is initiated at a randomly chosen seed node, and every other node may act as a transmitter to neighboring nodes.

One of the basic stochastic processes applicable to influence transmission is the Galton-Watson branching process (Iribarren, 2009). A neighboring node which receives influence transmitted from a node in the g-th generation is called an offspring node in the $(g+1)$-th generation. A neighboring node which transmits influence to a node in the g-th generation is called an ancestor node in the $(g-1)$-th generation. The number of offspring nodes to which a node transmits influence is a random number according to a fixed probability distribution function that does not vary from node to node. The process is Markovian. The topology along which the influence is transmitted is a chain, a hub-and-spoke, a tree, or any non-circulating shape. In a more complex stochastic process, both the number of offspring nodes and the time when the influence is transmitted to them are random numbers according to probability distribution functions. This is the non-Markovian Bellman-Harris branching process (Vazquez, 2006).

The network topology and the stochastic process that govern influence transmission are described by some probability parameters. First, the probability at which the influence is transmitted from a node n_j to a neighboring node n_k is r_{jk}. The influence is transmitted to multiple neighboring nodes independently in parallel. It is similar to the collaboration probability in trust modeling (Lavrac, 2007). The parameters must satisfy the constraints $0 \le r_{jk} \le 1$. The average number of offspring nodes to which a node n_j transmits the influence is $\sum_{k \neq j} r_{jk}$. Next, the quantity f_j is the probability at which the node n_j becomes a seed node. The parameter must satisfy the constraints $0 \le f_j$ and $\sum_{0 \le j < M} f_j = 1$. These parameters are defined for both the nodes in **O** and **C** in a social network.

It is, however, difficult to derive the mathematical formula that represents the probability for complicated influence transmission over an unlimited number of generations in general stochastic processes. Many theoretical works, therefore, employ an approximation where the number of generations

is limited. The nodes in a given cut-off generation g_C cease to transmit influence to offspring nodes, and influence transmission terminates at that moment. In this study, the statistical inference method is formulated when the cut-off generation is just $g_C = 1$. The collaborative activities among the nodes are then the consequence of a stochastic process which governs the hub-and-spoke influence transmission from an index node to multiple offspring neighboring nodes. The basic procedure is also presented to derive the corresponding formula when the cut-off generation is $g_C \geq 2$.

Node Discovery Solution

Heuristic Method

A conventional heuristic method (data crystallization) was first proposed by (Ohsawa, 2005) to explore a latent structure with dummy variables, and then streamlined by (Maeno, 2009). The method is reviewed briefly.

At first, a set of nodes that appear in the dataset $\{d_i\}$ is grouped into clusters c_l $(0 \leq l < C)$. The number of clusters is C, which may be calculated automatically from the dataset, or given based on the known characteristics of the problem. For the purpose of clustering, closeness between a pair of nodes is calculated by the Jaccard's coefficient (Liben-Nowell, 2004). It is used widely in link discovery, web mining, or text processing. The Jaccard's coefficient between the nodes n and n' is defined by Eq. (3). The function $B(s)$ in Eq.(3) is a Boolean function which returns 1 if the proposition s is true and 0 if s is false. The operators $\wedge$ and $\vee$ represent logical terms AND and OR, respectively.

$$J(n,n') = \frac{\sum_{i=0}^{D-1} B(n \in d_i \wedge n' \in d_i)}{\sum_{i=0}^{D-1} B(n \in d_i \vee n' \in d_i)} \quad (3)$$

The k-medoids clustering algorithm (Hastie, 2001) is employed in this study. It is an EM (expectation-maximization) algorithm similar to the k-means algorithm for numerical data. A medoid node locates most centrally within a cluster. It corresponds to the center of gravity in the k-means algorithm. The clusters and the modoid nodes are re-calculated iteratively until they converge into a stable structure. Other clustering algorithms such as hierarchical clustering or self-organizing mapping may substitute the k-medoids clustering algorithm.

Then, likeliness of every data d_i resulting from the influence of the covert nodes is evaluated with a ranking function $s(d_i)$. The ranking function returns higher value for data having larger likeliness. The strength of the correlation between the data d_i and the cluster c_l is defined by $w(d_i, c_l)$ in Eq. (4).

$$w(d_i, c_l) = \max_{n_j \in c_l} \frac{B(n_j \in d_i)}{\sum_{i=0}^{D-1} B(n_j \in d_i)} \quad (4)$$

The ranking function takes $w(d_i, c_l)$ as an input. Various forms of ranking functions can be constructed. For example, (Maeno, 2009) studied a simple form in Eq. (5) where the function $u(x)$ returns 1 if the real variable x is positive and 0 if x is 0 or negative.

$$s(d_i) \propto \sum_{l=0}^{C-1} u(w(d_i, c_l)) = \sum_{l=0}^{C-1} B(d_i \cap c_l \neq \phi) \quad (5)$$

The data that is assigned the i-th largest value of the ranking function is given by $d_{\sigma(i)}$ where $\sigma(i)$ is calculated by Eq. (6). The value of $s(d_{\sigma(i)})$ is always smaller than that of $s(d_{\sigma(i')})$ for any $i' < i$.

$$\sigma(i) = \arg \max_{i'' \neq \sigma(i') \text{for} \ \forall i' < i} s(d_{i''}) \quad (1 \leq i \leq D) \quad (6)$$

The computational burden of the heuristic method remains light even as the number of nodes and dataset increases. The method is expected to work generally for clustered networks even if the network topology and the stochastic process that govern influence transmission and generate datasets are not understood well. That is, the method works without the knowledge of the influence transmission and its parametric form with r_{jk} and f_j. The result, however, cannot be very accurate because of its heuristic nature. A statistical inference method, which carries a heavy computational burden but outputs more accurate results, is presented next.

Statistical Inference Method

A novel statistical inference method is founded on the following statistical theory. Link discovery (Liben-Nowell, 2004) relies on a dataset within which no data are defective. Data are considered to be an accurate record of collaborative activity patterns. That is, $d_i = \delta_i$ always holds. In the node discovery problem, however, a small portion of data is defective in that $d_i = \delta_i$ does not hold because of the relationship in Eq. (1). The defective data are outliers in that they do not conform to the normal behavior of the rest of the dataset for which $d_i = \delta_i$ always holds. The outliers within the dataset $\{d_i\}$ are, therefore, the targets to retrieve.

In statistical theory, outlier detection is a technique to assess whether data from a given dataset is likely to be spurious or not. For one-dimensional numerical data, Chauvenet's criterion (Taylor, 1996) is used widely. First, the mean and standard deviation of a given dataset are calculated. Then, the probability at which the data is obtained under the calculated mean and standard deviation is evaluated. Finally, the data is considered to be an outlier if the product of this probability and the amount of data in the dataset is less than a given threshold.

The calculation of the mean and standard deviation may be generalized to the calculation of the probability parameters r_{jk} and f_j in the stochastic process that are most likely to generate the dataset. The maximal likelihood estimation is a basic statistical method used for providing estimates for the parameters. The data is likely to be an outlier if the probability at which the data is obtained under the calculated parameters is small. That is, that data should be retrieved.

The statistical inference method employs maximal likelihood estimation to infer the value of r_{jk} and f_j and applies the outlier detection technique to retrieve the defective data for which $d_i = \delta_i$ does not hold. A single symbol $\boldsymbol{\theta}$ represents the parameters r_{jk} and f_j for the nodes in $\mathbf{O}$. $\boldsymbol{\theta}$ is the target variable, the value of which needs to be inferred from the dataset. The logarithmic likelihood function (Hastie, 2001) is defined by Eq. (7). The quantity $p(\{d_i\}|\boldsymbol{\theta})$ denotes the probability at which the dataset $\{d_i\}$ is obtained under a given $\boldsymbol{\theta}$.

$$L(\boldsymbol{\theta}) = \log(p(\{d_i\}|\boldsymbol{\theta})) \quad (7)$$

The data are assumed to be statistically independent on each other. Eq. (7) becomes Eq. (8).

$$L(\boldsymbol{\theta}) = \log(\prod_{i=0}^{D-1} p(d_i|\boldsymbol{\theta})) = \sum_{i=0}^{D-1} \log(p(d_i|\boldsymbol{\theta})) \quad (8)$$

Let the quantity $q_{i|jk}$ represent the probability at which the data d_i is obtained in case that the index node n_j transmits influence to the offspring neighboring node n_k. When the cut-off generation is just $g_c = 1$, the quantity $q^{[1]}{}_{i|jk}$ in Eq. (9) gives the probability $q_{i|jk} = q^{[1]}{}_{i|jk}$.

$$q^{[1]}{}_{i|jk} = \begin{cases} r_{jk} & (\mathbf{d}_{ik} = 1) \\ 1 - r_{jk} & (\mathbf{d}_{ik} = 0) \end{cases} \quad (9)$$

Eq. (9) is equivalent to Eq. (10) since the value of $\mathbf{d}_{ik}$ is either 0 or 1.

$$q^{[1]}{}_{i|jk} = \mathbf{d}_{ik} r_{jk} + (1-\mathbf{d}_{ik})(1-r_{jk}) \quad (10)$$

When the cut-off generation is $g_c = 2$, $q^{[2]}{}_{i|jk}$ in Eq.(11) is added to $q_{i|jk}$. The result is $q_{i|jk} = q^{[1]}{}_{i|jk} + q^{[2]}{}_{i|jk}$. The influence is transmitted across the intermediate node n_l.

$$q^{[2]}{}_{i|jk} = \begin{cases} 1 - \prod_{j \neq l \neq k} (1 - r_{jl} r_{lk}) & (\mathbf{d}_{ik} = 1) \\ \prod_{j \neq l \neq k} (1 - r_{jl} r_{lk}) & (\mathbf{d}_{ik} = 0) \end{cases} \quad (11)$$

When the cut-off generation is $g_c > 2$, the probability is given by Eq. (12).

$$q_{i|jk} = \sum_{g=1}^{g_c} q^{[g]}{}_{i|jk} \quad (12)$$

For $q^{[g]}{}_{i|jk}$, the multiplication over the intermediate nodes in Eq. (11) is replaced by Eq. (13). The influence is transmitted across the multiple generations of the intermediate nodes $n_{11}, n_{12}, \ldots, n_{1g-1}$.

$$\prod_{j \neq l_1 \neq l_2 \neq \cdots \neq l_{g-1} \neq k} (1 - r_{jl_1} r_{l_1 l_2} \cdots r_{l_{g-1} k}) \quad (13)$$

Again, when the cut-off generation is just $g_c = 1$, the probability $p(\{d_i\} | \boldsymbol{\theta})$ in Eq. (8) is expressed by Eq. (14).

$$p(d_i | \boldsymbol{\theta}) = \sum_{j=0}^{N-1} \mathbf{d}_{ij} f_j \prod_{0 \leq k < N \wedge k \neq j} q^{[1]}{}_{i|jk} \quad (14)$$

Eq. (14) takes an explicit formula in Eq. (15). The case $k = j$ in multiplication ($\prod_k q^{[1]}{}_{i|jk}$) is included since $\mathbf{d}_{ik}{}^2 = \mathbf{d}_{ik}$ always holds.

$$L(\boldsymbol{\theta}) = \sum_{j=0}^{N-1} \mathbf{d}_{ij} f_j \prod_{k=0}^{N-1} \{1 - \mathbf{d}_{ik} + (2\mathbf{d}_{ik} - 1) r_{jk}\} \quad (15)$$

The maximal likelihood estimator $\hat{\boldsymbol{\theta}}$ is obtained by solving Eq. (16). It gives the values of the parameters $\hat{r}_{jk}$ and $\hat{f}_j$. A pair of nodes n_j and n_k for which $r_{jk} > 0$ possesses a link between them.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) \quad (16)$$

A simple incremental optimization technique, the hill climbing method or the method of steepest descent, is employed to solve Eq. (16). Non-deterministic methods such as simulated annealing (Hastie, 2001) can be employed to strengthen the search ability and to avoid sub-optimal solutions. These methods search more optimal parameter values around the present values and update them as in Eq. (17) until the values converge.

$$\begin{cases} r_{jk} \rightarrow r_{jk} + \Delta r_{jk} \\ f_j \rightarrow f_j + \Delta f_j \end{cases} \quad (0 \leq j,k \leq N) \quad (17)$$

The change in the logarithmic likelihood function can be calculated as a product of the derivatives (differential coefficients with regard to r_{jk} and f_j) and the amount of the updates in Eq. (18). The update Δr_{jk} and Δf_j should be in the direction of the steepest ascent in the landscape of the logarithmic likelihood function.

$$\Delta L(\boldsymbol{\theta}) = \sum_{n,m=0}^{N-1} \frac{\partial L(\boldsymbol{\theta})}{\partial r_{nm}} \Delta r_{nm} + \sum_{n=0}^{N-1} \frac{\partial L(\boldsymbol{\theta})}{\partial f_n} \Delta f_n \quad (18)$$

The derivatives with regard to r are given by Eq. (19).

$$\frac{\partial L(\boldsymbol{\theta})}{\partial r_{nm}} = \sum_{i=0}^{D-1}[\, f_n \mathbf{d}_{in}(2\mathbf{d}_{im}-1) \times \frac{\prod_{k\neq m}\{1-\mathbf{d}_{ik}+(2\mathbf{d}_{ik}-1)r_{nk}\}}{\sum_{j=0}^{N-1}\mathbf{d}_{ij} f_j \prod_{k=0}^{N-1}\{1-\mathbf{d}_{ik}+(2\mathbf{d}_{ik}-1)r_{jk}\}}] \quad (19)$$

The derivatives with regard to f are given by Eq. (20).

$$\frac{\partial L(\boldsymbol{\theta})}{\partial f_n} = \sum_{i=0}^{D-1} \frac{\mathbf{d}_{in}\prod_{k=0}^{N-1}\{1-\mathbf{d}_{ik}+(2\mathbf{d}_{ik}-1)r_{nk}\}}{\sum_{j=0}^{N-1}\mathbf{d}_{ij} f_j \prod_{k=0}^{N-1}\{1-\mathbf{d}_{ik}+(2\mathbf{d}_{ik}-1)r_{jk}\}} \quad (20)$$

The ranking function $s(d_i)$ is the reciprocal number of the probability at which d_i is obtained under the maximal likelihood estimator $\hat{\boldsymbol{\theta}}$. According to the outlier detection technique, a higher return value is given to the data which are less likely to be obtained. The ranking function is given by Eq. (21).

$$s(d_i) = \frac{1}{p(d_i \mid \hat{\boldsymbol{\theta}})} \quad (21)$$

The data that is assigned the i-th largest value of the ranking function is given by $d_{\sigma(i)}$ where $\sigma(i)$ is calculated by Eq. (6).

Node Discovery Solution Test

Network model

Computationally synthesized networks and a real clandestine organization are employed to generate test datasets for a performance evaluation of the methods discussed.

The computationally synthesized networks include three network models, a Barabasi-Albert model (Barabasi, 1999), a random Erdos-Renyi network model (Erdos, 1959), and a clustered network model. In the Barabasi-Albert model, the probability at which a node n_k connects a link to another node n_j is proportional to the nodal degree of n_j ($p(k \to j) \propto K(n_j)$). The occurrence frequency of the nodal degree tends to be scale-free ($F(K) \propto K^a$). The model tends to be inhomogeneous. The clustering coefficient (Watts, 1998) averaged over all the nodes in the network $\langle W(n_j)\rangle$ is moderately large. The Gini coefficient of a nodal degree G is large. In economics, the Gini coefficient is a measure of inequality in income or wealth distribution. A larger Gini coefficient indicates lower equality.

In the random network model, where links are placed between randomly chosen pairs of nodes, $p(k \to j)$ does not depend on n_j. The nodal degree does not differ largely among nodes. The model tends to be homogeneous. Both the clustering coefficient and the Gini coefficient are small.

The clustered network in this study is an extension of the Barabasi-Albert model. Here, every node n_j is assigned a pre-determined cluster attribute $c(n_j)$ to which it belongs. The number of clusters is C. The probability $p(k \to j)$ is replaced by Eq. (22).

$$p(k \to j) = \begin{cases} \eta C K(n_j) & (c(n_j) = c(n_k)) \\ K(n_j) & (c(n_j) \neq c(n_k)) \end{cases} \quad (22)$$

The cluster contrast parameter η is introduced. Links between the clusters appear less frequently as η increases. The model returns to the original Barabasi-Albert model when $C = 1$ because $c(n_j) = c(n_k)$ always holds. The clustered network model in Figure 1 is an example of a more clustered network model. Its characteristics are C =5, η =50, M =101, $\langle K(n_j)\rangle$ =4, $\langle W(n_j)\rangle$ =0.42, and G =0.36. The clustered network model in Figure 2 is an example of a less clustered network model. Its characteristics are C =5, η =2.5, M =101, $\langle K(n_j)\rangle$ =4, $\langle W(n_j)\rangle$ =0.22, and G =0.37. Both

the clustering coefficient and the Gini coefficient are large when the cluster contrast parameter η is large. The node n_{12} in Figure 1 is a typical hub node. Hub nodes are those that have a nodal degree larger than the average. The node n_{75} in Figure 1 is a typical peripheral node. Peripheral nodes are those having a nodal degree smaller than the average. The model tends to be inhomogeneous.

The network in Figure 3 represents a real clandestine organization; a global mujahedin organization that was analyzed in (Sageman 2004). The mujahedin in the global Salafi jihad means Muslim fighters in Salafism (Sunni Islamic school of thought) who struggle to establish justice on earth. The organization consists of M =107 persons and 4 regional sub-networks. The sub-networks represent Central Staffs (n_{CSj}) including the node n_{ObL}, Core Arabs (n_{CAj}) from the Arabian Peninsula countries and Egypt, Maghreb Arabs (n_{MAj}) from the North African countries, and Southeast Asians (n_{SAj}). The network topology is not simply hierarchical. The 4 regional sub-networks are connected mutually in a complex manner.

The average nodal degree is $\langle K(n_j)\rangle$=5.1. The global mujahedin organization has a large Gini coefficient of the nodal degree (G=0.35) and a large average clustering coefficient ($\langle W(n_j)\rangle$=0.54). The values mean that the organization possesses hubs and a cluster structure. The node representing Osama bin Laden n_{ObL} is a hub ($K(n_{ObL})=8$). He is believed to be the founder of the organization, and said to be the covert wire-puller who provided operational commanders in regional sub-networks with financial support in many terrorism attacks including 9/11 in 2001. His whereabouts are unknown despite many efforts to locate and capture him.

Figure 1. Clustered Network Model for η=50

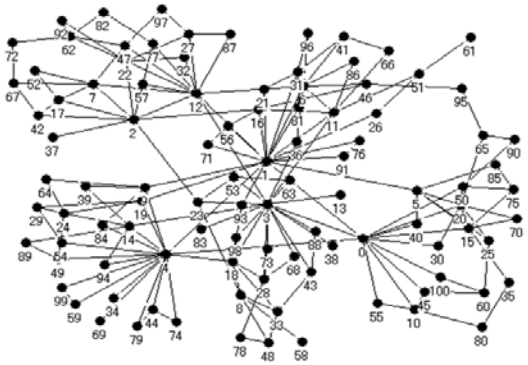


Figure 2. Clustered Network Model for η=2.5

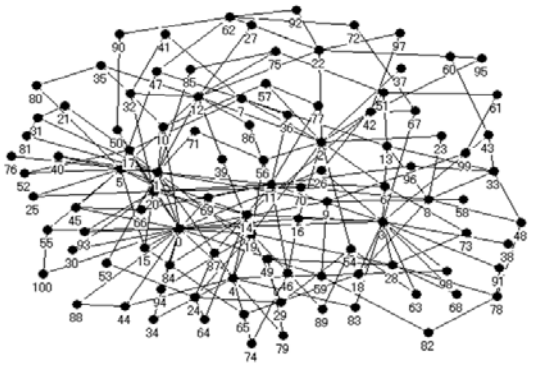
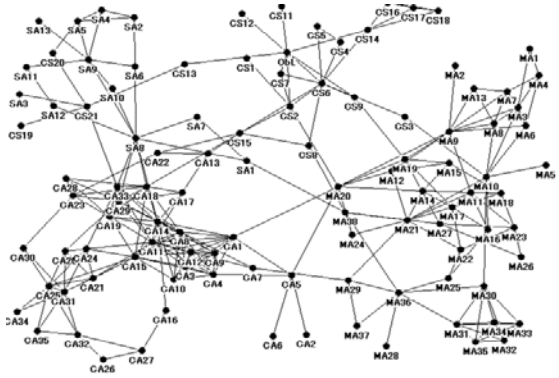


Figure 3. Network for a Global Mujahedin Organization



Test Dataset

The test dataset $\{d_i\}$ is generated from the above mentioned networks in the 2 steps below.

In the first step, the collaborative activity patterns δ_i are generated D times according to the influence transmission under the true value of $\boldsymbol{\theta}$. A pattern includes both a seed node n_j and multiple offspring neighboring nodes n_k. An example is $\delta_{\mathrm{EXl}} = \{ n_{\mathrm{CS1}}, n_{\mathrm{CS2}}, n_{\mathrm{CS6}}, n_{\mathrm{CS7}}, n_{\mathrm{CS9}}, n_{\mathrm{ObL}}, n_{\mathrm{CS11}}, n_{\mathrm{CS12}}, n_{\mathrm{CS14}}\}$ for the global mujahedin organization in Figure 3.

In the second step, deleting the covert nodes belonging to $\mathbf{C}$ from the collaborative activity patterns $\{\delta_i\}$ generates the dataset $\{d_i\}$. The example δ_{EXl} results in the data $d_{\mathrm{EXl}} = \{ n_{\mathrm{CS1}}, n_{\mathrm{CS2}}, n_{\mathrm{CS6}}, n_{\mathrm{CS7}}, n_{\mathrm{CS9}}, n_{\mathrm{CS11}}, n_{\mathrm{CS12}}, n_{\mathrm{CS14}}\}$ if the experimental condition is that Osama bin Laden is the target covert node to discover. That is, $\mathbf{C} = \{n_{\mathrm{ObL}}\}$. The covert node in $\mathbf{C}$ may appear multiple times in $\{\delta_i\}$. The number of the data to retrieve (D_{t}) is given by Eq. (23).

$$D_{\mathrm{t}} = \sum_{i=0}^{D-1} B(d_i \neq \delta_i) \quad (23)$$

In the performance evaluation, a few assumptions are made for simplicity. The probability f_j does not depend on the nodes ($f_j = 1/M$). The value of the probability r_{jk} is 1 if a link is present between nodes and 0 if a link is absent. It means that the number of possible collaborative activity patterns is bounded. The influence transmission is symmetrically bi-directional. That is, $r_{jk} = r_{kj}$.

Node Discovery Solution Performance

Performance Measure

Three measures, precision, recall, and Van Rijsbergen's F measure (Korfhuge, 1997), are used to evaluate the performance of the methods. They are commonly used in information retrieval such as search, document classification, and query classification. The precision p is used as evaluation criteria, which is the fraction of the number of relevant data to the number of all data retrieved by search. The recall r is the fraction of the number of the data retrieved by search to the number of all the relevant data. The relevant data refers to the data where $d_i \neq \delta_i$. They are given by Eq. (24) and Eq. (25). They are functions of the number of the retrieved data D_{r}. It can take the value from 1 to D. The data is retrieved in the order of $d_{\sigma(1)}$, $d_{\sigma(2)}$, to $d_{\sigma(D_{\mathrm{r}})}$.

$$p(D_{\mathrm{r}}) = \frac{\sum_{i=1}^{D_{\mathrm{r}}} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_{\mathrm{r}}} \quad (24)$$

$$r(D_{\mathrm{r}}) = \frac{\sum_{i=1}^{D_{\mathrm{r}}} B(d_{\sigma(i)} \neq \delta_{\sigma(i)})}{D_{\mathrm{t}}} \quad (25)$$

The F measure F is the harmonic mean of the precision and recall. It is given by Eq. (26).

$$F(D_{\mathrm{r}}) = \frac{2\,p(D_{\mathrm{r}})r(D_{\mathrm{r}})}{p(D_{\mathrm{r}}) + r(D_{\mathrm{r}})} \quad (26)$$

The precision, recall, and F measure range from 0 to 1. All the measures take larger values as the retrieval becomes more accurate.

Performance Comparison

The performance of the heuristic method and statistical inference method is compared with the test dataset generated from the computationally synthesized clustered network models.

First, Figure 4 shows the precision $p(D_{\mathrm{r}})$ as a function of the rate of the number of the retrieved data to the total number of the data (D_{r}/D). The experimental conditions are that the hub node n_{12} in the more clustered network

model in Figure 1 is the covert node ($\mathbf{C}=\{n_{12}\}$, $|\mathbf{C}|=1$, $|\mathbf{O}|=100$) and $D=100$. The three graphs show the results for [a] the statistical inference method, [b] the heuristic method with the prior knowledge of C =5, and [c] the heuristic method with the prior knowledge of C =10. The broken lines indicate the theoretical limit that is the upper boundary of the performance, and the random retrieval that indicates the lower boundary. The vertical solid line indicates the position where $D_r = D_t$. The values of precision at $D_r = D_t$ measure the typical ability of the methods. Figure 5 shows the recall $r(D_r)$ as a function of D_r / D. Figure 6 shows the F measure $F(D_r)$ as a function of D_r / D. The experimental conditions are the same as those for Figure 4. The accuracy of retrieval by the heuristic method is moderately good if the number of clusters (C =5) is previously known. If a wrong number of clusters such as C =10 is given, the accuracy of retrieval becomes worse. The statistical inference method always surpasses the heuristic method. The accuracy of the retrieval is close to the theoretical limit.

Next, Figure 7 shows the F measure $F(D_r)$ as a function of D_r / D. The experimental condition is that the hub node n_{12} in the less clustered network model in Figure 2 is the covert node. The two graphs show the results for [a] the statistical inference method and [b] the heuristic method with the prior knowledge of C =5. The performance of the statistical inference method is still good while that of the heuristic method worsens in a less clustered network.

Finally, Figure 8 shows the F measure $F(D_r)$ as a function of D_r / D. The experimental condition is that the peripheral node n_{75} in the more clustered network model in Figure 1 is the covert node. Figure 9 shows the F measure $F(D_r)$ as a function of D_r / D. The experimental condition is that the peripheral node n_{48} in the less clustered network model in Figure 2 is the covert node. The statistical inference method works fine in both experiments while the heuristic method fails.

Figure 4. Precision in Discovering a Hub Node in Figure 1

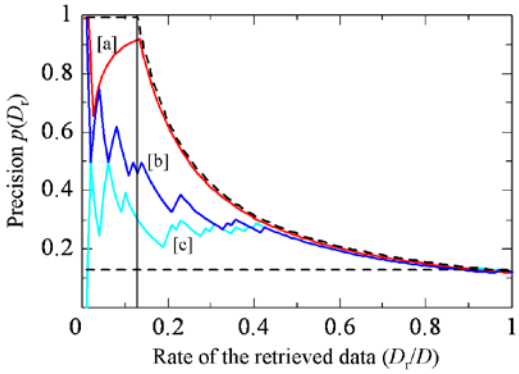


Figure 5. Recall in Discovering a Hub Node in Figure 1

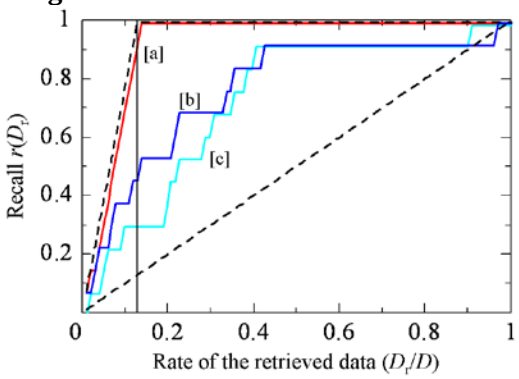


Figure 6. F Measure in Discovering a Hub Node in Figure 1

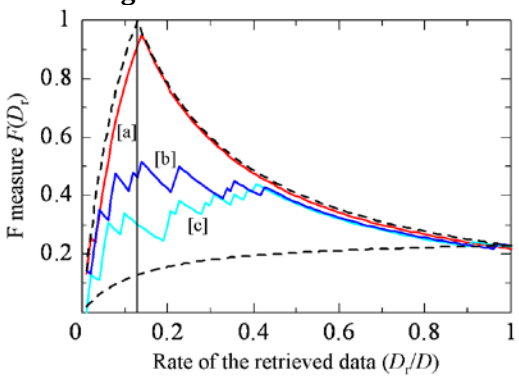


Figure 7. F Measure in Discovering a Hub Node in Figure 2

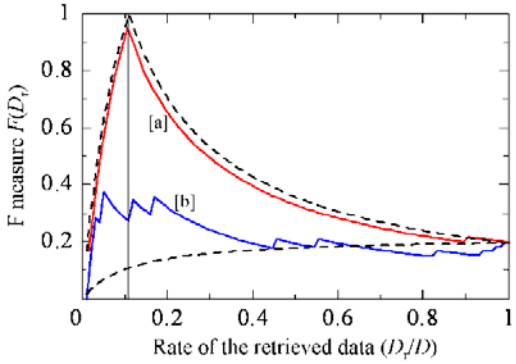


Figure 8. F Measure in Discovering a Peripheral Node in Figure 1

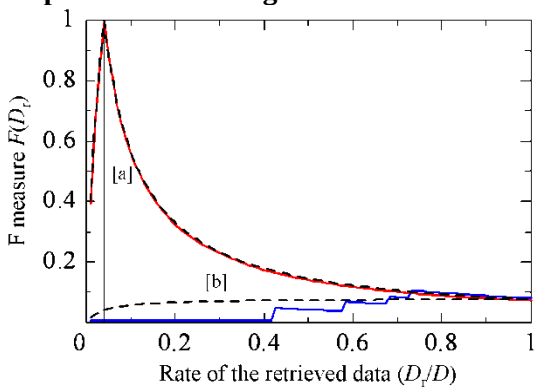
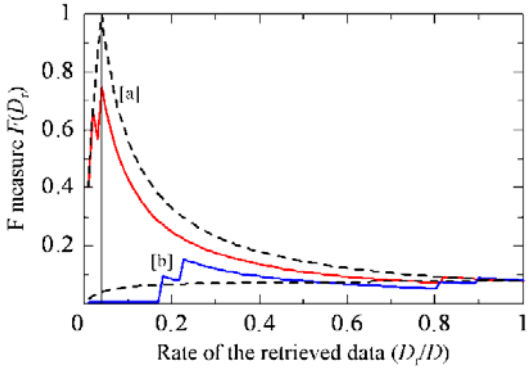


Figure 9. F Measure in Discovering a Peripheral Node in Figure 2



Network Comparison

The applicability of the statistical inference method is tested with the test datasets generated from many computationally synthesized network models.

At first, the average performance in discovering various covert nodes is measured. Figure 10 shows the F measure $F(D_t)$ (at $D_r = D_t$) as a function of the nodal degree of the covert nodes (K). The theoretical limit of the performance is $F(D_t) = 1$. Individual plots show the values averaged over all the possible experiments where a single node n_j having a given nodal degree ($K(n_j) = K$) is the covert node ($\mathbf{C} = \{n_j\}$). The two graphs show the results for [a] the more clustered network model in Figure 1 and [b] the less clustered network model in Figure 2. The overall performance is good. The F measure ranges from 0.6 to 0.8 when the nodal degree of the covert nodes is less than about 5. On the other hand, the F measure is more than 0.8 when the nodal degree is more than about 5. The F measure approaches the theoretical limit when the nodal degree is more than about 10. The results imply that the method is not subject to aberrations with respect to the types of covert nodes.

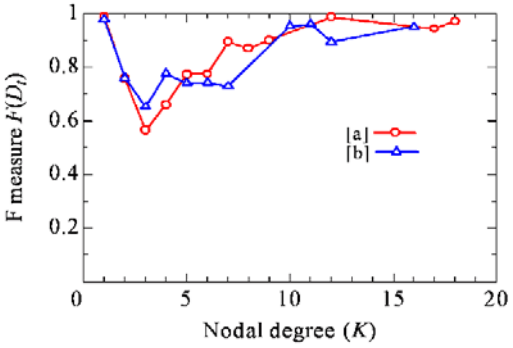
Second, the statistical inference method is applied to various network models of larger sizes. Figure 11 shows the F measure $F(D_t)$ as a function of the nodal degree of the covert nodes (K). The experiments are similar to those in Figure 10. Individual plots show the values averaged over all the possible experiments where a single node n_j having a given nodal degree is the covert node. The three graphs show the results for [a] the Barabasi-Albert model where M =201, $\langle K(n_j)\rangle$ =4, $\langle W(n_j)\rangle$ =0.14, and G =0.38, [b] the more clustered network model where M =201, C =8, η =50, $\langle K(n_j)\rangle$ =4, $\langle W(n_j)\rangle$ =0.43, and G =0.39, and [c] the random network model where M =201, $\langle K(n_j)\rangle$ =4, $\langle W(n_j)\rangle$ =0.095, and G =0.26. No nodes have a nodal degree larger than 10 in the random network model. Again, the overall performance is good. The three graphs nearly overlap. The method works for the

Barabasi-Albert model and the random network model as successfully as for the clustered network model. Note that the performance for the covert nodes having small nodal degree (less than about 5) is relatively bad. This tendency is similar to the graphs in Figure 10. The results demonstrate that the method is not subject to aberrations with respect to network topologies.

Finally, the average performance for larger and smaller networks is measured. Figure 12 shows the F measure $F(D_t)$ as a function of the total number of nodes M. The graph shows the values averaged over all the possible experiments where a single node is the covert node. The network is the Barabasi-Albert model where $\langle K(n_j)\rangle$=4. The method still works even as the number of nodes approaches M=1000. The results prove that the method is not subject to aberrations with respect to the network sizes.

These are the basic performance characteristics of the statistical inference method. The method can solve the node discovery problems for various types of covert nodes, network topologies, and network sizes.

Figure 10. F Measure for the Clustered Network Model



Application

How are investigators aided by the accurate retrieval of the statistical inference method? Let's assume that investigators have a dataset of the members of the global mujahedin organization except Osama bin Laden by the time of the attack. The situation is simulated

Figure 11. F Measure for Various Network Models

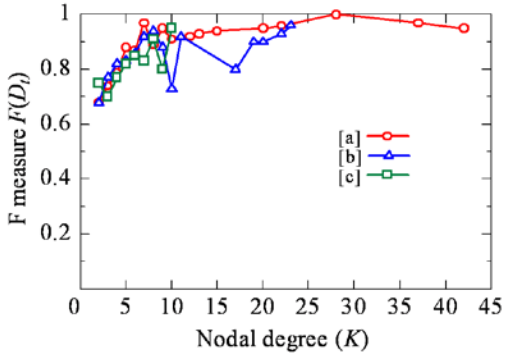
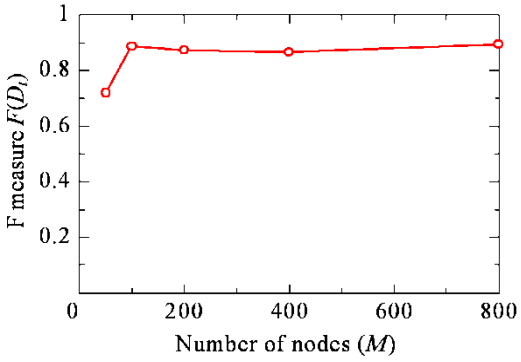


Figure 12. F Measure for Various Network Sizes



computationally similarly to the problems in Figures 4 through 12. In this case, the experimental condition is that n_{ObL} in Figure 3 is the covert node, $\mathbf{C} = \{n_{\mathrm{ObL}}\}$.

Figure 13. F Measure in Discovering Osama bin Laden in Figure 3

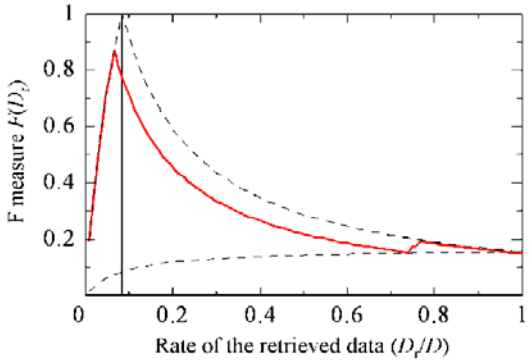


Figure 13 shows $F(D_r)$ as a function of D_r / D. The accuracy of retrieval is close to the theoretical limit. The data that is assigned the largest value of the ranking function ($d_{\sigma(1)}$) includes all and only the neighboring overt nodes n_{CS1}, n_{CS2}, n_{CS6}, n_{CS7}, n_{CS9}, n_{CS11}, n_{CS12}, and n_{CS14}. This encourages the investigators to search for an unknown wire-puller or a key conspirator near these 8 known neighbors who are likely to be the close associates.

The method, however, fails to retrieve two data $\delta_{FL1} = \{ n_{ObL}, n_{CS11} \}$ and $\delta_{FL2} = \{ n_{ObL}, n_{CS12} \}$. These nodes have a small nodal degree, $K(n_{CS11}) = 1$ and $K(n_{CS12}) = 1$. This shows that the data on the nodes having small nodal degree do not always provide investigators with many clues on the covert nodes. This is consistent with the results in Figures 10 and 11 that the F measure becomes slightly worse as the nodal degree of the target covert nodes gets smaller.

CONCLUSION

In this study, the node discovery problem for a social network is defined mathematically, and a novel statistical inference method is discussed and compared with a conventional heuristic method (data crystallization). The accuracy of retrieval (precision, recall, and F measure) is proven close to the theoretical limit in various problems. In the investigation of a clandestine organization, the method aids the investigators in identifying the close associates near whom an unknown wire-puller or a key conspirator should be searched for.

Three issues will be addressed as future works. The first issue is to derive the mathematical formula for the statistical inference method that is applicable to the influence transmission described by general non-Markovian stochastic processes.

The second issue is to develop a method to solve variants in the node discovery problem. Discovering fake nodes or spoofing nodes is also an interesting problem in uncovering the malicious intentions of a clandestine organization. A fake node is a person who does not exist in the organization, but appears in the dataset. A spoofing node is a person who belongs to an organization, but appears as a different node in the dataset.

The third issue is to exploit other application domains in social and business sciences. If a similar problem were encountered in studying friendship relationships, web communities, business organizations, or economic systems, the methods presented in this study would also contribute to such studies.

REFERENCES

Anderson, C., Wasserman, S., and Crouch, B., 1999. A p* primer, Logit models for social networks. Social Networks 21, 37-66.

Adamic, L. A., Adar, E,. 2003. Friends and neighbors on the web. Social Networks 25, 211-228.

Barabasi, A. L., Albert, R., and Jeong, H., 1999. Mean-field theory for scale-free random networks. Physica A 272, 173-187.

Clauset, A., Moore, C., and Newman, M. E. J., 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98-100.

Erdos, P. and Renyi, A., 1959. On random graphs I. Publicationes Mathematicae 6, 290-297.

Frank, O., and Strauss, D., 1986. Markov graphs. Journal of the American Statistical Association 81, 832-842.

Getoor, L., and Diehl, C. P., 2005. Link mining: a survey. ACM SIGKDD Explorations 7, 3-12.

Hastie, T., Tibshirani, R., Friedman, J,. 2001. The elements of statistical learning: Data mining, inference, and prediction (Springer series in statistics). Springer-Verlag, Berlin.

Iribarren, J. L. and Moro, E., 2009. Impact of human activity patterns on the dynamics of information diffusion. Physical Review Letters 103, 038702.

Klerks, P,. 2002. The network paradigm applied to criminal organizations. Connections 24, 53-65.

Krebs, V. E,. 2002. Mapping networks of terrorist cells. Connections 24, 43-52.

Korfhuge, R. R., 1997. Information storage and retrieval. Wiley.

Lavrac, N., Ljubic, P., Urbancic, T., Papa, G., Jermol, M., and Bollhalter, S., 2007. Trust modeling for social networks using reputation and collaboration estimates. IEEE Transactions on Systems, Man, & Cybernetics Part C 37, 429-439.

Lazega, E. and Pattison, P., 1999. Multiplexity, generalized exchange and cooperation in organizations. Social Networks 21, 67-90.

Liben-Nowell, D., and Kleinberg, J., 2004. The link prediction problem for social networks. Journal of American Society of Information Science and Technology 58, 1019-1031.

Lindelauf, R., Borm, P., and Hamers, H., 2009. The influence of secrecy on the communication structure of covert networks. Social Networks 31, 126-137.

Maeno, Y., and Ohsawa, Y., 2009. Analyzing covert social network foundation behind terrorism disaster. International Journal of Services Sciences 2, 125-141.

Morselli, C., Giguere, C., Petit, K,. 2007. The efficiency/security trade-off in criminal networks. Social Networks 29, 143-153.

Newman, M. E. J., and Leicht, E. A., 2007. Mixture models and exploratory analysis in networks. Proceedings of the National Academy of Sciences USA 104, 9564-9569.

Ohsawa, Y., 2005. Data crystallization: chance discovery extended for dealing with unobservable events. New Mathematics and Natural Computation 1, 373-392.

Palla, G., Derenyi, I., Farkas, I., and Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814-818.

Pattison, P., and Robins, G. L., 2005. Neighborhood-based models for social networks. Sociological Methodology 32, 301-337.

Pattison, P., and Wasserman, S., 1999. Logit models and logistic regressions for social networks. Britsh Journal of Mathematical and Statistical Psychology 52, 169-194.

Rabbat, M. G., Figueiredo, M. A. T., and Nowak, R. D., 2008. Network Inference from co-occurrences. IEEE Transactions on Information Theory 54, 4053-4068.

Robins, G. J., Pattison, P. E., and Elliott, P., 2001. Network models for social influence processes. Psychometrika 66, 161-190.

Robins, G. L., Pattison, P. E., and Wasserman, S., 1999. Logit models and logistic regressions for social networks. Psychometrika 64, 371-394.

Sageman, M., 2004. Understanding terror networks. University of Pennsylvania Press.

Silva, J., and Willett, R., 2009. Hypergraph-based outlier detection of high-dimensional co-occurrences. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 563-569.

Silva, R., Scheines, R., Glymour, C., Spirtes, P,. 2006. Learning the structure of linear latent variable models. Journal of Machine Learning Research 7, 191-246.

Singh, S., Allanach, J., Haiying, T., Pattipati, K., Willett, P,. 2004. Stochastic modeling of a terrorist event via the ASAM system. Proceedings of the IEEE International Conference on Systems, Man & Cybernetics, Hague, 5673-5678.

Taylor, J. R., 1996. An introduction to error analysis - The study of uncertainties in physical measurements. University Science Books.

Vazquez, A., 2006. Spreading dynamics on heterogeneous populations: Multitype network approach. Physical Review E 74, 066114.

Watts, D. J., Strogatz, S. H,. 1998. Collective dynamics of small-world networks. Nature 398, 440-442.

A Note on Creating Networks from Social Network Data

Steven Gustafson
GE Global Research, One Research Circle, Niskayuna, NY

Huaiyu (Harry) Ma
GE Global Research, One Research Circle, Niskayuna, NY

Abha Moitra
GE Global Research, One Research Circle, Niskayuna, NY

We are interested in the variance of social networks when using supporting evidence to define "friend" relationships, particularly in online social networks. Of related interest, relevant to this study, is the impact of various network sampling methods as well as the ability to capture the true structure of a network given incomplete or inaccurate data. Using empirical social network data, we explore the effect of requiring friendship relationships to be supported with communications between members, where friendship between members must also be corroborated with friendships to a third member. Ultimately, we hope to identify substantial relationships between members that may be capable of influencing behavior. A hypothetical scenario of measuring the vitality of an online community allows us to assess the effect on pertinent network metrics. Results indicate some amount of stability in certain measurements, but enough variance is present to suggest that in empirical networks, the presence of key nodes and edges reduces the robustness previously measured in random networks. Most significantly, the study demonstrates a possible way to identify the robustness of relationships within networks, as well as identify high-level groupings of communities as stable under different relationships, by increasing the amount of 'evidence' required to create ties between members, irrespective of the strength of the ties.

Authors: *Steven Gustafson, Ph.D. (University of Nottingham) is a computer scientist in the Computational Intelligence Lab at GE Global Research, focused on social networks, machine learning and data mining. He develops and applies research for understanding the behavior of people: from customer relational databases to Web data. He also serves as a Technical Editor-in-Chief of Memetic Computing, a journal dedicated to heuristic search and optimization.*

Huaiyu (Harry) Ma, Ph.D. (Rensselaer Polytechnic Institute) is a statistician in the Applied Statistics Lab at GE Global Research Center, performing data analysis, statistical computing, and time series analysis for risk management and customer behavior modeling in both online social networks and in the financial sector.

Abha Moitra, Ph.D. (Tata Institute of Fundamental Research) is a computer scientist in the Computational Intelligence Lab at GE Global Research. She has developed tools for automated generation of ontologies from technical documents, including clinical medical guidelines. She has also investigated semantic languages for modeling data provenance and information assurance.

Correspondence*: Contact Steven Gustafson at steven.gustafson@research.ge.com*

INTRODUCTION

Studies on real-world complex networks do not typically consider the existence of erroneous links in the observed social network. Prior work has assumed that large sample sizes are representative: that the impact of erroneous links will be washed out by the sheer amount of valid data. However, this is not always true, particularly in real-world network data that tend to exhibit power-law degree distributions. It is important to study the impact of these issues on the topology of the observed networks and any inferences drawn from the networks. Also, as the availability of data increases, and more beneficial applications of social network analysis are identified (Lazer et al., 2009), understanding data issues inherent in social networks will be critical.

Latapy and Magnien (2008) are among recent studies that validate the sample size assumption, showing that it is possible to distinguish between cases where this assumption is reasonable, and cases where it must be discarded. Prior work has examined this topic within simulated data, which we do not discuss here as our interest is primarily in empirical data. Latapy and Magnien (2008) conclude that the qualitative properties of some statistics do not depend on the sample size, as long as it is not trivially small. They find that some statistics, like average degree, can be used to infer other statistics. Other statistics like transitivity are generally unstable as sample sizes grow. Measures like transitivity appear to be more related to the "structural" aspects of the network and are somehow more related to other measures like maximal degree. While qualitative estimations of the more stable statistics, for example average degree, are possible, obtaining accurate estimations of statistics like transitivity remains difficult. Lin and Zhao (2005) present a study on the impact of erroneous links on degree distribution estimation and show that the degree distributions of power-law networks are still power-law for the middle range degrees, but can be greatly distorted for low and high degrees. Borgatti et al. (2006) show that centrality measures are surprisingly similar with respect to pattern and level of robustness to data errors and different types of errors have relatively similar effects on centrality robustness. The limitation of this last study is that it considers only random errors on random networks. Other related work can be found in Marsden (1990); Costenbader and Valente (2003); Rothenberg (1995); Bernard et al. (1984); Adar and Re (2007).

In this paper, we argue that the imprecision in the inferences drawn from the collected social networks requires rethinking the use of social network analysis techniques, and demands new statistical methods for analyzing data and making inferences. In the following section, we illustrate our point by showing a motivating example. We hypothesize that in most real-world networks, working at a scale that is realistic in collection and analysis size, the variations and instability of network metrics are greater than theoretical models would predict. To provide initial evidence for this hypothesis, we show how network measures vary as our sampling strategy is made more strict to include only ties that we believe are significant. Our results are limited in scale and scope, but provide a clear empirical demonstration of the variance in social network metrics when applied to real-world data.

METHODS

To demonstrate the impact of imperfect data, we study the friend networks of an online social site and examine the robustness and relevance of certain network measurements. For privacy reasons, we do not give the name of the site, but it is typical of what is understood to be an online social networking site, allowing people to connect with their friends and family, and exchange information and content. The social network site is organized into communities. Each community has a leader who is part of the network like any other member, but is responsible for maintaining the vitality of the community. Members can create "friend" relationships with other members by directing a

friend request to another member (but once accepted, the tie is non-directed), send private messages to other members, or comment on other members' profile pages. Figure 1 below shows an example of a friend network, where the large node represents the leader.

We are interested in understanding the friend networks of the communities and evaluating the leader's capability to be a central figure in the community. To maintain the global social network, the owners of the site are constantly looking to replace or assist leaders who are not central in their community. The lack of a strong friend network is one way to identify underperforming leaders. As in most online social networks with friend relationships, we expect a significant amount of friending is spurious and does not represent significant relationships that might influence positive behavior on the site. To filter out such spurious links, we add a constraint that a friend relationship between two members is valid only if there exists a third member who is a common friend of both.

To further validate a friend relationship we will require that some amount of communication take place between the friends. We have access to the online communication data of the community members, including messages sent between members and comments on member profile-like pages. We start out with the total friend network for a community, where only triads are allowed. We then remove ties that lack a certain amount of communication, causing additional edges and nodes to be removed when some triads are no longer

Figure 1. Friending Network

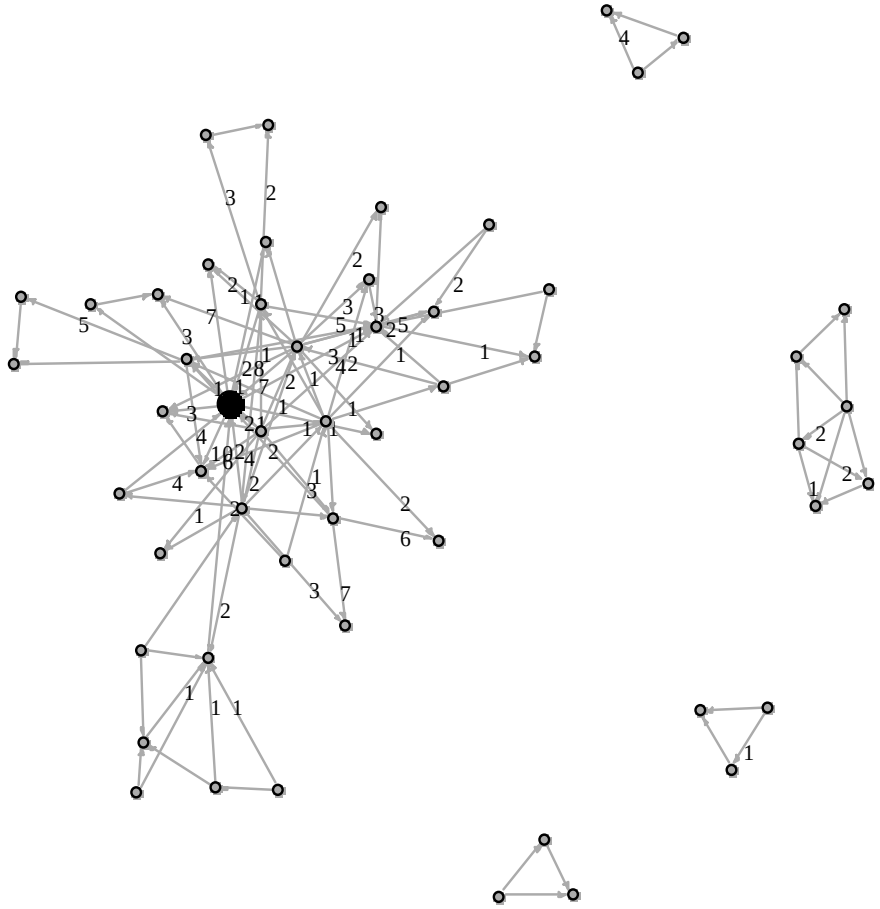


Figure 1. The friending network of an online social networking site. The large black node indicates the leader. Arrow indicates the initial friending request direction. Edge weights are the number of communications between nodes.

supported. In summary, to construct a network for analysis, the following process is carried out once: first we enforce triads on the original network, then we remove ties that lack a level of communication, and lastly, we again enforce the triad constraint. This network effect, the removal of a few edges percolating through the network, will potentially cause network metrics to vary in different communities. This removal of edges due to the lack of communication can be seen as a kind of reverse sampling. In the strictest case, where we require 3 or more communications between members to validate a friend tie, we are sampling fewer nodes and edges. As we relax this criterion, we are adding additional nodes and edges, sampling or uncovering more of the network, but possibly introducing error by including inaccurate ties that are not significant relationships.

Figure 1 illustrates the common structure we observed in these online communities. There is typically a large connected component that consists of most of the members, while the other members are grouped in a few islands. In Figure 1, the numbers on the edges denote the total number of communications between the members represented by the two end nodes. Our final friend definition is as follows: Member A and Member B are friends only if they communicate at least ***K*** times and share a common friend. The cutoff value ***K*** also has significant impact on the structure of the resultant networks, as one would expect. In Figure 1, the typical structure of our communities is apparent: the community leader is typically central in the big connected component. However, one can also see a few other nodes with very high degrees. The leader is by no means the only high degree node, or even the most important. Their role is to make sure the community is active and healthy.

Threshold

Figure 2 shows the impact of increasing the value of ***K*** on the number of nodes across all the communities in the social network. In general, increasing the value of ***K*** by 1 causes about half of the nodes to be removed in a community. Figure 3 shows the impact of increasing the value of ***K*** on the 10 largest communities in the social network, denoted by letters ***A*** to ***J***, which are the communities used in this study. In general, all the communities behave similarly as the average, except for community E which becomes very small for ***K=2***. Note that these communities' sizes, especially when ***K>0***, reflect fairly realistic sizes that one might expect not only in on-line communities but in real-world communities as well.

Figure 2. Average Community Size

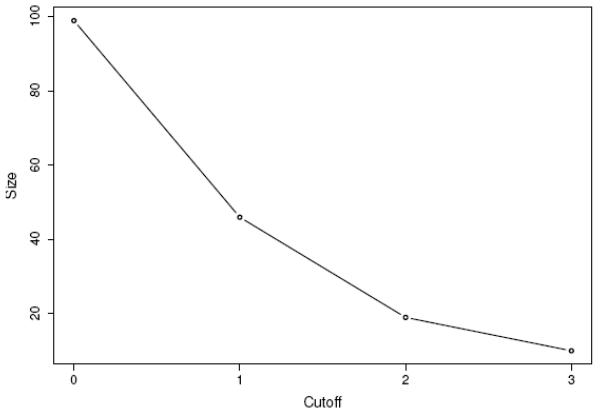


Figure 2. Average sizes of all communities in network under increasing values of *K*, the cutoff threshold

In our scenario measuring the performance of social network leaders, we consider the following network measurements to be informative in determining the health of a community and the performance of its leader:

- Edge-to-node ratio: number of edges / number of nodes.
- Transitivity (Clustering Coefficient): defined as the probability that the connecting nodes of a node are connected.
- Leader local transitivity.
- Leader betweenness: roughly defined by the number of shortest paths going through the node.

Figure 3. Sizes and Rank of the Top 10 Communities

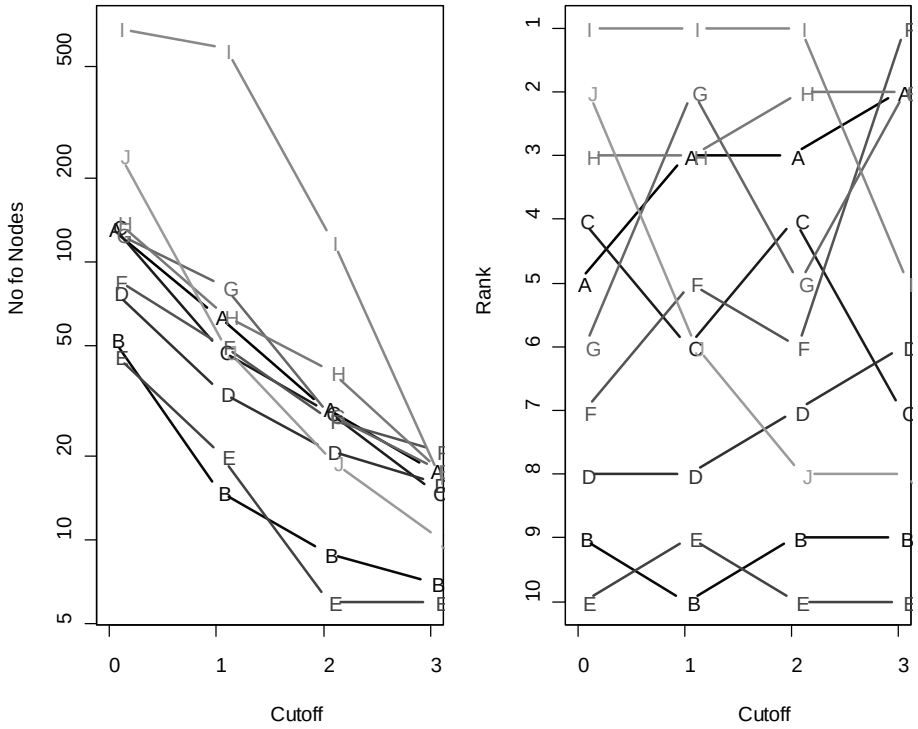


Figure 3. The sizes of the top 10 communities (on the left-hand side) and the rank of network sizes (on the right-hand side) with an increasing value of *K*, the cutoff threshold.

RESULTS

The plot on the left in Figure 4 shows the transitivity of ten communities with respect to different cutoff values $K \in \{0,1,2,3\}$. The ten communities are the top ten communities both in size and in the amount of cumulative communication. The plot on right displays the ranks instead of values. We can see that the value of transitivity generally increases with ***K*** for all ten communities. This is understandable since our friend definition implies that triangles are the smallest structural units of the networks. When edges without communication support are removed, the remaining ties between members are stronger, causing transitivity to increase, albeit with fewer nodes overall. The ranking of the communities in Figure 4, in general, is relatively stable across different cutoff values. We can see two groups of communities: the first is B, E and D, and the second is the other communities. The B-E-D community remains mostly top-ranked as ***K*** increases. Community I remains bottom ranked with consistently low transitivity. Some communities show a great amount of variance intransitivity. For example, community G is ranked as number 4 when ***K=0***, while it is ranked as number 9 when ***K=1***. The actual transitivity of G drops by more than 50%, which suggests that the transitivity measurement is not robust against systematic linking errors. This finding is opposite from the conclusion in Borgatti et al. (2006). We do understand, however, that they limit their results to random errors.

Figure 4. Transitivity and Rank of the Top 10 Communities

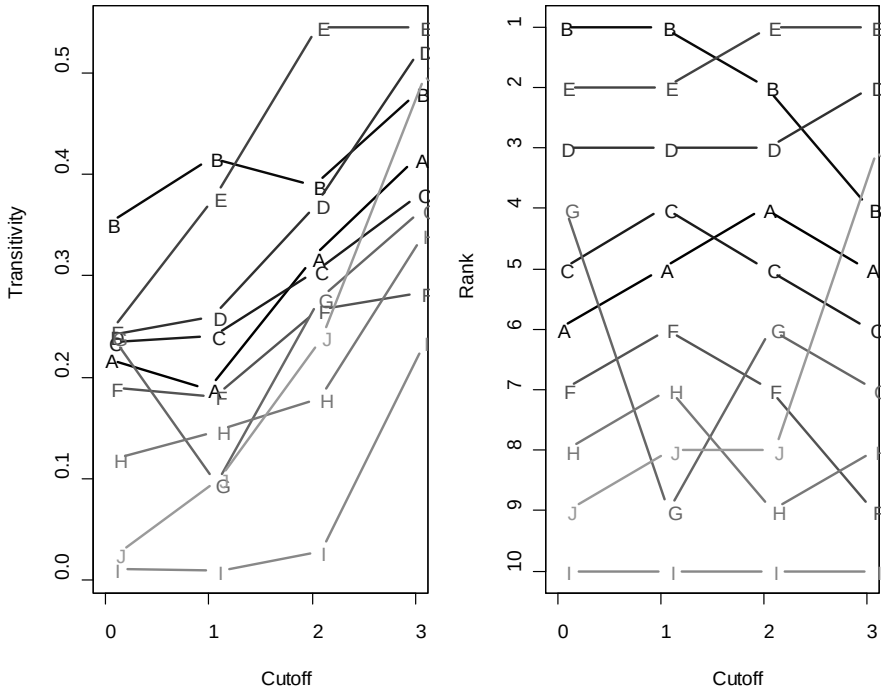


Figure 4. The transitivity of the top 10 communities (on the left-hand side) and the rank of the communities by transitivity (on the right-hand side) are shown with increasing values of K, the cutoff threshold.

Figure 4. Transitivity / Density and Rank of the Top 10 Communities

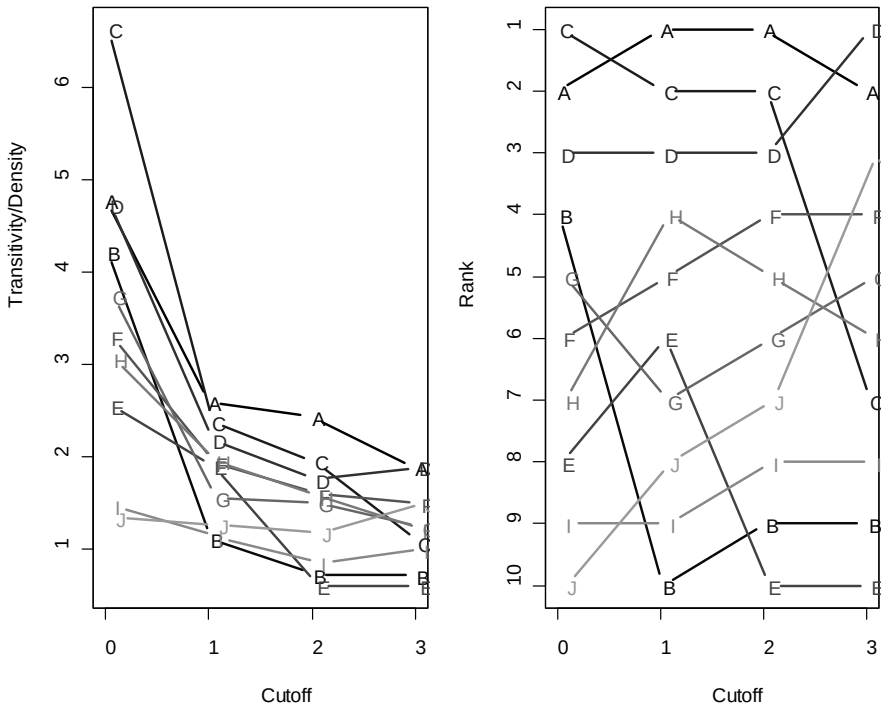


Figure 5. The transitivity and density ratio of the top 10 communities (on the left-hand side) and the rank of the communities by the transitivity and density ratio (on the right-hand side) are shown with increasing values of K, the cutoff threshold.

Latapy and Magnien (2008) suggest that the ratio between transitivity and density is more stable than transitivity, where density is defined by the number of edges over the number of possible edges. Note that transitivity represents average local connectedness; while density represents global connectedness. Our results in Figure 5 suggest otherwise, especially when we look at the ranks. The top-ranked group of communities is no longer B-E-D, but C-A-D, where the rank and value of C drops significantly with ***K=3***. Also, in the non-top-ranked communities, the variation amongst ranks is significant. However, we realize the study by Latapy and Magnien (2008) is more about the sample size problem and our study is more about data errors. While we can see a decrease in sample size as somewhat similar to an increase in ***K*** cutoff, our conclusion may not apply across different types of data issues.

We compare several other measurements of networks with different values of ***K***: the edge node ratios, the betweenness of leaders (a node level statistic), and the transitivity of leaders. By evaluating the ranks of communities and community leaders over increasing values of ***K***, we find further evidence to support our above results. While network measurements vary, the rank of the different objects being measured (communities or community leaders) can identify groups of objects that can maintain stability while having high or low ranks amongst the other objects.

DISCUSSION and CONCLUSIONS

Network effects, the percolation of simple changes in one location in the network that cascade throughout, are amplified in networks that exhibit power-law degree distributions, which are often found in real-world data. Unlike random networks that may contain a more uniform degree distribution, or have a more uniform distribution of high edge-to-node ratios, real-world networks often contain a few highly influential nodes or edges that play a major role in many network metrics. We used an empirical example to demonstrate that the application of social network analysis techniques in social networks with incomplete and erroneous data needs to be carefully evaluated.

In this study, we showed how instability in network metrics can arise when using behavior to validate self-elected friend relationships. Some of the metrics show the ability to maintain some consistency at the rank level when communities are grouped. Some communities tend to maintain strong rank, while others display a great amount of variance in their rankings. We suspect that there are some systematic issues that cause communities to show such behavior. Communities may tend toward a dichotomous outcome: friends are made more spuriously and lack any direct supporting communication, or friends are made between people that are already likely to communicate. Either way, it is of interest in our scenario of measuring communities to observe the 'health' of such communities.

While our scenario may seem limited, in terms of identifying the rank of communities or members, it is this kind of assessment which focuses the attention of the owners of such networks to communities that may require assistance in order to be successful. Also, we believe this kind of scenario will be practical in other real-world situations. To identify the communities or members who might be at-risk due to their social network would mean finding the members with the best or worst local structure that is relative to the community or network.

This study makes another contribution in demonstrating a technique that could be used to 'stress-test' the various possible definitions of relationships or ties within a social network. That is, by varying the degree to which a tie is validated with behavior data, like direct communication, one can see how stable various communities or groups are. In this study, by increasing the value of ***K***, the number of observed communications between members, and removing ties that did not have more than ***K***

communications observed, we can see that communities organized themselves into two groups. The first group consistently maintained a top rank under various network metrics, while the second group was consistently at the bottom rank. Within these two groups, the individual communities vary in rank for the different network metrics. In summary, by increasing the evidence required for friendship relationships, we see that one group of communities consistently has higher values of some network measurements. We might then be inclined to focus on the stability in rank, rather than on individual values of network measurements. This study is intentionally limited in scale and scope to demonstrate and support the hypothesis that subtle decisions made when creating social networks from real-world social network data can lead to varying stability in measurement outcomes. While the results may be intuitive, they are valuable in identifying how relative ranking of metric values aligns across metrics and in comparison to metric values.

REFERENCES

Adar, E. and Ré, C. (2007). Managing uncertainty in social networks. Data Engineering Bulletin, 30(2):23--31.

Bernard, H.R., Killworth, P., Kronenfeld, D., and Sailer, L. (1984). The problem of informant accuracy:the validity of retrospective data. Ann. Rev. Anthropology, 13:495--517.

Borgatti, S. P., Carley, K.M., and Krackhardt, D. (2006). On the robustness of centrality measures under conditions of imperfect data. Social Networks, 28(2):124--136.

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

Latapy, M. and Magnien, C. (2008). Complex network measurements: Estimating the relevance of observed properties. IEEE Conference on Computer Communications, 1660--1668.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computational social science. Science, 323(5915):721--723.

Lin, N. and Zhao, H. (2005). Are scale-free networks robust to measurement errors? BMC Bioinformatics, 6(1):119.

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

Rothenberg, R. B. (1995). Commentary: Sampling in social networks. CONNECTIONS, 18(1):104--110.

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 includes official board members and five committees:

Board Members

President:	George A. Barnett
Vice President:	Pip Pattison
Treasurer:	Thomas W Valente
Founder:	Barry Wellman
Members:	Phil Bonacich, Martin Everett, Katie Faust, Scott Feld, Anuska Ferligoj, Garry Robins, Ulrik Brandes, David Lazer

Committees

Finance Committee - Chaired by Treasurer Tom Valente

Conference Committee - Chaired by Sunbelt host Mario Diani

Web Committee - Chaired by webmaster (Chief Information Officer) Benjamin Elbirt

Publications Committee - Chair - TBD

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.

Awards Committee - The awards committee is comprised of four sub-committees:

- Visual Path Award, Dan Brass
- Freeman Award, Noshir Contractor
- Simmel Award, Russ Bernard, Chris McCarty, & John Skvoretz
- Microsoft Award**,** Janet Fulk, & Mario Diani