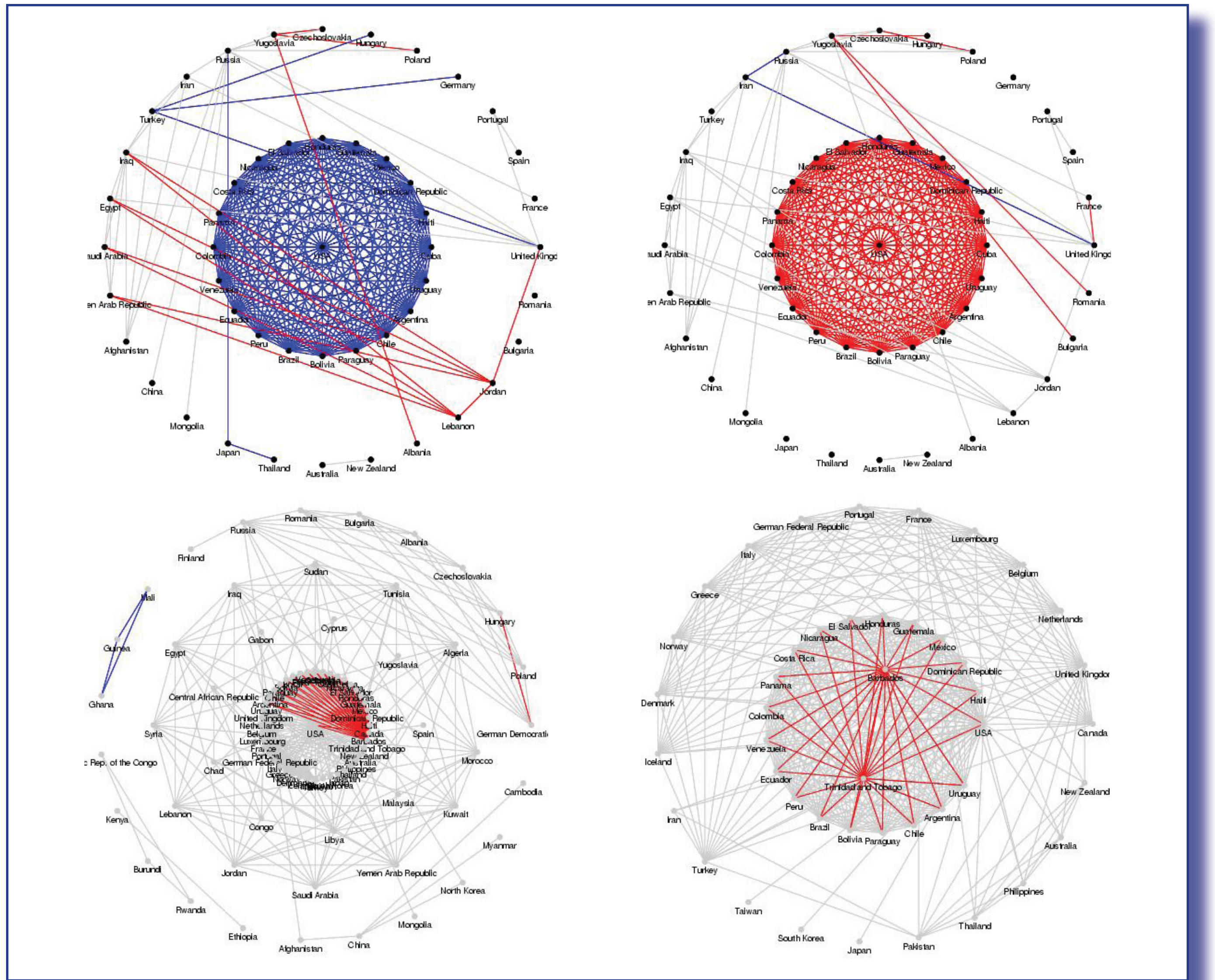
2008

CONNECTIONS

Volume 28 • Issue 2

CONNECTIONS

November 2008 *Volume 28 • Issue 2*



Inside this issue:

Looking at Social Capital through Triad Structures

Change in Connectivity in a Social Network over Time: A Bayesian Perspective

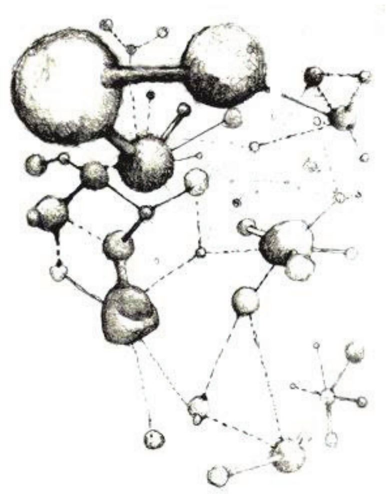
Determining An Actor's Network Capacity And Network Utilization:
A Markov Model Of Human Agency In Social Networks Scan Statistics for Interstate Alliance Graphs

Preference or Propinquity? The Relative Contribution of Selection and Opportunity to Friendship Homophily in College

Consent and Confidentiality: Exploring Ethical Issues in Public Health Social Network Research

Official Journal of the International Network for Social Network Analysis

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



***Front Cover*:** Images are from enclosed article titled "Scan Statistics for Interstate Alliance Graphs" by David Marchette and Carey Priebe. Graphs represent alliances between countries and show that not only are there more alliances among the countries allied with the central country, but there are also more alliances amongst the allies themselves. The top graphs depict the changes in the alliances between 1945 and 1946 (left), and between 1946 and 1947 (right). Blue lines denote edges that were removed, red lines are edges that were added, and lines edges are those which stayed the same. The central clique corresponds to the USA and countries in South and Central America. The bottom two graphs are illustrations of scan statistic methodology to detect unusual increases in the number of alliance among small sets of countries. Graphs represent different layouts of alliances in 1967, with the right graph corresponding to the large connected component.

Artwork on this page provided by Nicholas Coronges.

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. INSNA holds an annual conference that takes place in the United States for two years and in Europe every third year. Updated information about the **Sunbelt Social Network Conferences** can be found on INSNA's website at *www.insna.org*. Conferences bring together researchers from many disciplines to share current theoretical, empirical, and methodological findings around social networks.

INSNA includes official board members and five committees:

Board Members
George A. Barnett, President
Pip Pattison, Vice President
Thomas W Valente, Treasurer
Barry Wellman, Founder
Phil Bonacich, Martin Everett, Katie Faust, Scott Feld, Anuska Ferligoj, Garry Robins, John Skvoretz, Stan Wasserman

Committees

Finance Committee **-** Chaired by Treasurer Tom Valente

Program/Conference Committee **-** Co-chaired by the host(s) Chris McCarty and John Skvoretz of the Sunbelt Conference and Russ Bernard

Web Committee - Chaired by a webmaster (chief information officer) Benjamin Elbirt.

Publications Committee - Chair - TBD.
Composed of current and former and 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 additional members will be selected by the President.

Awards Committee - The awards committee would be composed 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

Looking at Social Capital through Triad Structures

Christina Prell
Sociological Studies, University of Sheffield, UK
John Skvoretz
USF College of Arts & Sciences, University of South Florida, USA

Abstract
The concept of social capital, which has gained wide currency in the literature, examines how actors' ties to others advantage or disadvantage them and the groups to which they belong. Two conceptually distinct types of social capital, closure and brokerage in Burt's (2005) terms, have been identified. In this paper, we propose a method by which brokerage and closure can be distinguished using a census of patterns of ties in triads of actors. We apply this method to network data gathered on 24 non-profit organizational actors. Our findings show when a network is characterized by brokerage or closure and how that network coincides with the presence/absence of trust and reciprocity. We conclude with a discussion on the nature of non-profits, and how the larger social context of network actors, in this case non-profits, play a role in interpreting the network structures uncovered via social network analysis.

Keywords: social capital, social networks, triad census, non-profits

Acknowledgements: We would like to thank Stephen Borgatti, Martin Everett, Richard Jenkins and David Phillips for helpful comments at various stages in this work.

Christina Prell is a Lecturer in Sociology at the University of Sheffield. Her research interests include social capital, computer-mediated-communication, stakeholder involvement in natural resource management, and more recently, using agent-based modeling for exploring the links between social capital and small worlds. She is currently working on a book on social network analysis for SAGE Publications.

John Skvoretz is Dean of the College of Arts and Sciences at the University of South Florida. His research areas include stratification and mobility, social network theory, group processes, power in exchange networks and the structure of social action systems. John has published over 70 research articles, edited books, and book chapters and has authored several computer programs used in network research and analysis.

For correspondence, please contact Christina Prell at the Sociological Studies Department , University of Sheffield, Elmfield Building, Northumberland Road, S10 2TU Sheffield, UK. email: c.prell@shef.ac.uk

Introduction

Social capital's rise in popularity in recent years is a phenomenon many have noted (Kadushin 2004; Portes 1998). Social capital consists of a network of relations and the resources embedded in those relations. As Putnam (1997) notes, social capital consists of the "features of social life—networks, norms of reciprocity and trust—that facilitate cooperation and coordination for mutual benefit" (pg. 31). Research on social capital shows it is not just the presence or absence of a relational tie that matters, but the overall structuring of the network (Baker 1990; Nahapiet and Ghoshal 1998). It is this issue of *how* a network is structured, and how a network characterized by that structure coincides with the presence/absence of trust and reciprocity, that our paper primarily addresses.

While social capital is typically used in studies of networks of persons, the concept is general enough that it can be applied to networks of other entities, such as groups or organizations. At this level of analysis, one would consider how some groups or organizations are connecting to other organizations, how these relations are structured, and what kinds of benefits these relations and network structures provide. The data explored in this paper represent various types of relations among a set of non-profits. These relations, reflecting the three main aspects of social capital, include trust, communication (or social relations), and the exchange of resources. Research shows that certain kinds of relations should coincide with one another; for example, actors who trust one another are more likely to help one another (Wellman and Frank, 2001).

In addition, how certain relations are structured should also coincide with the accessibility and ease in which resources get exchanged. This last issue of the structuring of relations is not as clear cut as it may seem for two reasons. First, there are two major and seemingly contradictory ideas about which structures embody social capital – in Burt's (2005) terms, is it closure or is it brokerage? We elucidate this contrast in the next section. Second, while various measures of social capital have been proposed (see Borgatti, Jones and Everett 1998), these measures are purely descriptive and do not address the question of whether a particular network exhibits more or less of a structural attribute than would be expected by chance. If we can determine that, in a given set of actors, a structural attribute is statistically significant then the co-presence of the exchange of resources for that set of actors would give us a firmer foothold upon which to state the claim that it is indeed the network structure which contributes to the presence of a particular beneficial outcome.

In succeeding sections, we briefly review the social capital literature, paying particular attention to drawing the distinction between closure and brokerage, and how these two structural attributes relate to social capital. We then propose a method based on the census of triads to identify when a network has more or less of a certain structural attribute than expected by chance. A triad census refers to how all triads in a network are distributed over different types of patterns of ties in the triad. There is a long history in social network analysis of using the triad census to detect systematic patterns that structure a network. We then introduce the data, in particular, the organizations that were examined and the methods used for gathering and analyzing these data. Finally, we present and discuss our results regarding social capital: to what extent can we determine the presence and/or absence of particular network structure and link the network containing that structure to the presence/absence of trust and reciprocity?

Social capital: Closure or Brokerage?

Within the social capital debate, two distinct network structures have drawn the most attention, these being cohesive networks, also referred to as 'network closure,' and networks composed of bridges and structural holes, also referred to as 'brokerage' (2005; Burt 2001). *Closure* refers to networks where actors are tied to one another through mutually reciprocated, strong ties. In addition to considering the strength of tie, measuring closure involves looking at the overall network structure. For example, common measure for closure is *density*.[1] This dense, closed structure is argued as enabling certain group behaviour and attitudes. For example, Coleman (1988; 1990) discusses closed, dense networks as conducive to social capital as they create feelings of mutual obligation and trust among members of the network. Coleman's (1988) work on social capital and school children suggests that a cohesive network made up of parents, teachers, and neighbours creates a supportive social structure resulting in more children within this structure completing their education. These findings are in keeping with Putnam's (2001) description of "bonding" social capital. For Putnam, bonding social capital refers to strong ties within a more or less closed, homogenous community that help community members get by, but not ahead.

Other social capital theorists do not consider closure in such optimistic terms. For example, Burt (2001) argues that norms can emerge from such networks that constrain social behaviour and inhibit innovation. Others note how closure might enable less socially-desirable groups to become stronger, for example, the Mafia and neo-Nazi groups. Finally, closure might work to keep socially isolated groups, such as immigrant communities and/or urban ghettoes, from becoming more integrated within mainstream society (Huysman and Wulf 2004; Narayan 1999). Thus, because closure is not always seen as helpful, a competing view of group structure has emerged that builds upon Granovetter's (1973) strength of weak ties argument. Granovetter (1973) argues that weak ties between actors are more likely than strong ones to carry non-redundant information across the disconnected segments of a network (Granovetter 1973; 2005). In doing so, weak ties are actually more important than strong ties for building social cohesion within heterogeneous networks.

One implication of this argument picked up by social capital theorists is the importance of weak ties as bridges, linking together more tightly bound, clique-like sections of a network. Putnam's (2001) notion of 'bridging social capital' reflects this idea: bridging social capital consists of ties that link across different community groups, thus providing actors access to diverse resources. Thus, bonding capital can help individuals or communities get by, but bridging capital is what helps individuals and communities get ahead. Similarly, Burt (2005; 2001; 1992; 2000) develops the concept of a "structural hole" to convey how bridging ties can benefit an actor. Structural holes are instances where two actors or two groups have no ties between them, but there is a third actor or group with ties to both of them, thus creating a "hole" opportunity for that third entity.

[1] Density is the proportion of possible ties in a network that are actually present.

In occupying a hole position, a third actor performs the role of a 'broker' between the two disparate entities. Such a broker position provides this actor key benefits,[2] but in addition, the two disparate entities linked through the broker also potentially benefit in being brought together. Without 'brokerage', these two entities would remain separate, and thus not gain access to one another's knowledge and other resources. Thus, while the broker has the strategic advantage of controlling when and how these two entities will interact, the fact remains that these two now have an indirect connection where none existed previously. This notion of 'brokerage' as a means of advantaging both the individual and the network as a whole has become an alternative form of social capital, and for Burt, one potentially more powerful than closure social capital.

Over time, empirical research has supported both of these competing views of social capital, and now scholars are arguing that a mixture of closure and brokerage is preferable for communities and groups. For example, Narayan (1999) argues that healthy societies need a combination of cohesive micro units (his examples are the family or tribal clan) that are then linked together through both weak and strong ties. Woolcock and Narayan (2000) make a similar argument for policy development purposes, noting that both bonding and bridging social capital are needed for communities to be truly healthy. Finally, Burt (2001) makes a similar argument for organizations, saying that "while brokerage across structural holes is the source of added value, closure can be critical to realizing the value buried in the structural holes" (pg. 52).

The above discussion summarizes the closure-brokerage argument. There are a number of measures for closure and brokerage that have been proposed by various scholars (see Borgatti, Jones, and Everett, 1998 for a good review). Yet, one criticism that can be made about all these measures is that all are purely descriptive, and thus do not address the question of whether a particular network exhibits more or less of a structural feature than would be expected by chance. Such a gap in the literature requires a baseline against which to calibrate the tendency to inhibit or enhance.

Thus, the question before us is how can we tell whether a particular network exhibits more or less of one or the other types of social capital than we would expect by chance? Similar questions have been addressed throughout the literature on network analysis. Holland, Leinhardt, and Davis, in a series of papers in the 1970s, for instance, proposed a way of investigating whether complex network patterns can be the result of local, triad structures (Davis 1977; Davis and Leinhardt 1972; Holland and Leinhardt 1970; 1972). The authors' method compared a census of observed types of triads to a census of triads expected by chance. Different types of triads were weighted differently in calculating a summary score, with the different weights determined by the property of interest. The method then determined whether the observed summary score was sufficiently different from the summary score expected by chance.

Davis, Holland, and Leinhardt's work linked such network structural features as partial orderings, ranked clusters, and transitivity to the presence or absence of

[2] Benefits could include such things as access to different sources of knowledge and resources.

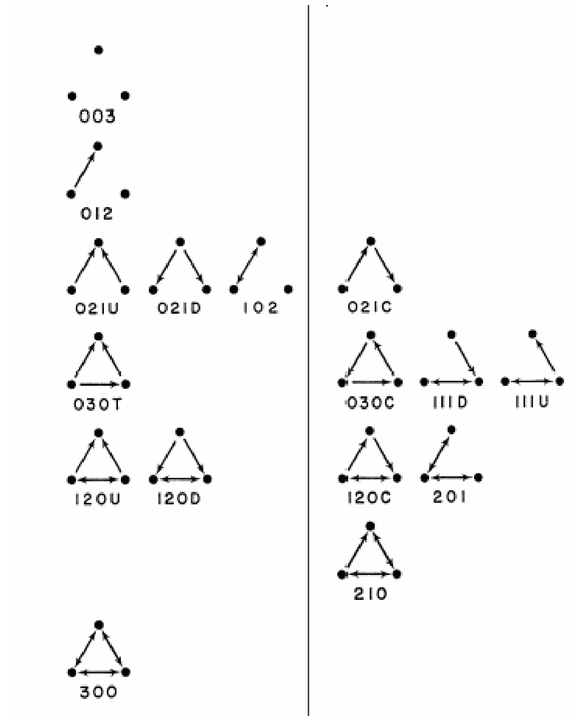
particular triad structures.[3] Their work was complemented by Granovetter's (1973) work on weak ties, in which the author argued that cohesion in large, heterogeneous networks could be explained 'more precisely' by looking at whether the network's triads contained strong, weak, or absent ties (pg. 1363). Since this series of papers, many scholars have come to view triads as valid 'building blocks' of larger, more complex, network structures (Laumann and Marsden 1982; Robins, Pattison and Woolcock 2005).

Based on the bulk of this previous work, we will now look at a triad census and describe how this census can be used for analysing brokerage and closure on the triad level.

Triad Census

A triad census refers to a census of all the possible types of triads that could be found in a given network (Holland and Leinhardt 1970). Such a census is depicted below in Figure 1**.**

Figure 1: Triad census[4]



This triad census codes the different triads according to their number of mutual, asymmetric, and null dyads. A mutual dyad refers to instances where one actor nominates another actor as someone they share a tie with, and this second actor reciprocates that nomination. An asymmetric dyad refers to instances where an actor nominates another, but this nomination is not reciprocated. Finally, null dyads refer to instances where neither actor nominates the other. For example, in triad 012 in Figure 1, the triad contains zero mutuals, one asymmetric, and two null dyads within its structure. In those instances where triads have the same number of dyads, the census uses letters to indicate the direction of the ties in the dyads. For example, in triad 021U, the "U" refers to ties directed upwards, whereas 021D refers to

[3] For a full description of each of these structural features, please refer to Wasserman and Faust, 1994.

[4] Adapted from Holland & Leinhardt, 1970.

ties directed downwards, and 021C refers to ties structured in a cyclic formation.

Linking this triad census to social capital, both closure and brokerage can be related to these triad structures. Closure can be taken to mean the number of triads which are fully connected, or 'closed.' In our triad census, such triads would be 030T, 030C, 120U, 120D, 120C, 210, and 300. The idea that social capital pertains to brokerage – meaning an actor rests between two unconnected actors – brings to prominence all the triad structures with one null dyad while the other two dyads have at least one connection. Such triads include 021U, 021D, 021C, 111D, 111U, and 201. Thus, for a network rich in closure social capital, one would expect there to be more of the closed set of triads than one would expect by chance. Similarly, a network rich in brokerage social capital should have more triads from the second set outlined above than expected by chance.

Linking these triad structures to our previous discussion on the differences between closure and brokerage, we can now distill our arguments down into the following:

Brokerage argument: weak ties tend to operate as bridging ties. Thus, networks composed of weak social ties should display a significantly high number of open triads. These open triads reflect the notion of brokerage. Networks characterized by brokerage should also be characterized by a high level of reciprocity. Trust is not an important factor for the brokerage argument.

Closure argument: strong ties tend towards closure. Thus, networks composed of strong social ties should display a significantly high number of closed triads and, also, be characterized by a high level of reciprocity and trust.

In what follows, we will investigate these two arguments using data gathered from an ongoing case study in the social capital of non-profit organizations.

Connected Kids

To explore these two arguments, we analyzed network data gathered on 24 non-profit organizations in Troy, New York. These non-profits were participating in a project lead by City Hall and the local university, Rensselaer Polytechnic Institute (RPI), to build an IT system for the local population. The project leaders of this IT initiative were two professors from RPI, and they selected representatives from these non-profits who were either a) administrators who had an overview of the sorts of programs and services their organization provided, or b) managers of youth programming within the organization. These organizational representatives dealt directly with youth service and programming issues and were thus in a more likely position to know how their organizations collaborated with others.

Network data were gathered through structured interviews in each respondent's own work setting. Respondents were handed a roster of the other 23 actors and their respective organizations, and questions were posed to respondents on their relationships with these actors' organizations. The data gathered reflected the three important aspects of social capital, social networks, trust, and reciprocity. These data, as they were conceptualized and measured, are described below:

1) *Social networks*. The literature on social capital notes how social capital is embedded in social relations (Coleman 1988; 1990; Foley and Edwards 1999; Lin 2001). Thus, we mapped out social relations among these 24 nodes, and conceptualized a social relation as any form of communication contact existing between

two organizations. To measure communication contact, we devised a frequency of communication contact item, whereby respondents were asked to rate how often they had any sort of communication contact with the 23 other organizations listed on the roster. Respondents could rate their frequency of communication contact from 1 to 7, with 1 as ‘not much communication contact’ and 7 as ‘a great deal of communication contact.’ These data, organized into a matrix, were then dichotomized to create two separate matrices: a less frequent communication contact matrix (compiled from data containing a value of 3 or less for communication contact) and a more frequent communication contact matrix (compiled from data containing a value of 4 or more). Thus, networks for strong and weak communication were created.

2) *Trust*. In conceptualizing trust, we are keeping stride with notions from the social capital literature, drawing on Coleman’s (1990) notion that trust arises from, and thus exists within, social relations. We, thus, see trust as specific to a relationship between two actors, and to measure this, we devised two attitudinal trust items adapted from Tsai’s (2000) network study on social capital. The first asked respondents to nominate those on the roster with whom they were willing to collaborate without a contract. The second item asked respondents to nominate those on the list whose information regarding youth respondents found trustworthy. Respondents’ answers were coded using 1 for those who were nominated as trustworthy and 0 for those who were not nominated.

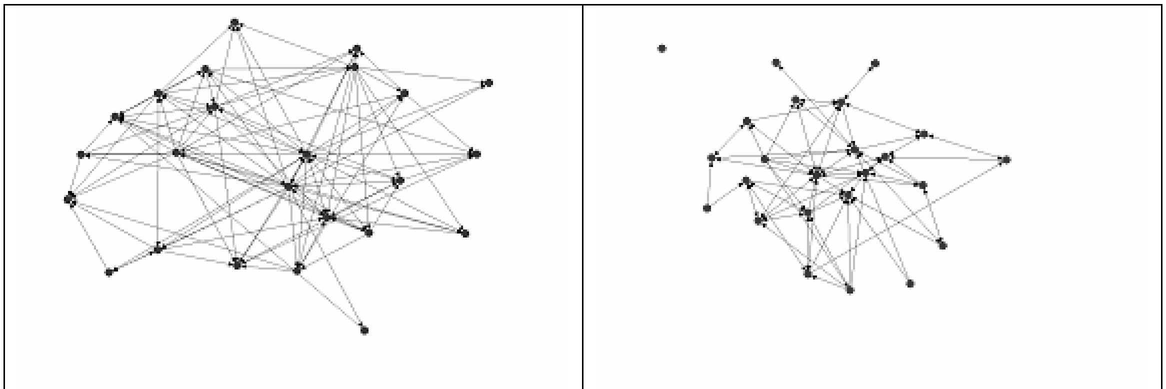
3) *Reciprocity (or ‘resource exchange’)*. In the social capital literature, the notion of reciprocity pertains to either a ‘norm of reciprocity’ (e.g. 1995; 2001; e.g. Putnam 1993) or the act of doing favors for others and exchanging resources with others (Coleman, 1990; Foley and Edwards, 1999). Within the network view of social capital, resources are seen as embedded within social structures, and the structuring of these social relations (e.g. size, level of cohesion, structural holes, etc.) determine how these resources get exchanged (Bourdieu 1986; Burt 2001; Coleman 1990). With this view of reciprocity as resource exchange among actors, we asked respondents questions about the giving and receiving of resources germane to the non-profit youth-service community within Troy. Respondents were asked the following: a) to whom they gave funds, b) from whom they received funds, c) with whom they shared clients, and d) with whom they had performed some sort of joint programming within the past year.

All actors receiving nominations were recorded as 1’s on the data sheet, and all other actors listed who were not nominated were coded as 0s.

Network Characteristics

Figure 2 displays graphs for the social networks based on strong and weak communication.

Figure 2: Graphs of Strong and Weak Communication Networks



Weak communication | **Strong communication**

Looking at Figure 2, neither network appears to have any sub-groups linked together through bridging ties. In addition, weak communication appears to have more ties linking the actors together than strong communication does. Strong communication, in addition to holding fewer ties, also contains an isolate. Taken together, we may tentatively conclude that weak communication is more cohesive than strong communication. This result is not surprising: it is easier for an actor to hold many weak ties, as strong ties are more demanding and energy intensive. Thus, a network can be more or less held together through weak ties, just as Granovetter (1973) discussed. From the graphs alone, we can not say anything about the overall distribution of ties, or the structure of these networks.

Triad Census Results

A triad census for each network was conducted,[5] testing for the two types of social capital mentioned earlier: closure and brokerage. The analysis involved counting the observed triads in each network and then comparing these observations with expected counts from a random network having the same number of dyad types (mutual, asymmetric, and null). As stated earlier, we expect that if a network is rich in closure, it should have more of the closed triads, i.e. more triads where every dyad contains at least one tie, than expected by chance. If a network is rich in brokerage, it should have more triads, expected by chance, in which exactly one of the dyads is null.

We tested these ideas by computing a weighted score in which each triad containing the key attribute (closure or

[5] Skvoretz and Agneessens's (2004) SPSS program script was used for this analysis.

brokerage) is weighted equally. This weighted score was computed from the observed distribution and from the expected distribution; the difference between the observed and expected score was taken and expressed relative to the standard deviation of the expected score. In effect, we computed two z-scores, one to test for the overabundance of connected triads, and another to test for the possible overabundance of triads with a structural hole.

Table 1: Tests for closure and brokerage for strong and weak communication

	STRONG	WEAK
CLOSURE	5.1*	4.2*
BROKERAGE	-0.2	3.7*

Note: * indicates $p < 0.01$

Strong communication shows significantly more closure triads than expected by chance, but the observed number of brokerage triads does not differ from chance. In this case, closure does not occur at the expense of brokerage. Weak communication ties exhibits both types of triads occurring at significantly levels greater than chance. That is, both closure and brokerage are present in this network, although perhaps to different subsets of actors. These findings are partly in keeping with what we would expect from the literature; we expect a network composed of strong ties to be characterized by closure, not brokerage. Similarly, we expect a network composed of weak ties to be characterized by brokerage, not closure. In this case, the network composed of weak ties comprises a significantly high level of *both* types of triads. One possible reason for this would be that all these organizations are located within the city boundaries. As Troy is not a large city, it is possible for nearly all organizations to have some sort of tie with one another. The strong ties, however, seem to be more dear. This is a point we shall return to shortly.

Trust and Reciprocity

For these data, we have uncovered the extent to which strong ties correspond to closure and weak ties, to brokerage. To what extent, then, are these social networks, composed of strong and weak ties, linked to trust and reciprocity? As a reminder, we are expecting a network characterised by strong ties and closure to be one in which trust and reciprocity are also present. A network characterized by weak ties and brokerage should also be one where reciprocity is present. Trust, in other words, should not figure strongly in the 'brokerage' network.

Below, we have performed QAP correlations[6] (Krackhardt, 1987) on the different relations on which we gathered data, which together reflect the three aspects of social capital. These correlations are found below**:**

[6] The correlation procedure used here is the QAP procedure, which is used to test the association between relations. As relations data have interdependencies that traditional case by variable data do not, calculating statistics for such data needs to account for these interdependencies. Thus the QAP procedure involves first computing Pearson's correlation coefficient between the corresponding cells of the two data matrices (relational data is stored and structured as matrix data). In the second step, it randomly permutes rows and columns of one matrix and recomputes the correlation. This is done hundreds of times in order to compute the proportion of times that a random coefficient is larger than or equal to the observed coefficient calculated in the first step. A low proportion (where $p < 0.05$) suggests a statistically strong relationship between the two matrices

Table 2: Correlations Across the Eight Relations

	Strong Communication	Weak Communication
1	0.38**	0.15**
2	0.47**	0.17**
3	0.17**	0.08*
4	0.28**	0.12**
5	0.12*	0.07
6	0.17**	0.07*

Note. * indicates $p < 0.05$ and ** indicates $p < 0.001$.

TRUST
1 Trustworthy information
2 Willingness to collaborate without a contract
RESOURCE EXCHANGE
3 Sharing clients
4 Joint programming
5 Giving funds
6 Receiving funds

Overall, Table 2 shows social, trust, and resource relations being intercorrelated. Both strong and weak communication ties show a similar pattern: both correlate with relations of trust and resource exchange, although stronger communication correlates slightly more with trust and resource exchange than weaker communication. Thus, stronger communication ties seem to be doing more of the work with this set of organizations.

The closure argument that trust and reciprocity are found in networks composed of strong ties receives support from these data (Coleman 1988; Coleman 1990). These non-profits seem to be utilizing their strong ties more so than their weak ties for accessing resources and forming joint programs, although both strong and weak ties are showing significant correlations with both trust and resource exchange.

Although more closure seems apparent in these findings, re-thinking the nature of this research site and these data could suggest otherwise: these relations are all, first and foremost, inter-organizational ones. As such, actors involved in these ties are embedded in different social circles, and a tie established between two such actors can be viewed as a bridging tie between two cohesive sub-groups. Thus, one might argue, a strong tie that bridges together different non-profits could, arguably, reflect the mixture of ‘closure’ and ‘brokerage’ advocated by certain scholars (e.g. Burt, 2001; Narayan, 1999).

Thinking still further about the nature of this research site, however, one could also argue that the prevalence of closure over brokerage is not surprising: non-profits based in local communities have a direct interest in developing long-term relationships with all members of their community. This means forming strong ties with clients, other local non-profits and municipal agencies. In this context, ‘getting ahead’ is not so much about forming ties with the outside world in order to bring in new resources and ideas as about forming strong ties within one’s world and making good use of the resources and ideas found therein. Thus, in certain contexts, in particular community-based non-profits, success and getting ahead are defined by the extent to which bonding and closure can be attained.

Summary and Further Thoughts on the Nature of Non-profits

These results have led us to some interesting findings: using the triad census approach helps us ascertain the extent to which closure and brokerage are statistically significant. In doing so, the census also helps us better ascertain the extent to which strong and weak ties correspond to closure and brokerage.

Once we established the extent to which strong and weak ties corresponded to closure and brokerage, we then assessed the extent to which these networks related to trust and resource exchange. We uncovered that a strong communication network not only contains more closure, but also relates to more trust and resource exchange. Our weak communication network contained an abundance of both types of triads, a finding we attribute to these non-profits all existing in close geographical proximity to one another. In addition, this network did correlate to trust and resource exchange, but not as strongly as the network based on strong ties. These findings both supported and were slightly at odds with the social capital literature, and we have interpreted this difference to the nature of our research site. Thus, a strong tie between two organizations can be interpreted differently than one, say, in a single organization. Additionally, in the context of non-profit community organizations, the goal of 'getting ahead', which is normally linked to weak ties and brokerage, seems out of place; such organizations have goals traditionally associated with closure, i.e. building strong community-based ties. Thus, searching for brokerage social capital among non-profits in a community might be an inappropriate research endeavor.

This last comment regarding the usefulness of brokerage within the context of non-profit research calls for more discussion on the nature of non-profits. In this study, non-profits had historical tensions and larger structural inequalities than our network analysis shows. Such tensions emerged in the qualitative data gathered alongside the network analysis. For example, the vast majority of non-profits in this study expressed a struggle with finding time, space and energy to form and maintain ties to other non-profits. This struggle was for a variety of reasons: non-profits lacked the staff numbers to network properly, which in itself was a result of small budgets, and they also needed to spend more time reporting back to a large number of external bodies, e.g. government and funding agencies. These pressures inevitably took their toll. As one respondent said to me in an interview, "we don't have time to network. None of us do. We're too understaffed (interview with non-profit employee, June 2001)." In the world of non-profits, where the realities of low staff numbers and pinched funds make networking difficult, an organizational actor might be more strategic about forming ties, only forming ones where a clear payback is evident. Thus, these actors might put more energy into fewer ties, relying more on their strong ties for the resource exchanges they need to survive.

The network capital literature tends to look at the social networks of individuals (Lin, Fu and Hsung 2001; Van Der Gaag and Snijers 2005; Wellman and Frank 2001), schools and business/organizational contexts (Burt, 2001; 2005; Coleman, 1988; 1990), and geographically-bound communities (Huysman and Wulf, 2004; Narayan, 1999). These are unique settings, and what tends to get overlooked is how such settings might play a role in the presence or absence of social capital. To what extent *ought* we pay attention to these unique contexts in which social networks are embedded? Paying attention to the wider context implies attention to additional structural features such as institutions, cultures, local histories, and socio-economic environments, things which social network analysts tend to ignore (Portes 1998). It is these features which might be the very influences we need to focus our attention on in order to gain a fuller sense of the role of networks within the social capital debate.

Research has long shown that non-profits need a different consideration from other organizational settings (Newman and Wallender 1978). For example, the 'corporate ethos' found in for-profit organizations emphasizes marketing and management strategies geared towards making profits that can be distributed to shareholders. Non-profits, on the other hand, are not accountable to shareholders, but rather to external funders, political, and governmental bodies, which constitute a larger range of external influencing bodies than those dealt with by for-profits. Thus non-profits must contend with greater external scrutiny of their activities and a greater degree of public accountability. They must balance more goals and services than those primarily guided by the for-profit motive (Potter 2001; Schwenk 1990). They also play a major role in all aspects of public policy (Bryce 2006). Thus, non-profits are, indeed, different than for-profit organizations. As such, the social capital discussion might need to account for this difference.

In conclusion, we have managed to push the social capital research forward a slight bit on methodological grounds through our use of the triad census. However, more research on social capital is still needed, both conceptually and methodologically, for exploring the links between larger social structures and the ones found via social network analysis.

References

Baker, W. E. . 1990. "Market Networks and Corporate Behavior." *American Journal of Sociology* 96, 589–625

Borgatti, S.P., C. Jones, and M.G. Everett. 1998. "Network measures of social capital." *Connections* 21:1-36.

Bourdieu, P. 1986. "The forms of capital." in *Handbook of theory and research for the sociology of education (pp.185-206)*, edited by J.G. Richardson (Ed.): New York: Greenwood Press.

Bryce, H.J. 2006. "Nonprofits as social capital and agents in the public policy process: toward a new paradigm." *Nonprofit and voluntary sector quarterly* 35:311-318.

Burt, R. 2005. Brokerage and Closure: An Introduction to Social Capital: Oxford: Oxford University Pres.

Burt, R. . 2001. "Structure Holes versus Network Closure as Social Capital." Pp. 31-56 in *Social Capital: Theory and Research*, edited by K. Cook N. Lin, and R. Burt (eds.) New York: Aldine de Gruyter.

Burt, R.S. 1992. *Structural Holes: The Social Structure of Competition*: Cambridge, MA: Harvard University Press.

Burt, R.S. 2000. "The network structure of social capital." in *Research in organizational behavior (pp. 31-56).* edited by R.I. Sutton & B.M. Staw (Eds.): Greenwich, CT: JAI Press. .

Coleman, J. 1988. "Social Capital in the Creation of Human Capital." *American journal of sociology* 94:95-120.

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

Davis, J. A. . 1977. "Sociometric triads as multi-variate systems." *Journal of Mathematical Sociological* 5:41-60.

Davis, J. A., and S. Leinhardt. 1972. "The structure of positive interpersonal relations in small groups." in *Sociological Theories in Progress*, edited by M. Berger, J. Zelditch, and B. Anderson. New York: Houghton-Mifflin.

Foley, M.W., and B. Edwards. 1999. "Is it time to disinvest in social capital? ." *Journal of public policy* 19:141-173.

Granovetter, M. 1973. "The strength of weak ties." *American journal of sociology* 78:1360-1380.

Granovetter, M. . 2005. "The Impact of Social Structure on Economic Outcomes." *Journal of Economic Perspectives* 19:33-50.

Holland, P., and S. Leinhardt. 1970. "A Method for Detecting Structure in Sociometric Data." *American Journal of Sociology* 76:492-513.

Holland, P. W., and S. Leinhardt. 1972. "Some evidence on the transitivity of positive interpersonal sentiment." *American Journal of Sociology* 72:1205-1209

Huysman, M. , and V. Wulf (Eds.). 2004. *Social Capital and Information Technology* Boston: MIT press.
Kadushin, C. . 2004. "Too Much Investment in Social Capital?" *Social Networks* 26:75-90.
Krackhardt, D. 1987. "'QAP Partialling as a Test of Spuriousness,'." *Social Networks* 9:171-186.
Laumann, E. O., and P. V. Marsden. 1982. "Microstructural analysis in interorganizational systems." *Social Networks* 4:329-48.
Lin, N. . 2001. "Building a Network Theory of Social Capital." Pp. 3-33 in *Social Capital: Theory and Research*, edited by N. Lin, K. Cook, and R. Burt. New York: Aldine De Gruyer.
Lin, N., Y. Fu, and R. Hsung. 2001. "The Position Generator: Measurement Technique for Investigations in Social Capital'." in *Social Capital: Theory and Research.*, edited by N. Lin, Cook, K., and R. Burt (Eds.): New York: Aldine de Gruyter.
Nahapiet, J., and S. Ghoshal. 1998. "Social capital, intellectual capital, and the organizational advantage " *Academy of Management Review* 23:242-266.
Narayan, D. 1999. "Bonds and Bridges: Social Capital and Poverty." in *Policy Research Working Paper 2167*. Washington, D.C.: Worldbank.
Newman, W., and H. Wallender. 1978. " 'Managing Not-For-Profit Enterprises'." *The Academy of Management Review* 3:24-31.
Portes, A. . 1998. "Social Capital: Its Origins and Applications in Modern Sociology." *Annual Review of Sociology* 22:1-24.
Potter, S. 2001. "'A Longitudinal Analysis of the Distinction between For-Profit and Not-For-Profit Hospitals in America' " *Journal of Health and Social Behavior* 42:17-44.
Putnam, R. 1997. "Democracy in America at Century's End." Pp. 27-70 in *Democracy's Victory and Crisis*, edited by Axel Hadenius. New York: Cambridge University Press.
Putnam, R.D. 1995. "Bowling alone: America's declining social capital." *Journal of democracy* 6:65-78.
Putnam, R.D. 2001. Bowling Alone: the collapse and revival of American community. : London: Simon & Schuster.
Putnam, R.D.. 1993. *Making democracy work: Civic traditions in modern Italy.* : Princeton, NJ: Princeton University Press.
Robins, G., P. Pattison, and J. Woolcock. 2005. "Small and other worlds: Global network structures from local processes." *American Journal of Sociology* 110:894-936.
Schwenk, C. . 1990. "Conflict in Organisational Decision-Making: an Exploratory Study of the to Sex in For-Profit and Not-For-Profit Organisations," *Management Science* 36: 436-448.
Van Der Gaag, M. , and T. Snijers. 2005. "The Resource Generator: Social Capital Quantification with Concrete Items." *Social Networks* 27:1-29.
Wellman, B., and K. Frank. 2001. "Network Capital in a Multi-level World: Getting Support in Personal Communities." Pp. 233-73 in *Social Capital: Theory and Research*, edited by N. Lin, Cook, K., and R. Burt (eds.) New York, Aldine de Gruyter.
Woolcock, M. , and D. Narayan. 2000. "Social Capital: Implications for Development Theory, Research, and Policy." *The World Bank Research Observer* 15:225-249.

Change in Connectivity in a Social Network over Time: A Bayesian Perspective

Susan Adams **Nathan Carter** **Charles Hadlock**
Dominique Haughton - *Bentley University, Waltham, Massachusetts*
George Sirbu - *Thomson Medstat, Ann Arbor, Michigan*

Abstract

In this paper, we propose a Bayesian methodology for examining differences between statistics of a social network at two distinct points in time. The problem has been of interest for some time in the social networks community because it is quite difficult to test whether differences over time in statistics such as overall network connectivities are significant. Several issues make this problem challenging: links in a social network tend to be dependent, and the networks at the two different points in time are likely to be dependent as well. This implies, for example, that bootstrapping a social network to address this problem may be impractical. This paper expands on a previously published Bayesian version of the p_1 model for social networks with random effects, which allows for dependence between the edges of the networks. We use the software Winbugs to obtain posterior distributions for the difference in connectivity over time and for the correlation between each actor's connectivities in the network at both points in time. We assume that this correlation is the same for all actors. We illustrate our methods with the case of a social network of collaborations (joint publications) between departments of a business university where interdisciplinary work was actively promoted. Our methods allow us to compare the tendency to make collaborative links across departments before and after the administrative initiatives.

Keywords: *Social Networks Over Time, Connectivity, p_1 models, Bayesian analysis*

Acknowledgments: We would like to express our sincere gratitude to the authors Gill and Swarz (2004) for kindly making the Winbugs code they wrote for their models available to us. It was very helpful in building the Winbugs code needed in this paper. The Winbugs code we used to simulate the posterior distributions is available on request from the authors.

Susan M. Adams (PhD Georgia Institute of Technology) is Associate Professor of Management at Bentley University where she teaches leadership and management consulting courses. Her current research focuses on how organizations affect the careers of executives and professionals, particularly in changing environments. She is a former chair of both the Careers and Management Consulting Divisions of the Academy of Management. **Nathan Carter** (PhD Indiana University) teaches mathematics at Bentley University and publishes in diverse areas, including mathematical logic, group theory visualization, and social network analysis. His book Visual Group Theory will be published by the Mathematical Association of America in 2009.
Charles Hadlock (PhD University of Illinois) is an applied mathematician with a combination of industrial and academic experience. He is currently Professor of Mathematical Sciences and Trustee Professor of Technology, Policy, and Decision Making at Bentley University. His current research focuses on complexity theory and the collapse of complex organizations and societies.
Dominique Haughton (PhD M.I.T.) is Professor of Mathematical Sciences at Bentley University. Her research interests include statistics and marketing, the analysis of living standards data, international statistics and data mining. She is editor-in-chief of the journal Case Studies in Business, Industry and Government Statistics (CSBIGS).
George Sirbu (PhD Michigan State University) is a statistician at Thomson Medstat in Ann Arbor, Michigan, where he specializes in the analysis of health care data.

For correspondence, please contact Dominique Haughton at dhaughton@bentley.edu

Introduction

The comparison of two networks at two (or more) different points in time has been approached in social network literature in the following ways. Wasserman and Iacobucci (1988) attempted to test the equality of model parameters across two or more time periods by where extensions of the p_1 model are used, but where dyads at a given point in time are still assumed to be independent. To try to overcome the difficulty of both relaxing the assumption of independent dyads (links between pairs of factors) and the observations of the same relation at two or more time points (which of course are not likely to be independent), Snijders (1996) proposed to model the longitudinal network as a simulated Markov chain, with parameters estimated via the method of moments. However, this method does not readily lend itself to testing the significance of a difference in parameter values across time periods, and is quite difficult to implement.

The paper by Faust and Skvoretz (2002) goes in another direction by introducing methods which rely on the p^* model proposed by Wasserman and Pattison (1996) and attempts to compare networks with different sizes, across different time periods and even different entities. The paper also includes a useful review of the literature to date on issues of comparing different relations on the same actors, or the same relation at several points in time, or across different groups, inclusive of the papers mentioned immediately above. However, the emphasis in the Faust and Skvoretz article is not on testing whether a particular aspect of the network has changed significantly over time.

It is this last issue – investigating shifts in a particular network parameter across two points in time – which is the focus of this article. The problem is a challenging one, because, for instance, attempts to use procedures such the bootstrap to obtain estimates of standard errors hit against the problem that dyads are not independent, so that it is impractical to bootstrap a socio-matrix of ones and zeros which are not independent, even when only one network is under consideration (at one fixed point in time). Bootstrapping independent and identically-distributed quantities such as regression residuals is possible (and is indeed common), but, unfortunately, to bootstrap a socio-matrix, one would need to respect the dependency structure of the ones and zeros while sampling, which is very complicated. Even more intractable is bootstrapping two dependent socio-matrices made up of dependent ones and zeros.

We, therefore, propose to follow a Bayesian approach introduced in the context of social networks by Wong (1987) and extended by Gill and Swartz (2004). As outlined in the latter article, the p_1 model, with fixed effects proposed by Holland and Leinhardt (1981), was improved by Wong (1987) by incorporating random effects. The main difference is that the original p_1 model assumes independence between the links, whereas with random effects, some dependence structures can be incorporated.

The article is organized as follows: section 2 describes the data and the context that will be used to illustrate our approach; section 3 presents the Bayesian model we propose for our networks; and section 4 gives results and conclusions.

Data

In the Fall of 2001, our institution, a large, independent business school in Massachusetts, began a program specifically intended to encourage interdisciplinary research collaborations among the faculty.

The institution has been accredited by the Association to Advance Collegiate Schools of Business International (AACSB) since 1991 and has made major investments to establish a strong teaching and research capability, especially emphasizing the intersection of business and information technology.

In the period leading up to and immediately following the initial AACSB accreditation, the institution was entering a transition phase from a predominantly undergraduate teaching institution to a more comprehensive university. Buoyed by a measurable growth in reputation, rankings, and student quality, a development objective evolved into transforming the institution into a business university. It became clear that faculty could reach a new level of reputation and contribution only by focusing on major current issues facing the business world, rather than spending much of their time in isolated academic niches that might have much less impact on the practice of business. Such major current issues are almost invariably interdisciplinary in nature.

The institution decided to work proactively to facilitate the formation of faculty research teams focused on major issues that suited their interests and backgrounds, and the administration first issued a request for proposals (RFP) to the faculty near the end of 2001. Funds could be requested for a variety of purposes, such as course reductions, summer stipends, and various expense items[1]; even to the point where an intensive faculty research seminar might substitute for part of the teaching load of major participants.

Considering the faculty as a social network, we let each faculty department be a node (vertex) in the network, and connections among departments be determined by faculty in a department doing scholarly work with faculty in other departments. Therefore we consider the network of co-publication between departments (1 if at least one article was jointly authored by at least one member of each department, 0 if not) at two distinct periods of time, in ’00-’01 and in ’03-’04. The reason for this particular split is because of the time when some programs were launched (e.g. RFP program) and what authors believe to be a turnaround point in the university development. Note that the network is undirected.

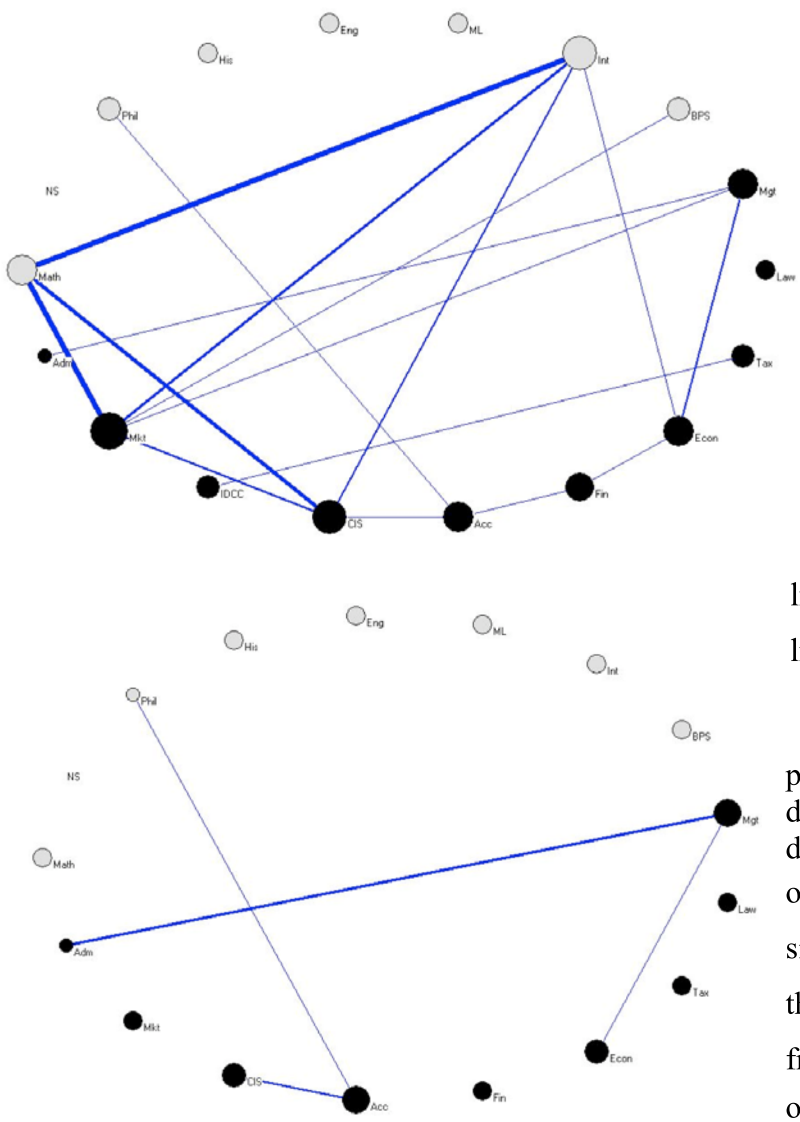
We obtain our data from our faculty research database, which had recently been updated in connection with our re-accreditation from the AACSB. The visualization of the two networks at the two points in time is given in Figure 1. We only took into consideration journal articles, since we felt that journal articles were the best category of scholarly work to account for the quality and quantity of the faculty scholarship activity.[2]

Visual evidence from Figure 1 would seem to indicate a clear increase in the connectivity of the network. We will now examine this issue more formally and describe our Bayesian model in the next section.

[1] Expense items include travel, data acquisition, student assistants, ongoing research seminars, etc.

[2] Since the RFP initiative had the underlying intent to encourage a better interaction among internal faculty resources, our data set does not record the departments for authors external to the institution.

Figure 1: Collaboration Network in 2000-2001 (top) and 2003-2004 (bottom)



Gray vertices are Arts and Sciences departments, black vertices are Business departments or Administration. The thickness of the lines is proportional to the number of joint publications between the two nodes it joins; the size of a vertex is proportional to the node's degree, including itself twice. Abbreviations for the departments are as follows: NS for Natural Sciences, ML for Modern Languages, Int for International Studies, BPS for Behavioral and Political Science, CIS for Computer Information Systems, IDCC for Information Design and Corporate Communication, Adm for Administration.

Model

Since in our case the ties are non-directional, the p_1 model can be written in the following way:[1]

$Y_{ij11}^{k} = 1$ if i and j are connected, 0 otherwise

$Y_{ij00}^{k} = 1$ if i and j are not connected, 0 otherwise

$$\ln P(Y_{ij00}^{k} = 1) = \lambda_{ij}^{k}$$
$$\ln P(Y_{ij11}^{k} = 1) = \lambda_{ij}^{k} + \theta^{k} + \alpha_{i}^{k} + \alpha_{j}^{k} \ .$$

The index k denotes the time period and indices i and j refer to departments. Because each pair of departments (i, j) can be either linked or un-linked, the matrix Y_{ij00}^{k} is a simple opposite of the matrix Y_{ij11}^{k} in the sense that Y_{ij00}^{k} can be obtained from Y_{ij11}^{k} by replacing zeros with ones and ones with zeros. The matrix Y_{ij11}^{k} is often referred to as the socio-matrix, with its ones indicating where a link occurs. The probability $P(Y_{ij11}^{k} = 1)$ represents the probability of a link occurring between departments i and j, at time k, and $P(Y_{ij00}^{k} = 1)$ represents the probability that no such link exists.

We will make the convention that $k = A$ for the '00-'01 social network and $k = B$ for

[1] We adopt the Wasserman and Faust (1994) formulation.

the '03-'04 social network. The parameter θ^k is called the choice parameter and is a measure of the overall connectivity of the network, and α_i^k is the attractiveness parameter correspondent to node i for period k. We observe also that λ_{ij}^k is not a parameter, but rather a fixed constant subject to the constraint, $P(Y_{ij00}^k = 1) + P(Y_{ij11}^k = 1) = 1$.

In a Bayesian analysis, unknown parameters are considered random variables with a distribution referred to as the prior distribution, which reflects knowledge we might have about the parameters even before any data are collected. The analysis produces a posterior conditional distribution of the parameters given the data, using a MCMC (Monte Carlo Markov Chain) simulation procedure, the details of which are well documented (Congdon 2001 or Winbugs 2006). The posterior distribution is proportional to the product of the likelihood function and the prior density, and as such takes into account both the prior distribution and the data (i.e. the observed networks). Posterior densities typically *cannot be calculated in closed form*, which makes simulating from them difficult; progress in MCMC methods has made it possible to simulate from such posterior densities without fully computing their densities, and has helped give rise to the fast development of Bayesian applications. We consider the following prior distributions for the parameters of interest:

Given Σ,

$$\begin{pmatrix} \alpha_i^A \\ \alpha_i^B \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right)$$

$$\begin{pmatrix} \theta^A \\ \theta^B \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma \right),$$ independently of

$$\begin{pmatrix} \alpha_i^A \\ \alpha_i^B \end{pmatrix},$$

with Σ given by

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma^2 \end{pmatrix}$$

distributed as

$$\Sigma^{-1} \sim \mathrm{Wishart}\left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, 2 \right),$$

where ~ denotes "is distributed as". The choice of our prior distributions is standard for this particular situation and implies no particular prior knowledge about where parameter values might be concentrated. For instance, the bivariate normal distribution we have chosen for the vector of attractiveness values of a node (for both periods), with a precision matrix[2] distributed according to a Wishart distribution as above is standard (Congdon, 2001). Note that we have assumed that the vectors with components α_i^A and α_i^B are independent a-priori of the vector with components θ^A and θ^B given Σ, because there is no particular reason for our prior belief about the overall connectivity of the network to be related to our prior belief about the attractiveness of each department. On the other hand, it seems sensible to assume that the precision

[2] Precision matrix is the inverse of the covariance matrix.

of our prior knowledge (represented by Σ^{-1}) is the same for the vectors with components α_i^A and α_i^B and the vector with components θ^A and θ^B, and that the a-priori correlations between α_i^A and α_i^B are the same for all *i*, and equal to the correlation between θ^A and θ^B.

We are interested in the difference $\Delta = \theta^B - \theta^A$ in the choice parameter θ from the former social network '00-'01 to the latter one '03-'04. We expect to find that the posterior distribution of Δ is concentrated for most of its range in the set of positive numbers. We are not arguing that this change can be entirely attributed to the programs mentioned in the introduction but merely observe that the difference Δ is a-posteriori likely to be positive. However, it is our belief that the success of the program was part of the positive change that can be seen from the analysis.

In the next section, we report the results of using an MCMC procedure, such as implemented in the software package Winbugs to generate random draws from the posterior distribution of parameters of interest. Note that Winbugs does not require that the user provide expressions for auxiliary distributions used in the procedure, only that the model which generates the data and the prior distribution be specified.

Results and Conclusions

In the table and figures below we present the results of the MCMC analysis from the model outlined in the previous section with graphs representing kernel densities for the posterior distributions of the two main parameters of interest. The statistics presented in Table 1 and the data which were used to create kernel densities for the posterior distributions of parameters of interest arose from a Winbugs analysis where we generated 200,000 iterates of the MCMC procedure. The first 4,000 iterations were used as a “burn-in”, so the summary statistics in Table 1 are in fact based on the 196,000 remaining iterates. This is necessary because the MCMC chain becomes stationary typically only after a certain number of iterations, so it is safe to compute posterior moments (means and standard deviations) from iterations arising once the chain has become stationary (further discussion of the convergence of the process is given below).

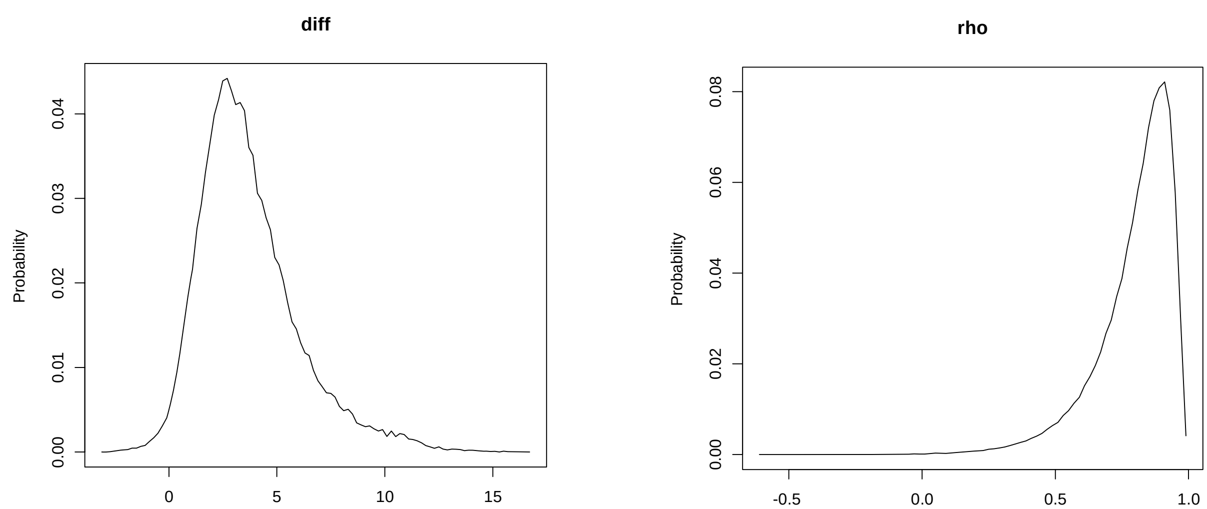
The parameters we focused on are the difference $\Delta = \theta^B - \theta^A$, the choice parameters θ^A and θ^B, their standard deviations σ_A and σ_B, and the correlation ρ between θ^A and θ^B. We also included summary posterior moments for the attractiveness of a particular department at both points in time.

Table 1: **Summary Statistics of the Posterior Distribution of Parameters of Interest, for 196,000 Iterates**

Parameter	Mean	SD	MC error	2.50%	Median	97.50%
Δ	3.69	2.266	0.0824	0.282	3.313	9.346
θ^A	-7.15	2.705	0.1082	-13.64	-6.722	-2.962
θ^B	-3.46	1.286	0.0489	-6.231	-3.389	-1.096
ρ	0.798	0.1425	0.0016	0.425	0.835	0.964
σ_A	4.39	1.758	0.0471	2.1	4.022	8.838
σ_B	2.645	0.7466	0.0119	1.568	2.518	4.454
a[1,1]	-0.07125	0.7972	0.02401	-1.643	-0.07444	1.518
a[1,2]	2.257	1.486	0.05393	-0.2261	2.092	5.683

As we can see in Table 1, the 2.5% percentile (0.282) from the posterior distribution of the difference Δ is above zero so we can conclude that it is a-posteriori very likely that the choice parameter for interdisciplinary research in the university has increased from '00-'01 to '03-'04. This is also clear from Figure 2, left panel, where the posterior density of Δ is displayed.

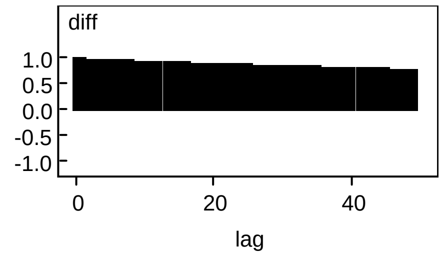
Figure 2. Posterior Distributions for Δ and ρ



It is also interesting to notice that the correlation between choice/attractiveness parameters in '00-'01 and '03-'04 is quite high with its posterior mean estimated at 0.798 and its left-skewed posterior distribution (Figure 2, right panel). This makes sense given the nature of the data, where relations among the actors in the same network are likely to be preserved over time. Nevertheless due to the increased activity in the network in the latter period, the observed increase in the estimates, from 2.645 to 4.39, for the posterior mean standard deviation of the expansiveness/attractiveness parameter is quite natural; in the latter period, departments became more diverse in their propensity to engage in joint research with other departments.

When using MCMC techniques to sample from posterior distributions, an issue arises about the stationarity of the sequence of draws, which one would wish to occur after a number of “burn-in” draws. A visual examination of the history of the draws will usually suffice to conclude to stationarity; however there is always the risk that that MCMC sampler might get stuck in a sub-region of the parameter space instead of exploring the sample space, particularly if successive draws are strongly auto-correlated. For that reason, it is desirable that the auto-correlation between successive draws should decrease rapidly with the lag for each parameter of interest. It is not uncommon, and happens in the case for example of our parameter Δ on the difference between the choice parameters (referred to as diff on Figure 3), that this auto-correlation in fact dies down slowly. This can be seen clearly on Figure 3.

Figure 3. Auto-correlation Between Draws of the Posterior Distribution of Δ, with All 196,000 Draws Included



In that case, one can sample one out of, for example, 100 draws of the posterior, which, typically, will remove the auto-correlation problem, and then compare the posterior summaries of the smaller sample with the full sample. We are grateful to an anonymous referee for suggesting this very useful idea. In Figure 4, we can see that the auto-correlation problem has now disappeared.

Figure 4. Auto-correlation Between Draws of the Posterior Distribution of Δ, with One Out of 100 Draws (sample with 1960 draws)

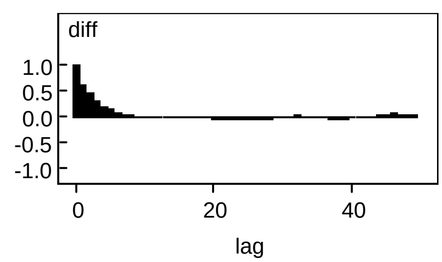


Figure 5 indicates that stationarity is not in question, particularly since the draws on Figure 5 hover about a very similar level to the level featured on a history graph for the full sample.[1] This is further supported by the large number of iterations we used. We conclude from examining this set of graphs

[1] A history graph of the full sample is not given here because it is very similar to Figure 5.

that it is quite unlikely that the MCMC chain was trapped in a sub-region of the sample space.

We have presented auto-correlation graphs for Δ and history graphs for Δ and ρ, but graphs for the remaining parameters of interest show that stationarity holds for them as well.

Figure 5: History of draws from the posterior of Δ and ρ, on the basis of one out of 100 draws (sample with 1960 draws)

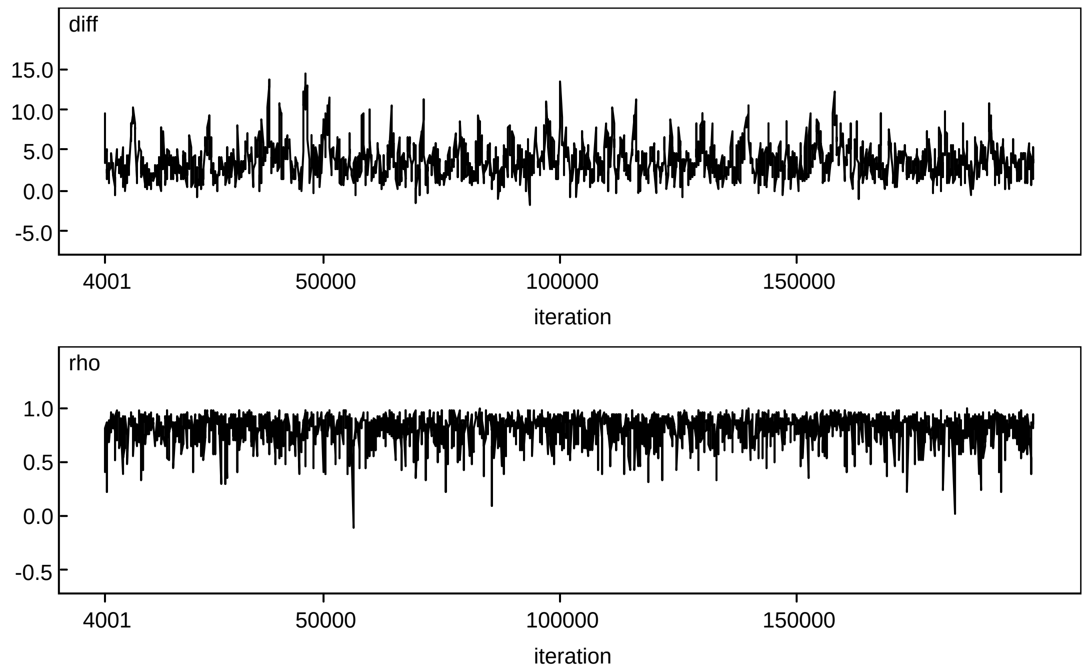


Table 2 reveals that summary statistics of the posterior distribution of the parameters of interest evaluated on the sub-sample of one out of 100 draws are close to those in Table 1 evaluated for the whole sample.

Table 2. Summary Statistics of the Posterior Distribution of Parameters of Interest, on the Basis of the One Out of 100 Sample of 1960 Draws

Parameter	Mean	SD	MC error	2.50%	Median	97.50%
Δ	3.671	2.263	0.1144	0.2885	3.329	9.289
θ^A	-7.134	2.7	0.171	-13.52	-6.714	-2.912
θ^B	-3.463	1.285	0.0804	-6.248	-3.385	-1.161
ρ	0.8005	0.1391	0.003079	0.4407	0.8372	0.9668
σ_A	4.385	1.707	0.06254	2.1	4.017	8.774
σ_B	2.649	0.7586	0.02124	1.566	2.534	4.441

To sum up, we have found that the Bayesian methodology provides a convenient and effective way of deciding whether a parameter of interest in a social network has changed over time. Of course, because of its flexibility, the methodology lends itself to other situations which might prove intractable otherwise, and a whole variety of social network models can be formulated in the Bayesian framework. For instance, Tallberg (2003) proposes a Bayesian approach to uncovering blocks within networks of actors which are similar in the sense that their probabilities of forming links with other actors are the same.

It is therefore quite likely that Bayesian methods will find further applications to problems of interest to social network researchers where no other method is readily available. However, a caveat is in order when using Bayesian methods (or even when using more classical likelihood methods). Adding too many parameters carries with it the risk of over-parameterization of a model. Over-parameterization occurs when several values of the parameters give rise to the same value of the likelihood function, leading to a situation where some parameters may not be identifiable. It is desirable to avoid over-parameterization because it can lead to some convergence problems in the MCMC procedure, even if the problems can to some extent be overcome by the choice of suitable priors. However, it is not always easy to know if a model is over-parameterized; one may be alerted to it only by unusual behavior in the MCMC iterations.[1]

We also note that when attempting to compare connectivities of two networks over time, it is advisable to make sure that the networks have about the same size, as is the case here (the number of actors differs by only one between the two time periods), since several network parameters are known to depend rather critically on network size (Anderson, Butts and Carley 1999).

[1] We refer the reader to papers by O'Neill (2005) and Rannala (2002) where the issue is discussed.

References

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

Congdon, P. 2001. Bayesian Statistical Modelling. New York: Wiley.

Faust, K. and Skvoretz, J. 2002. Comparing networks across space and time, size and species. *Sociological Methodology*, 32: 267-299.

Gill, P. S. and Swarz, T. B. 2004. Bayesian analysis of directed graphs with applications to social networks. *Applied Statistics*, 53(2): 49-260.

Holland, P. W. and Leinhardt, S. 1981. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76: 33-65.

O'Neill, B. 2005. Consistency and identifiability in Bayesian analysis. Preprint School of Finance and Applied Statistics, Australian National University, available at http://ecocomm.anu.edu.au/research/papers/pdf/05-09.pdf.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of Phylogeny. *Systematic Biology*, 51(5): 754-760.

Snijders, T. 1996. Stochastic-oriented models for network change. *Journal of Mathematical Sociology*, 21(1-2): 149-172.

Tallberg, C. 2003. A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology*, 29(1): 1 – 23.

Wasserman, S. and Faust, K. 1994. Social network analysis: Methods and applications. Cambridge, England: Cambridge University Press.

Wasserman, S. and Pattison, P. 1996. Logit models and logistic regressions for social networks: An introduction to Markov graphs and p^*. *Psychometrika*, 61(3): 401-425.

Wasserman, S. and Iacobucci, D. 1988. Sequential social network data. *Psychometrika*, 53(2): 261-282.

Winbugs. 2006. The Bugs project, Winbugs, http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

Wong, G. Y. 1987. Bayesian models for directed graphs. *Journal of the American Statistical Association*, 82: 140-148.

Determining An Actor's Network Capacity And Network Utilization: A Markov Model Of Human Agency In Social Networks

Binod Sundararajan
School of Business Administration, Dalhousie University

Abstract
The focus of this paper has been to put forth a Markov model that will provide information to actors in a network about the optimum capacity of alters in their social networks and how to maintain social-temporal relations with their contacts and resources with optimum efficiency. This model takes advantage of the similarities between the concept of human agency and Markov random processes. It takes into account the fact that present experiences are the sum total of past iterational and habitual experiences and the present practical-evaluative capacity to evaluate these past experiences. The model then adapts the Markov process and the Erlang blocking formula used in telephony, when it uses the present practical-evaluative experience to create a projective capacity toward the future state of the social network. This will then provide the actor (ego) with information about the efficient utilization of his/her channel/network capacity and the number of contacts or resources he/she would need to maintain to achieve the future stability of the network contacts and resources.

Binod Sundararajan is an Assistant Professor for Business Communication at the School of Business Administration in Dalhousie University, Halifax, Nova Scotia, Canada. His research interests include Organizational, Professional & Business Communication, Computer & Technologically Mediated Communication, Computer-Supported Collaborative Learning & Work, Social Network Analysis, Small Group Communication and use of information and communication technologies in diverse learning, work and entrepreneurial efforts and environments. He has pursued a hybrid approach to his research employing quantitative and qualitative methodologies to reach across disciplines to develop theoretical frameworks, perform historical data analysis and social network analysis of social movements, blog analysis, and some usability studies using eye-tracking equipment, and tactile devices for the visually impaired. He his starting his academic career and has about 14 conference presentations to his credit and four journal publications.

For correspondence, contact Binod Sundararajan, School of Business Administration, Faculty of Management, Dalhousie University; Email: binod@dal.ca.

Introduction

Human agency as described by Emirbayer and Mische (1998), is conceptualized as a temporally embedded process of social engagement, informed by the past (in its "iterational" or habitual aspect), but also oriented toward the future (as a "projective" capacity to imagine alternative possibilities) and toward the present (as a "practical-evaluative" capacity to contextualize past habits and future projects within the contingencies of the moment). Emirbayer and Mische (1998) further make a proposition that actors in the midst of changing situations and contexts that demand the reconstruction of temporal perspectives can expand their capacity for imaginative and or deliberative response. This leads to viewing actors as much more than agents alone in a social situation, in that any agency that they employ in the maintenance of their social structures is both within themselves and simultaneously embedded in their social structures.

This definition of agency provides for the scope that Isocrates (436-338 B.C.E.) a contemporary of Plato in ancient Greece, advocated (Bizzell and Herzberg, 2001) in his definition of rhetoric, involving the use of history to solve present problems depending on the context (following from the Sophistic traditions of Kairos) and making useful contributions for the future in all public and private human affairs. This approach, also taken by Emirbayer and Mische (1998), allows one to view human agency not only from the temporal perspective, but also from the relational dimension of sociality by viewing the embeddedness of actors in multiple cultural, social-structural and social-psychological contexts.

One can approach the task of identifying the settings and situations that tend to keep actors engaged in patterns of behaviors and communication processes from various theoretical perspectives, such as action theory and normative theory which are discussed by Emirbayer and Mische (1998), but the aim of this paper is to propose a theoretical model based on Markov processes to model this human agency so prevalent and embedded in social networks. In socio-temporal relations where kinship relations are generally constant over time and friendship or entrepreneurial networks change with time, the structure of a group is often a function of time. We can then model these relations stochastically in order to better explain the social and network behavior of actors in the social structure.

If we consider the formation of relationships and ties over time between people in a social structure, we can consider this to be both a deterministic process (with kith and kin) and a stochastic process (friends, co-workers, entrepreneurial acquaintances, etc.). These socio-temporal relations depend on the past actions and behaviors of the people, which affects their present actions and their future relationships with one another. In a general random process we have a set of times $0=t_0 < t_1 < \ldots < t_n$ and a set of states $s_i \in S$ so the probabilities $P(Xt_n . s_n \mid Xt_{n-1} = s_{n-1}, \ldots , Xt_0 = s_0)$ depend on the entire history of events from t_0 to t_n. A stochastic process is a *Markov process* if the probability of the next state depends upon the current state and not the previous states. The current state is the sum total of all past states or in the case of social relations, all past experiences, and is used as a predictor of future states or behavior. We can thus see how the Markovian approach lends itself to the modeling of human agency in social networks and the rest of the paper will be devoted to exploring more of the relevant literature and the development of the model.

Organizations, Networks and Agency

Since we are looking specifically at the stochastic nature of these socio-temporal relations, i.e. between friends, co-workers and entrepreneurial acquaintances, I begin with a discussion of some social network analysis studies, which looked at organizational structures in the context of computer-mediated communication. This is relevant to the understanding and modelling of human agency. One of the questions we would like answered is the one about the contexts that constrain or enable the capacity of people in a social structure like an organization for communicative deliberation, by means of which they judge the particular actions most suitable for resolving the practical dilemmas of emergent situations (Emirbayer and Mische, 1998). Communication and communication patterns then could be viewed in this context of both emergent network structure and emergent situations, such as those that arise in organizations.

Burkhardt and Brass (1990) used social network analysis to examine the organizational impacts of a new information technology, specifically the relationship between *centrality, power, and the timing of adoption* of a new distributed computing system. They reasoned that a new technology would increase uncertainty, raising the power of those able to mitigate that uncertainty, while increasing the need to communicate about that uncertainty and thus altering the social communication network. They found that early adopters increased their power and centrality to a greater degree than later adopters. They also observed changes in the network structure as a result of the new technology.

Rice and Aydin (1991) used social network analysis to examine the mechanism by which individual attitudes toward an information system were influenced by the attitudes of socially proximate others. They identified three mechanisms for proximity: *relational, positional and spatial*. Relational and positional proximity were defined following the discussion above, while spatial proximity represented physical location. They found that attitudes towards an information system are socially influenced and that relational and positional proximity have greater influences than traditional occupational roles and spatial proximity.

By mapping the social networks and the computer-mediated communication (CMC) networks for each organization, Zack and McKenney (1995) examined how existing social structure influences the way in which an organization appropriates electronic messaging systems. They were able to make direct comparisons between networks and between organizations. By comparing both networks within organizations, they found that the CMC network closely reflected the social structure. Comparing networks across organizations, they found that where the social structure reflected open, collaborative communication and a participatory management style, CMC was used to broaden the communication networks and make them more responsive. Where the social structure reflected relationships in conflict and a strict, centralized hierarchy, CMC was appropriated in a way that reinforced the hierarchy. They were able to relate performance effectiveness to particular communication patterns to show why the technology enabled effective performance in some organizations while not in others.

Zack (2000) studied the key impact of organizational systems and new information technologies and how they enable new organizational forms - the structural features or patterns of relationships and information flows of an organization. His study also proposed social network analysis as a highly appropriate and useful method for framing and describing the effects of organizational and communication systems on organizational forms and structures. The key finding was that the social structure influenced the way the technology was appropriated, and therefore, mediated its impact on organizational performance.

Wellman, Garton, and Haythornwaite (1997) showed the utility of the social network approach for studying CMC, in either a computer-supported network, in a virtual community, or in less bounded systems like the Internet. However, Ahuja & Carley (1998) empirically measured the structure of a virtual organization and found evidence of hierarchy in the virtual organization, much like in traditional organizations. They recommend the retaining of individuals who are at the center of information exchange networks, designing reward structures so those individuals acting as knowledge centers on specific topics can be retained and promoted, and giving incentives for these individuals to share their expertise with other organization members. Ahuja & Carley go further to say that it is critical to develop and train other individuals who can assume the network positions occupied by other individuals as they are promoted so that the communication structures can remain stable despite the turnover.

The above studies have highlighted the various facets of human agency in organizations, specifically with regard to information flow and the resultant emergent network structures. They bring to the fore both relational and positional issues involving actors in an organization and how these impact their social structures. Emirbayer and Mische (1998) ask whether the changes in agentic orientations allow actors to exercise different forms of mediation (face-to-face and CMC to leverage relational social capital, Lin, 1999) over their contexts of action, and if actors who feel blocked in encountering problematic situations can actually be pioneers in exploring and reconstructing contexts of action.

To attempt an answer to these questions I have proposed to look at these socio-temporal structures from a stochastic modeling perspective. Socio-temporal relations like kinship relations, are generally constant over time, while friendship or entrepreneurial networks change with time, often making the structure of a group a function of time. If we consider the formation of relationships and ties over time between people in a social structure, we can consider this to be both a deterministic process (with kith and kin) and a stochastic process (friends, co-workers, entrepreneurial acquaintances etc.). These socio-temporal relations depend on the past actions and behaviors of the people, which affects their present actions and their future relationships with one another.

Researchers have proposed several such time dependent statistical models which are both deterministic and stochastic (Bernard and Killworth, 1979; Rapoport, 1963; Wasserman and Faust, 1994). Wasserman and Iacobucci (1988, 1991) proposed loglinear approaches to model network changes and Markov chains to represent stochastic block models for structural equivalence and brokerage (Burt, 1992), while Katz and Proctor (1959) applied discrete Markov chain theory to longitudinal sociometric data to demonstrate the possible use of Markov models to explain the dynamic nature of social network structures. Sorensen and Hallinan (1977) applied continuous time discrete state Markov chains to the study of the evolution of triads over time and their model analyzes the tendency toward transitivity at the triad level affecting the social structure at the macro level.

A network triad consists of an unordered set of three nodes and the ties between them. An ordered set of three nodes is called a triplet and a triplet (i,j,k) is defined as transitive from i's perspective if the presence of arcs (directional lines or degrees) from i to j and from j to k implies the presence of an arc from i to k. Sorensen and Hallinan (1977) reported that triads tend to move away from intransitivity over time with the inconclusive/inconsistent assumption that triads behave independently. Snijders (1996) proposed stochastic actor-oriented models for network evolution which combined a rational choice approach with a continuous-time Markovian approach, while Holland and Leinhardt (1977a, 1977b) also proposed a continuous-time Markov approach to model structural change, starting from the dyad level with each dyad following a four state Markov process.

More recently, Robins and Pattison (2003) have proposed a generalized graphical modelling approach of *p** (Wasserman and Faust, 1994) social influence models to develop discrete time models for the temporal evolution of social networks. They report that systematic temporal processes are construed as effects that are homogeneous across the network, and that reflect dynamics inherent in a particular social relation. Any one actor cannot control these dynamics, especially given that non-dyadic configurations may be implicated, for instance, tendencies for various triadic configurations to be constructed or collapsed over time. Robins and Pattison (2003) further report that non-systematic processes, may pertain to the changing nature of a particular dyadic tie, or to change involving a particular socio-temporal neighbourhood of the network. Non-systematic processes are inhomogeneous across time and across the network, and are modelled as random. To separate non-systematic from systematic temporal processes Robins and Pattison (2003) use the constant tie assumption – whereby ephemeral ties are assumed not to have influence across time. They illustrate these models with an analysis of the Freeman EIES data, and then with data from a newly-formed small training group involving trust and friendship networks.

Random or Stochastic Processes

A stochastic process is a family of random variables $\{X_t \mid t \in T\}$ where T is a parameter space indexing the set. We can consider $\{X_t \mid t \in T\}$ to be the path of a particle moving randomly. The particles' position at a time t is X_t. For example, Brownian motion can be analyzed as such a random process as can data packets moving in a computer network or information flow in a social network. T can belong to $[0,\infty]$ or T can be $\{0,1, \dots \}$. Therefore, the indexing can be done for all real numbers, and in such a case we have a continuous process; alternatively, T can be the set of non-negative integers, and we have a discrete process.

We can characterize a joint cumulative distribution function (CDF) as $F\mathbf{x}(\mathbf{s};\mathbf{t})$ for a given set of random variables $\{X_{t1}, X_{t2}, \dots , X_{tn}\}$ as follows. Given the parameter vector $\mathbf{t}=(t_1,\dots,t_n)$, t_i non negative, real or integer and $t_i < t_{i+1}$ with state vector $\mathbf{s}=(s_1,\dots,s_n)$ then $Fx(s;t) = P(x_{t1} < s_1, x_{t2} < s_2,\dots x_{tn} < s_n)$ and the joint density function is given by $fx((s;) = (\partial^n)/(\partial_s \dots \partial_{sn}) * Fx(s;t)$. Stochastic modeling allows us to do the following:

- Find $P(x_t \in s')$ where s' is a particular state or set of states, and t is a particular time. We usually let t

$\rightarrow \infty$ (t tend to infinity), which is steady state. So we wish to find the probability of being in a particular state at a particular time.

- For t_i, $t_j \in T$ the relationship between x_{ti} and x_{tj}. That is, determine the relationship between the values of random variables at two times.
- Find $P(x_t = s_i)$ where s_i is a particular state. That is, the probability that the system enters a particular state.
- Since a system may enter into a state many times, one frequently wishes to know the first time of entry into a state.

Often, we have increments that occurs at times t_i and t_{i+1}. Within these increments we can have independent increments. For $t_i < t_{i+1} \in T$ then each $(x_{ti+1} - x_{ti})$ for $i < i+1$, is independent. This says that an event, such as an arrival, occurring at increments $[t_i,t_{i+1})$ was not influenced or affected by events in (t_{i-1},t). We can consider this event to be the arrival of a data packet at a network switch or the establishing of a contact between two actors in a social network by way of an email or a visit.

Markov Processes and Markov Chains

As mentioned earlier, in a general random process we have a set of times $0=t_0 < t_1 < \ldots < t_n$ and a set of states $s_i \in S$ so the probabilities $P(X_{tn} . s_n \mid X_{tn-1} = s_{n-1}, \ldots , X_{t0} = s_0)$ depend on the entire history of events from t_0 to t_n. A stochastic process is a *Markov process* if the probability of the next state depends upon the current state and not the previous states.

Several of the most powerful analytic techniques for evaluation of computer system performance (and many other systems) are based on the theory of *Markov chains*. A Markov chain is a special case of a *Markov process*, which is itself a special case of a *random process*. Random (stochastic) process as discussed above is a family of (ordered set of related) random variables *X*(t) where *t* is an indexing parameter (usually time). There are many kinds of random processes. Two of the most important distinguishing characteristics of a random process are whether or not the values that the random process can take on are continuous over some interval(s) and whether or not the indexing parameter is continuous or discrete.

Markov chains

A *Markov chain* is a discrete-state random process in which the only state that influences the next state is the current state. To be more precise:
X_{n+1} depends only on X_n and not on any X_i , $1 \leq i < n$
$\Pr [X_{n+1} = s_i \mid X_n = s_j, X_{n-1} = s_k, \ldots X_1 = s_1] = \Pr [X_{n+1} = s_i \mid X_n = s_j]$. This equation is referred to as the *Markov property*.

<u>Continuous-time Markov chain:</u>

Consider a continuous-time random process in which the number of times the random variables *X*(t) change value (the process changes state) is finite or countable. Let $t_1, t_2, t_{3,\ldots k}$ be the times at which the process changes state. If we ignore how long the random process remains in a given state, we can view the sequence $\{X_{t1}, X_{t2}, X_{t3,\ldots,K}\}$ as a discrete-time process embedded in the continuous-time process. Thus a continuous-time Markov chain is a continuous-time, discrete-state random process such that the embedded discrete-time process is a discrete-time Markov chain, and the time between state changes is a random variable with a memory-less distribution.

A distribution function $FT(.)$ is memory-less if and only if $FT(t) = FT(t + \tau | T > \tau)$. This says that the distribution of the time until the next state change is not a function of the time since the last state change. This can be restated as $FT(t) = \Pr[T \leq t + \tau | T > \tau]$. Using the definition of conditional probability,

$$\boldsymbol{FT(t) = \Pr[T \leq t + \tau \ \& \ T > \tau] / \Pr[T > \tau]}$$

$$= \frac{FT(t+\tau) - FT(\tau)}{1 - FT(\tau)} \quad (1)$$

Dividing both sides by t and taking the limit as $t \rightarrow 0$, we get a linear first order differential equation with the solution $FT(t) = 1 - e^{-FT(0)t}$. Hence, the only continuous-time, memory-less distribution is the exponential distribution, and the time between state changes in a continuous-time Markov chain is exponentially distributed. For discrete-time Markov chains, the next state may be the same as the current state: $X_{n+1} = X_n$. If p is the probability that the current state is as described above, then the probability that X_{n+1} is different from X_n is (1-p). Also, the probability that X_{n+1} is the same as X_n and X_{n+2} is different from X_{n+1} is p(1-p). Therefore, the number of state transitions between state changes is geometrically distributed.

One special type of Markov chain is a *birth and death* process, in which the states take on all non-negative integer values on a (possibly infinite) range. In this case, we can just refer to s_i as i and define a birth and death process as: if $X_n = i$, then $X_{n+1} = i + 1$, i, or $i - 1$, i.e., state transitions are always between neighboring states. If the inter-arrival times of data packets or agentic contact between two actors in a social network are independent and identically distributed (IID) and also exponentially distributed as shown above then the # of arrivals, n, over a given interval $(t, t+x)$ has a Poisson distribution with a mean rate of arrival, $\boldsymbol{\lambda}$, and a service rate of $\boldsymbol{\mu}$. The properties of the Poisson distribution allow the modeling of information flow in a communication channel like a telephone or computer network or, in our case, the social network of information flows in an organization or society in general.

Markov Chains and Agency

Suppose there are N communication terminals. In our case, N actors in a social network require a connection when the terminal (actor) becomes active (ready to make contact). Suppose there are C connections available. Normally, there will be fewer connections than actors, because not all terminals (actors) are active at the same time. It is also possible that all connections are temporarily busy and a terminal is blocked (the actor is inaccessible or has not yet made the social connection or the tie is weak). Then, the blocked terminal (actor) goes back to idle state without reattempts (will retry at a different time) or the blocked terminal (actor) is put on hold until the connection becomes available in which case the actor may need to establish a stronger tie with the alter. We can model the problem with birth-death Markov process.

- If state i represents the number of active terminals (i.e., the number of connections used),
- When i terminals are active, the rate at which these terminals become idle is given by i_m. The other $N-i$ idle terminals may become active with a total rate of $(N-i)\ \boldsymbol{\lambda}$. We can depict this Markov process as shown in Figure 1.

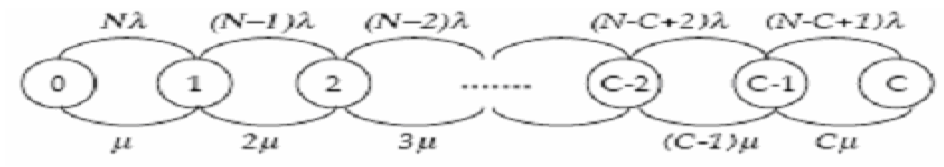


Figure 1: Markov Chain for Engset Blocking Formula – Human Agency

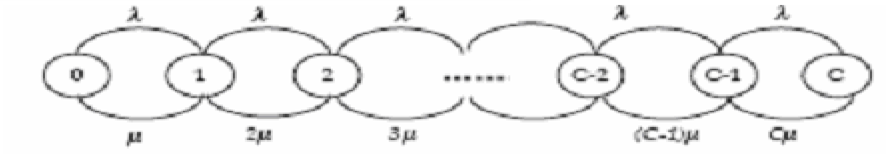


Figure 2: Markov Chain for Erlang Blocking Formula – Human Agency – Utilization and Efficiency

To simplify the analysis and observe that only adjacent states are connected, then the probability flow between states i and i-1 must be balanced. Using the fact that probability sums up to 1, we get **Probability of blocking = P(other N-1 terminals are using C connections).** This means that the probability of the actor's agentic attempt to establish the connection with the 'alter' in his/her social network is given by the above equation. In telephony this is called the Engset Blocking formula.

If we do not restrict the population size and let N go to infinity,[1] while keeping the call arrival rate[2] at a constant **λ** and taking limits for the binomial terms of the Engset distribution, the blocking formula becomes

$$\mathbf{B(C,\rho)} = \frac{[\rho^{C}/C!]}{\sum_{k=0}^{C} [\rho^{k}/k!]} \quad \mathbf{(2)}$$

Equation 2 is called the Erlang blocking formula, depicted in Figure 2 and in telephony provides information about trunk utilization and efficiency, while in the case of the social network, we can get information about the actor's agentic capacity and the ability to maintain his/her social network efficiently.

- Offered load = $\boldsymbol{\rho = \lambda / \mu = \lambda\, E[x]}$, where **λ** is the mean arrival rate; **μ** is the service rate over a given interval of (t, t+x).
- Blocking probability, $\boldsymbol{P_b = B(C, \rho)}$
- The blocked load is $\boldsymbol{\rho P_b}$ & the carried load is $\boldsymbol{\rho(1 - P_b)}$
- Efficiency = trunk utilization
- Average number of trunks (connections/ties) in use divided by total number of trunks (connections/ties) gives
- Utilization = $\boldsymbol{\rho(1 - P_b)/C}$ **(3)**
- Where C = total number of trunks (connections/ties).

This means that given a particular load (number of actors in the network), the ego would need a specific number of trunks

[1] N going to infinity means increasing the number of contacts for the actor in the social network.
[2] The call arrival rate is the actor's attempt to either build newer and newer entrepreneurial contacts or maintain the existing contacts.

(connections), C, to meet the target blocking probability (e.g., 1%). From these, you can compute the trunk utilization or efficiency. Based on past experiences, the actor can take decisions on current states, and this is the iterational process as described by Emirbayer and Mische (1998). The actor can then go further and predict his/her future human agentic capability and capacity to make newer friend or entrepreneurial contacts.

The following is an example of this process. Let there be 5 (the offered load **ρ**) actors in a network. Then, total number of ties or connections in the network is C = n(n-1)/2 (Scott, 1991). C then equals 5(5-1)/2 = 10. Assuming a target probability of blocking ($\mathbf{P_b}$) to be 1% (0.01), we can now look at the trunk or connection utilization of the central actor from equation 3 above as

$$\mathbf{U = (5*(1-0.01))/10 = 0.495}$$

This means that with five actors in the network, the efficiency of utilization of the connections or ties is 49.5%. If you now increase the number of actors in the network to 10, the total number of connections rises to 45 and the connection utilization efficiency drops to 22%, keeping the target probability of blocking constant at 1%. Lowering the probability of blocking will only lower the efficiency, so the central actor will then have to determine his ideal or optimal connection utilization and based on this percentage, limit or manage his/her contacts. Thus, for an actor in an organizational or entrepreneurial network, it becomes imperative to know which contacts to retain, which to let go, and which to keep in abeyance by delegating or occasional communiqués so that she/he can best benefit by the efficient utilization of the network's resources. Since everything depends on the flow of information between actors and the social relations that they have developed, each actor will be in a position to allocate a priority to the contact based on the history of their previous interactions with that actor, their present relationship and determine their individual optimum for connection utilization in the future with that actor in the network.

Example From a Recent Study

In a recent study involving distance students using a computer-supported collaborative learning (CSCL) environment, a network of the usage of the CSCL's instant messenger (IM) system by the students revealed a few central actors, one of them being the instructor himself. Of the others who also participated in IM discussions, N = 11, it was found that they were also active in the electronic bulletin boards, regularly posting messages (task related, administrative and social). Figure 3 depicts the IM social network.

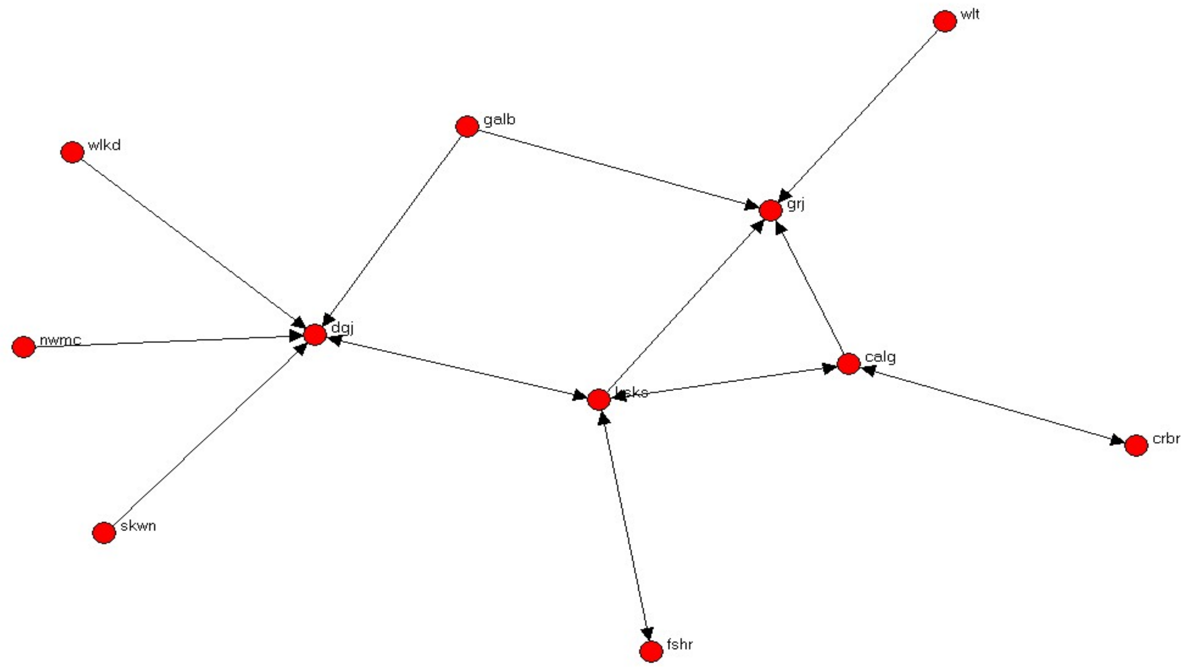


Figure 3: Social Network of Students in a CSCL IM Discussion (Ucinet 6.0 - Borgatti, Everett, and Freeman, 2002)

As you can see from Figure 3, actor dgj is the instructor and has the highest degree of connections, while actor grj is second, and actor ksks, third, with higher in-degree than actor calg. Actor ksks however, had highest flow betweenness centrality score (32.5) compared to dgj (20) and calg (11), while actor grj not being connected to dgj directly had a flow betweenness centrality score of (0). Surprisingly, in the discussion board posts, spanning eleven sessions over a six-week period, grj had the highest number of message posts (54 out of a total of 211) while actor calg had 12 posts and actor ksks had only six message posts.

This interesting phenomenon can be better explained when we use the modified Erlang Blocking formula described earlier (Utilization = **ρ (1- P_b)/C --** eq. 3). We can work with the number of actors in this network. The offered load **ρ** (actors in a network) is 11. The total number of ties or connections in the network is C = n(n-1)/2 (Scott, 1991). C then equals 11*(11-1)/2 = 55. Assume a target probability of blocking (**P_b**) to be 1 % i.e. 0.01, we can now look at the trunk or connection utilization of the central actor from equation 3 above as

$$\mathbf{U = (11*(1-0.01))/55 = 0.198}$$

Table 1 gives the network utilization for the whole network and for each of the actors calculated based on their in-degree and out-degree being the offered load **ρ** with the probability of blocking set at 1% (0.01). The offered load for each actor, when calculating the in-degree includes the ego and all alters to which the ego is connected. This is done in the case of the out-degree count also. Table 2 gives the network utilization for the whole network and for each of the actors calculated based on their in-degree and out-degree being the offered load **ρ** with the probability of blocking set at 10 % (0.1).

Table 1: Actor's network utilization for in-degree and out-degree with target blocking probability set at 1 %

Actor	ρ - indegree # of actors	P_b - prob	C - # of ties	U - Utilization	Actor	ρ - outdegree # of actors	P_b - prob	C - # of ties	U - Utilization	# of posts in Discussion Boards
Whole	11	0.01	55	0.198	Whole	11	0.01	55	0.198	100
ksks	4	0.01	6	0.66	ksks	5	0.01	10	0.495	6
Dgj	6	0.01	15	0.396	dgj	2	0.01	1	1.98	0
Calg	3	0.01	3	0.99	calg	4	0.01	6	0.66	12
Grj	5	0.01	10	0.495	grj	0	0.01	0	no value	54
Fshr	2	0.01	1	1.98	fshr	2	0.01	1	1.98	3
Crbr	2	0.01	1	1.98	crbr	2	0.01	1	1.98	3
Galb	3	0.01	3	0.99	galb	3	0.01	3	0.99	6
nwmc	2	0.01	1	1.98	nwmc	2	0.01	1	1.98	7
skwn	2	0.01	1	1.98	skwn	2	0.01	1	1.98	6
wlkd	2	0.01	1	1.98	wlkd	2	0.01	1	1.98	1
Wlt	2	0.01	1	1.98	wlt	2	0.01	1	1.98	2

Table 2: Actor's network utilization for in-degree and out-degree with target blocking probability set at 10 %

Actor	ρ - indegree # of actors	P_b - prob	C - # of ties	U - Utilization	Actor	ρ - outdegree # of actors	P_b - prob	C - # of ties	U - Utilization	# of posts in Discussion Boards
Whole	16	0.1	120	0.12	Whole	17	0.1	136	0.1125	100
ksks	4	0.1	6	0.6	ksks	5	0.1	10	0.45	6
Dgj	6	0.1	15	0.36	dgj	2	0.1	1	1.8	0
Calg	3	0.1	3	0.9	calg	4	0.1	6	0.6	12
grj	5	0.1	10	0.45	grj	0	0.1	0	no value	54
fshr	2	0.1	1	1.8	fshr	2	0.1	1	1.8	3
crbr	2	0.1	1	1.8	crbr	2	0.1	1	1.8	3
galb	3	0.1	3	0.9	galb	3	0.1	3	0.9	6
nwmc	2	0.1	1	1.8	nwmc	2	0.1	1	1.8	7
skwn	2	0.1	1	1.8	skwn	2	0.1	1	1.8	6
wlkd	2	0.1	1	1.8	wlkd	2	0.1	1	1.8	1
wlt	2	0.1	1	1.8	wlt	2	0.1	1	1.8	2

From the two tables we can see that the actors who had higher centrality from the network depicted in Figure 3 have differing values of network utilization **U**. While actor grj, who is central in that she had a higher in-degree than other students (with the exception of dgj the instructor), her out-degree is zero. Actor ksks, with in-degree = 4 and out-degree = 5, has **U** = 0.6, $\mathbf{P_b}$ = 10% for in-degree (0.66 with the target probability of blocking set at 1 %), and **U** = 0.45, $\mathbf{P_b}$ = 10% for out-degree (0.495 with target probability of blocking set at 1%) respectively. Coupled with his higher flow betweenness centrality score (32.5, the highest) and his judicious use of discussion board message postings (only 6), actor ksks is better placed in the network to not only utilize it in an optimum manner, but also leverage his network position in a manner that does not greatly reduce his efficiency. He was able to manage the capacity of actors and information in the network, and thus, used it to his benefit (found from self-reported survey questions, where he found the CSCL system Elluminate™, with its IM, videostreaming, voice-in and whiteboard features suitable for content delivery and that he found it facilitated his learning process). He also reported that it helped him learn new conceptual knowledge and gain new insights into collaborative distance work; he was completely satisfied with his performance in the course, felt he had the respect of his distance classmates, and was confident of getting an A in the course.

Though the other actors in the network have a greater utilization value, they only have 2 actors in their degree calculations, and, as the number of connections 'C' is in the denominator of equation 3, fewer connections will definitely give higher utilization values, but this does not mean that they are utilizing the network well, as can be seen from both their positions in the network (Figure 3) and their activity in the discussion board message postings. Actor, grj though very active in the message postings, and occupying a good network position, because of her lack of out-degree, her own self-reported views on the CSCL system and that she considered herself to be the most knowledgeable person in the network, her network utilization values are lower and apparently not beneficial to her.

Discussion

This modeling of human agency provides one way to answer the questions raised by Emirbayer and Mische (1998), namely, what kinds of contexts provoke or facilitate actors toward gaining imaginative distance from those multi-cultural and socio-cultural responses and thereby reformulating past patterns through the projection of alternative future trajectories? And, what sorts of contexts constrain or enable their capacity for communicative deliberation, by means of which they judge which particular actions are most suitable for resolving the practical dilemmas of emergent situations?

Depending on the situation, actors may decide to use past values or information and change as the need arises by establishing newer communication patterns as they seek to imagine alternative futures for a problematic present. However, certain sets of actors might resist change and hold tightly to past routines (such as local or national traditions) in an attempt to ward off uncertainty. By looking at periods of stability and change, as do telephony engineers when monitoring peak and off-peak telephone traffic, much insight into such processes can be gained by looking at these agentic orientations. Thus, the socio-temporal dimension of actors engaged in emergent events sees them positioned between the old and the new, and forces them to develop new ways of integrating past and future perspectives by understanding and using their embedded human agency in multiple cultural, social-structural, and social-psychological contexts.

Emirbayer and Mische (1998) state the implication of Rose Laub Coser's (1975, p. 239) missive that actors who are located in more complex relational settings must correspondingly learn to take a wider variety of factors into account, to reflect upon alternative paths of action, and to communicate, to negotiate, and to compromise with people of diverse positions and perspectives. All of these qualities, she argues, support more autonomous personal and occupational identities (and, by extension, more imaginative and reflective engagements with the contexts of action).

In this paper, I began with a review of organizational communication practices and how social network analysis reveals the emergent network structure in organizational information flows. DiMaggio (1991) argues that the creation of a professional environment at the inter-organizational level leads to more critical discourse, formal equality, and purposeful search for alternatives. This is in contrast to the routine, hierarchy and scripted forms of rationality that predominate inside organizations, highlighting the variation in agentic capacity to institutional complexity. Other researchers have looked at how choice-making and careers are embedded in complex network interactions (Abbott and Hrycak, 1990; Pescosolido, 1992), and the model proposed in this paper may shed light on how differently structured networks and careers support variable agentic orientations. These agentic orientations take into account actors' roles and position, (i.e. in brokerage, central and/or boundary spanner roles) to leverage these socio-temporal relational contexts and develop greater capacities for creative and critical intervention).

We are aware that entrepreneurs and actors in organizations embed their business decisions in social structures (Borch, 1994; Hansen, 1995; Larson & Starr, 1993; Reynolds, 1991; Starr & MacMillan, 1990). Social networks are not fixed; they are the social context of businesses and can be activated according to different needs (Granovetter, 1985; Burt, 1992). To fit their organizational needs, entrepreneurs bring both those that are closer and distant to them into their business decisions. However, kinship relations are usually stable over time and can be modeled deterministically.

The focus of this paper has been to put forth a Markov model that will aid in providing information to entrepreneurs on the capacity of their social networks and how to maintain social-temporal relations with their contacts and resources, despite their tendency to connect to a "friend of a friend" or the ability to select new contacts (acquaintances) from their networks.

This model takes advantage of the similarities between the concept of human agency and Markov random processes. It takes into account the fact that present experiences are the sum total of past iterational and habitual experiences and the present practical-evaluative capacity to evaluate these past experiences. The model then adapts the Markov process when it uses the present practical-evaluative experience to create a projective capacity toward the future state of the social network, by providing the actor (ego) with information about the efficient utilization of his/her channel/network capacity, and the number of contacts or resources he/she would need to maintain to achieve the future stability of the network contacts and resources.

Conclusions

From Rapoport (1963), we view the creation of new network ties as a trade-off between two opposing forces: "order" and "randomness." "Order" is defined in terms of a *triadic closure bias* (Rapoport, 1963), i.e., the tendency of an individual to connect to a "friend of a friend." "Randomness," on the other hand, means that new acquaintances are to be selected by drawing uniformly from the population at large.

Further research in this vein would involve designing a study which would collect the present N of the ego (entrepreneur), the present and past number of contacts, and make a prediction about the capacity of the ego's network and the future capacity and network utilization in order to have a healthy and productive association with the contacts and resources. We can then bring in notions of other factors like "network trust" (Burt, 1992), the "order & randomness" (Rapoport, 1963) that play a role in the selection, and maintenance of network contacts.

A longitudinal study may also estimate transition probabilities and contingencies for transition to another phase or dropping out of the establishment process. Further, longitudinal network data on successful and unsuccessful entrepreneurs, and the conditions forcing entrepreneurs to drop out of the establishment process, would help shed more light on the efficacy of the model. We need more research to describe the development and composition of efficient social structures that are conducive to entrepreneurship and the integral role played in these networks by human agency.

References

Abbott and Hrycak 1990. "Measuring Resemblance in Sequence Data: An Optimal Matching Analysis of Musicians' Careers." *American Journal of Sociology*, 96: 144-185.

Manju K. Ahuja , Kathleen M. Carley, (1998). Network Structure in Virtual Organizations. Journal of Computer Mediated Communication , 3 (4) http://jcmc.indiana.edu/vol3/issue4/ahuja.html

Bernard, H. R. and Killworth., P. D. 1979. Informant accuracy in social network data II. *Human Communication Research*, 4: 3-18.

Bizzell, P. and Herzberg 2001. The Rhetorical Tradition – 2nd ed. Bedford/St. Martin's, 2001.

Borch, O.J. 1994. The process of relational contracting: Developing trust-based strategic alliances among small business enterprises. In P. Shrivastava, A. Huff & J. Dutton eds., Advances in Strategic Management, 10B: 113-135.

Borgatti, S.P., Everett, M.G. and Freeman, L.C. 2002. Ucinet for Windows: Software for Social Network Analysis. Harvard: Analytic Technologies.

Burkhardt, M. E. and Brass, D. J. 1990. "Changing Patterns or Patterns of Change: the Effects of a Change in Technology on social Network Structure and Power", *Administrative Science Quarterly*, 35 (1): 104.

Burt, R.S. 1992. *Structural Holes: The Social Structure of Competition* Cambridge, MA: Harvard University Press.

Coser, R.L. 1975. "The Complexity of Roles as a Seedbed of Individual Autonomy." Pp. 237–63 in *The Idea of Social Structure: Essays in Honor of Robert Merton,* edited by Lewis Coser. New York: Harcourt Brace Jovanovich.

DiMaggio, P.J. 1991. "Constructing an Organizational Field as a Professional Project: U.S. Art Museums, 1920–1940." Pp. 267–92 in *The New Institutionalism in Organizational Analysis,* edited by Walter W. Powell and Paul J. DiMaggio. Chicago: University of Chicago Press.

Emirbayer, M. and Mische, A. 1998. What Is Agency? *New School for Social Research,* 103 (4): 962–1023.

Granovetter, M. (1985). "Economic Action and Social Structure: The Problem of Embeddedness." American Journal of Sociology, 91(November): 481-510.

Hansen, E.L. 1995. Entrepreneurial network and new organization growth. *Entrepreneurship: Theory & Practice*, 194: 7-19.

Holland, P. and Leinhardt, S. 1977a. A dynamic model for social networks. *Journal of Mathematical Sociology*, 5: 5-20.

Holland, P. and Leinhardt, S. 1977b. Social structure as a network process. *Zeitschrift f ur Soziologie*, 6: 386-402.
Katz, L. and Proctor, C.H. 1959. The con_guration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika*, 24: 317.
Larson, A. and Starr, J.A. 1993. A network model of organization formation. *Entrepreneurship: Theory and Practice*, 172: 5-15.
Lin, N. 1999. Building a network theory of social capital. *Connections*, 221: 28-51.
Pescosolido, B. 1992. Beyond Rational Choice: The Social Dynamic of How People Seek Help. *American Journal of Sociology*, 97(4): 1096-1138.
Pug, G.Ch. 1990. Non-asymptotic con_dence bounds for stochastic approximation algorithms with constant step size. *Monatshefte f ur Mathematik*, 110: 297-314.
Rapoport, A. 1963. Mathematical Models of Social Interaction. Handbook of Mathematical Psychology. R. D. Luce, R. R. Bush and E. Galanter. New York, Wiley. 2: 493-579.
Reynolds, P.D. 1991. Sociology and entrepreneurship: Concepts and contributions. *Entrepreneurship: Theory & Practice*, 162: 47-70.
Rice, R.E. and Aydin, C. 1991. "Attitudes toward New Organizational Technology: Network Proximity as a Mechanism for Social Information Processing." *Administrative Science Quarterly*, 36: 219-244.
Robins, G. and Pattison, P. 2003. Random Graph Models For Temporal Processes. In Social Networks Department of Psychology, University Of Melbourne.
Scott, J. 1991. *Social Network Analysis: A Handbook*. London: Sage Publications.
Snijders, T.A.B. 1996. Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology,* 21: 149-172. Also published in Doreian and Stokman, 1997.
Sorensen, A.B. and Hallinan, M.T. 1977. "A Reconceptualization of School Effects." *Sociology of Education*, 50: 272-89.
Starr, J. and MacMillan, I.C. 1990. Resource cooptation via social contracting: Resource acquisition strategies for new ventures. *Strategic Management Journal*, 11 **(**Summer): 79-92.
Wasserman, S. and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.
Wasserman, S. and Iacobucci, D. 1988. Sequential social network data. *Psychometrika*, 53: 261-282.
Wasserman, S., & Iacobucci, D. (1991). Statistical modeling of one-mode and two-mode networks: Simultaneous analysis of graphs and bipartite graphs. British Journal of Mathematical and Statistical Psychology, 44, 13–44.
Wellman, B., Garton, L., and Haythornwaite, C. 1997. "*Studying Online Social Networks*." JCMC, 3 (1).
Zack, M.H. 2000, January. "Researching Organizational Systems using Social Network Analysis." Proceedings of the 33rd Hawaii International Conference on System Sciences: Maui, Hawaii.
Zack, M.H. and McKenney, J.L. 1995, "Social Context and Interaction In Ongoing Computer-Supported Management Groups." *Organization Science*, 6 (4): 394-422.

Scan Statistics for Interstate Alliance Graphs

David J. Marchette
Naval Surface Warfare Center, Dahlgren, Virginia

Carey E. Priebe
Johns Hopkins University, Baltimore, Maryland

Abstract
This paper discusses work on graphs defined in terms of alliances between countries. Scan statistics are used to investigate years in which there are an unusual number of agreements, not just between one country and its allies, but amongst the allies themselves. This is related to work on email ``chatter'' discussed in Priebe et al. (2005). The scan statistic detects unusually high (or low) values for a graph invariant within a local region of the graph (an induced subgraph). Thus, without a priori knowledge of where in the graph the detection might occur, we seek to detect a region of the graph that is very different from the other regions. We will use a particular graph invariant, the size, or number of edges in the graph, to help detect interesting changes in the alliance graphs that we investigate. We will be more precise below, but the idea is as follows: A detection at scale 0 corresponds to a single country making an unusually large number of alliances; a detection at scale 1 corresponds to a country and its allies making a large number of alliances among themselves. This can be a measure of the cohesiveness of the group; a detection at scale 2 (and higher) corresponds to a larger spreading of the alliances. It means that not only are there more alliances among the countries allied with the central country, but among their allies there are more alliances. This paper seeks to perform two tasks: the first is to introduce scan statistics to those in the social network community not familiar with this work; the second is to determine whether, in the case of interstate alliances, there are any interesting detections at scales above 0. We will demonstrate that sometimes this type of behavior is interesting.

Acknowledgements: This work was funded in part by the Office of Naval Research under the In-House Laboratory Independent Research program.

David J. Marchette is a Principal Scientist at the Naval Surface Warfare Center. His research interests are in computational statistics and pattern recognition, with a particular focus on kernel and mixture models, classification and clustering, and the analysis of high dimensional, complex data. He has applied these methodologies to text processing, graph theory, network analysis and computer security. He is the author of two books and over 50 journal articles.

Carey E. Priebe is a professor in the Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, holding joint appointments in the Department of Computer Science, the Center for Imaging Science, and the Whitaker Biomedical Engineering Institute. His research interests include computational statistics, kernel and mixture estimates, statistical pattern recognition, statistical image analysis, dimensionality reduction, model selection, and statistical inference for high-dimensional and graph data.

For correspondence, please contact David J. Marchette at: dmarchette@gmail.com.

Introduction

There is a rich literature on social network analysis for understanding international relationships (Smith and White, 1992, Barnett, 2001, Ward et al., 2003, Moaz et al., 2004, Ward et al., 2005, Moaz, 2004, and Mahutga, 2006, and the references therein.). The scan statistic technique considered in this paper is intended to supplement this work, and to become one more tool for use in the analysis of social networks. This paper is thus intended to define the scan statistic and illustrate its utility within the domain of international relations.

Scan statistics have a long history in signal and image processing.[1] A typical application is the detection of cancer in mammograms. Certain types of cancer present as clusters of micro-calcifications that show up in x-rays as small white dots. An unusually large number of white dots in a small area is an indication of breast cancer. The scan statistic approach is to slide a small window around the image counting the number of dots within the window. An unusually large value for this statistic is then used for detection. In this paper, we describe an extension of this idea to graphs, in particular to a time series of graphs. We are interested in finding regions of unusual activity in the graph, with the constraint that we have no a priori information about where such a detection might reside within the graph. In graphs, the analog for the window will be the neighborhood (or *k*-neighborhood), and the local statistic will be the size of the induced subgraph. Any invariant on the induced subgraph could be used in its place, but we will use size to illustrate the idea.

Our approach normalizes the scan statistic within a window on the time series of graphs. This provides a mean-centering and scaling that allows for the detection of anomalies (change points) using a single threshold. Essentially this is a residual calculation: we compare the current value with the predicted (expected value within the window) and flag a detection if the current statistic is above a threshold.[2] These typically put a probabilistic structure on the edges of the graph, while we are modeling a particular statistic, which is an extremum of a local graph invariant.

The paper is organized as follows: in the next section, we provide the basic terminology of graphs, and the definitions of the scan statistics; then, we describe the data used; this is followed by a discussion of the results; and finally, we discuss the conclusions and give suggestions for future research. In the results section, we will discuss the detections at each scale, illustrating the types of detections possible at the different scales. We will show that for the interstate alliance data there are interesting detections the can be found through the scan statistic approach that cannot be found through simply looking at the degrees of the vertices.

Graph Terminology and Notation

A graph G is a pair (V,E) where V is a set of vertices (also called nodes or actors) and E is a set of unordered pairs of elements of V (the edges). We call the order of the graph $n=|V|$ and the size of the graph $s=|E|$ (Bollobas, 2001). We will denote the edge

[1] See Glaz et al., 2001 for a detailed discussion of the mathematics behind the traditional methods.

[2] See Snijders (2005) for a nice discussion of other models of time-series applied to social networks.

from v to w as vw. For $v, w \in V$ the distance $d(v,w)$ is defined to be the minimum path length from v to w in E. The (closed) k^{th}-order neighborhood (or k-neighborhood) of a vertex v is the set of vertices of distance at most k from v:

$$N_k(v) = \{w \in V : d(v,w) \leq k\}$$

The degree of a vertex v is the number of edges incident on v. The subgraph induced by a set of vertices S, denoted $\Omega(S)$, is the graph with vertex set S and edge set $\{vw \in E : v, w \in S\}$, that is, the subgraph on the vertices S containing all edges between these vertices that exist in the original graph.

A random graph is a graph valued random variable. For the purposes of this paper, we will assume the vertices are fixed, and the random component is contained entirely in the edges. One of the simplest (and most common) types of random graphs is the Erdos-Renyi random graph. In this model, each edge has a probability p of being in the graph, independent of all the other edges. This model has been well studied (Bollobas, 2001). While this is the simplest and most studied random graph model, it is inadequate for modeling social networks. We will not make this type of independence assumption in our random graphs. Instead of dealing directly with the random graph model, we will use the scan statistics, defined below, to extract a time series of statistics from the time series of random graphs.

We will investigate a time series of graphs defined in terms of interstate alliances: each graph corresponds to the alliances in place within a calendar year; each vertex is a country and there is an edge between two vertices if there was an alliance between the countries during the current year.

Scan Statistics

Scan statistics are commonly used in the investigation of random fields (for example, a spatial point pattern or an image of pixel values) for the possible presence of a local signal (Glaz, 2001). These are sometimes referred to in the engineering literature as "moving window analysis"; the idea is to scan a small window over the data, X, and calculate a local statistic ("locality statistic") for each window. In point patterns, this locality statistic might be the number of events in the window; for image analysis, it might correspond to some statistic (e.g. the average, or the number of white dots) applied to the pixels in the window. The maximum of these locality statistics is known as the scan statistic which we denote $M(X)$. Under some specified homogeneity null hypothesis on X (a Poisson point process or a Gaussian random field), one specifies a critical value for which deviation above this value has probability α under the null hypothesis. If the maximum observed locality statistic is larger than, or equal to, this critical value, then the inference can be made that there exists a nonhomogeneity, a local region with a statistically significant signal.

An intuitive approach to testing these hypotheses involves the partitioning of X into disjoint subregions. This approach can have poor power characteristics when there is no prior knowledge of the location and geometry of potential nonhomogeneities. Essentially, one wishes to select the window location and geometry to maximize the statistic. In the absence of prior knowledge this cannot be accomplished via disjoint subregions, and thus scan statistics are recommended.

Scan Statistics on Graphs

For a non-negative integer k (the *scale*) and vertex v (the *location*), consider the closed k^{th}-order neighborhood of v in G, $N_k(v)$. We define the (scale k) *scan region* to be the induced subgraph of $N_k(v)$, denoted $\Omega(N_k(v))$ with vertices

$$V(\Omega(N_k(v))) = N_k(v)$$

and edges

$$E(\Omega(\mathrm{N_k}(\mathrm{v}))) = \{(v,w) \in E : v,w \in \mathrm{N_k}(\mathrm{v})\}$$

A *locality statistic* at location v and scale k is any specified graph invariant $\Psi_k(v)$ of the scan region $\Omega(N_k(v))$. In this work (as in the previous work reported in Priebe, 2005) we use the size invariant, $\Psi_k(v) = |E(\Omega(N_k(v)))|$, and for convenience define the scale 0 locality statistic to be the degree. In the case of a weighted graph, the invariant is the sum of the edge weights. Notice, however, that any graph invariant (e.g. density, domination number, etc.) may be employed as the locality statistic as dictated by application. The "scale-specific" *scan statistic*, M_k, is given by some function of the collection of locality statistics $\{\Psi_k(v)\}$ taken over all v in V. We will use the maximum locality statistic over all vertices,

$$M_k = \max \Psi_k(v)$$

where the max is taken over all v in V (Priebe, 2004; Priebe et al., 2005).

Under a null model for the random graph G (e.g. the Erdos-Renyi random graph model) the variation of $\Psi_k(v)$ can be characterized, and a large value of $\mathrm{M_k}$ indicates the existence of an induced subgraph (scan region) $\Omega(N_k(v))$ with excessive activity. A test can be constructed for a specific alternative of interest concerning the structure of the excessive activity anticipated. However, if the anticipated alternative is, more generally, some form of "chatter" in which one (small) subset of vertices communicates amongst themselves (in either a structured or an unstructured manner) then our scan statistic approach promises more power than other approaches.

Time is incorporated through the implementation of a sliding window with standardization of the $\Psi_k(v)$ (we indicate the temporal dependence by subscripting with time: $\Psi_{k,t}(v)$). First, we perform vertex standardization by subtracting a recent mean and dividing by a recent standard deviation. Let $\tau > 1$ be a given window width. Then

$$\widetilde{\Psi}_{k,t}(v) = \frac{\Psi_{\mathrm{k,t}}(\mathrm{v}) - \hat{\mu}_{k,t,\tau}(\mathrm{v})}{\max(\hat{\sigma}_{k,t,\tau}(\mathrm{v}),1)}$$

where

$$\hat{\mu}_{k,t,\tau}(\mathrm{v}) = \frac{1}{\tau}\sum_{s=t-\tau}^{t-1}\Psi_{\mathrm{k,s}}(\mathrm{v})$$

and

$$\hat{\sigma}^2{}_{k,t,\tau}(v) = \frac{1}{\tau-1}\sum_{s=t-\tau}^{t-1}(\Psi_{\mathrm{k,s}}(\mathrm{v}) - \hat{\mu}_{k,t,\tau}(\mathrm{v}))^2 .$$

We also standardize the scan statistic in a similar manner. Given $L>1$, the window width for the scan statistic is defined as

$$S_{k,t} = \frac{\widetilde{M}_{\mathrm{k,t}}(\mathrm{v}) - \widetilde{\mu}_{k,t,L}(\mathrm{v})}{\max(\widetilde{\sigma}_{k,t,L}(\mathrm{v}),1)}$$

where

$$\widetilde{M}_{k,j} = \max(\widetilde{\Psi}_{k,t}(v))$$

and $\widetilde{\mu}_{k,t,L}(\mathrm{v})$ and $\widetilde{\sigma}_{k,t,L}(\mathrm{v})$ are the running mean and standard deviation estimates of $\widetilde{M}_{k,t}$ based on the most recent L time steps, in a manner similar to those above. In both of these scalings, the denominators are constrained to be at least 1 in order to eliminate fragility due to small variations.

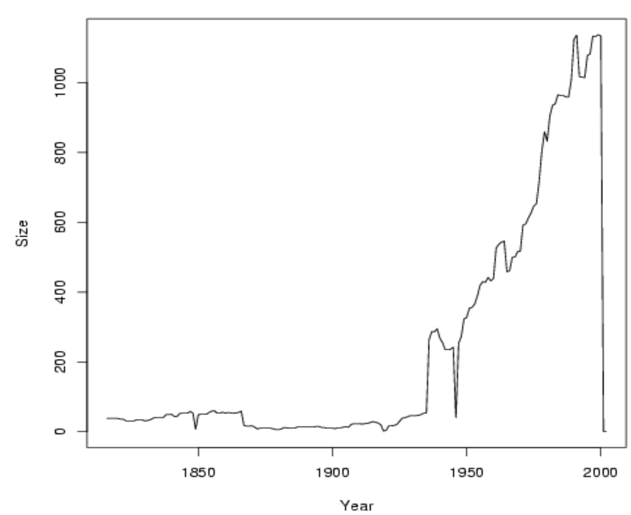
The Data

We consider a time series of graphs defined in terms of alliances. The alliance data represents alliances between a total of 214 nations collected from 1816-2000 (Gibler, 2004).[3] For each nation pair, alliance is coded as in Table 1. While the edges are colored by alliance type we will consider only the simplified graph with binary edges: existence or absence of an alliance. Utilizing the colored edges is straightforward: it can be done by modifying the locality statistic to produce a statistic (taking the type of edge into account) or by constructing multiple graphs, each for a given edge type, and analyzing these. As in any statistical inference problem, the key is to select the statistic that best captures the information of interest. For the purposes of illustration, we will use the size statistic on the binary graph.

Table 1: Alliance Codes in the Alliance Dataset

0	No Alliance	
1	Defense Pact	Intervene militarily if partner attacked
2	Neutrality	Remain militarily neutral if partner attacked
3	Non-aggression Pact	Consultation and/or cooperation in a crisis

For each year we form the graph with the nations as vertices, and the alliances between nations which define the edges. The alliance encoding is not obviously ordered. It is easy to argue that in certain scenarios a non-aggression pact is (or is not) stronger than a defense pact. Therefore, we will focus on the binary version of alliance/no alliance. Thus, there is an edge in the graph if there was an alliance of type 1, 2 or 3 between the two countries. Figure 1 depicts the sizes of the graphs.



[3] The data are available at http://correlatesofwar.org

Figure 1: The size of the graphs defined by the alliances.

As we can be see, the number of alliances increases dramatically after the mid 1930's (the big jump occurs in 1936). In Figure 2, we scale the size by the number of vertices in the graph (defined by first removing those countries which have no alliances with any other countries during that year). There are four major change points evident in these two graphs (particularly in Figure 2):

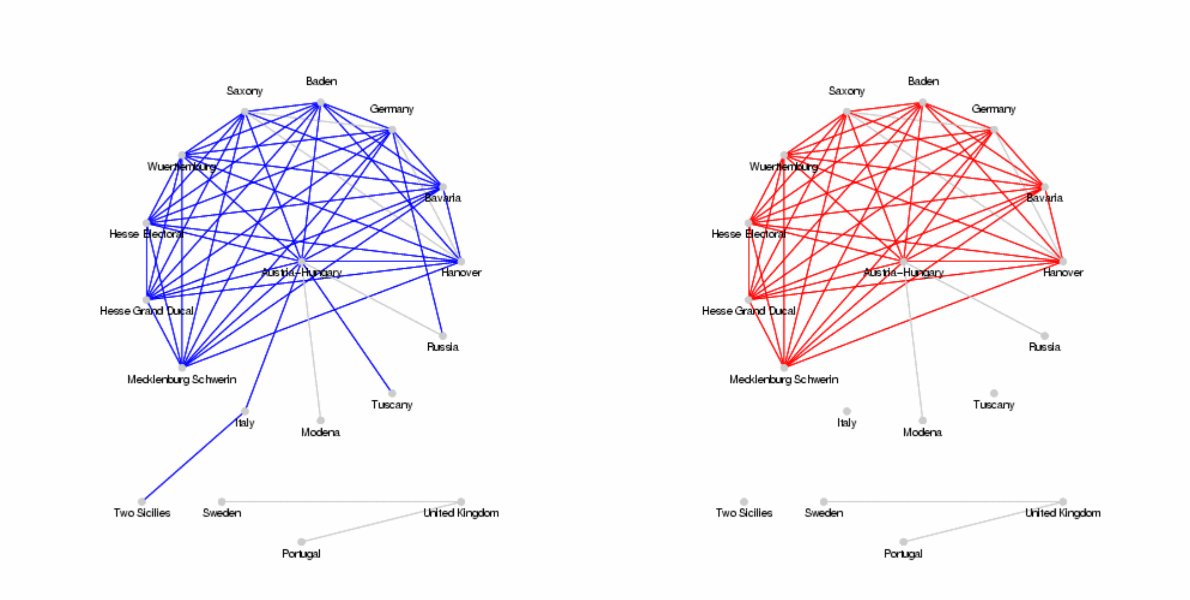
1. 1849 – a sudden dip in density.
2. 1867 – a drop in density.
3. 1936 – an increase in density.
4. 1946 – a sudden dip in density.

The graphs associated with the 1849, 1867 and 1946, are depicted in Figures 3, 4 and 5. The 1936 change point also shows up in the scan statistics, so we will deal with it later in the paper. The other obvious changes in the size distribution illustrate some features that are easily detectable from this global measurement on the graph.

Figure 2: The Density of the Graphs Defined by the Alliances

The density is computed on the subgraph formed after isolated vertices are removed. The arrows show two dips that are probably the result of missing data, and a drop in the density which is a result of the formation of the Austro-Hungarian Empire.

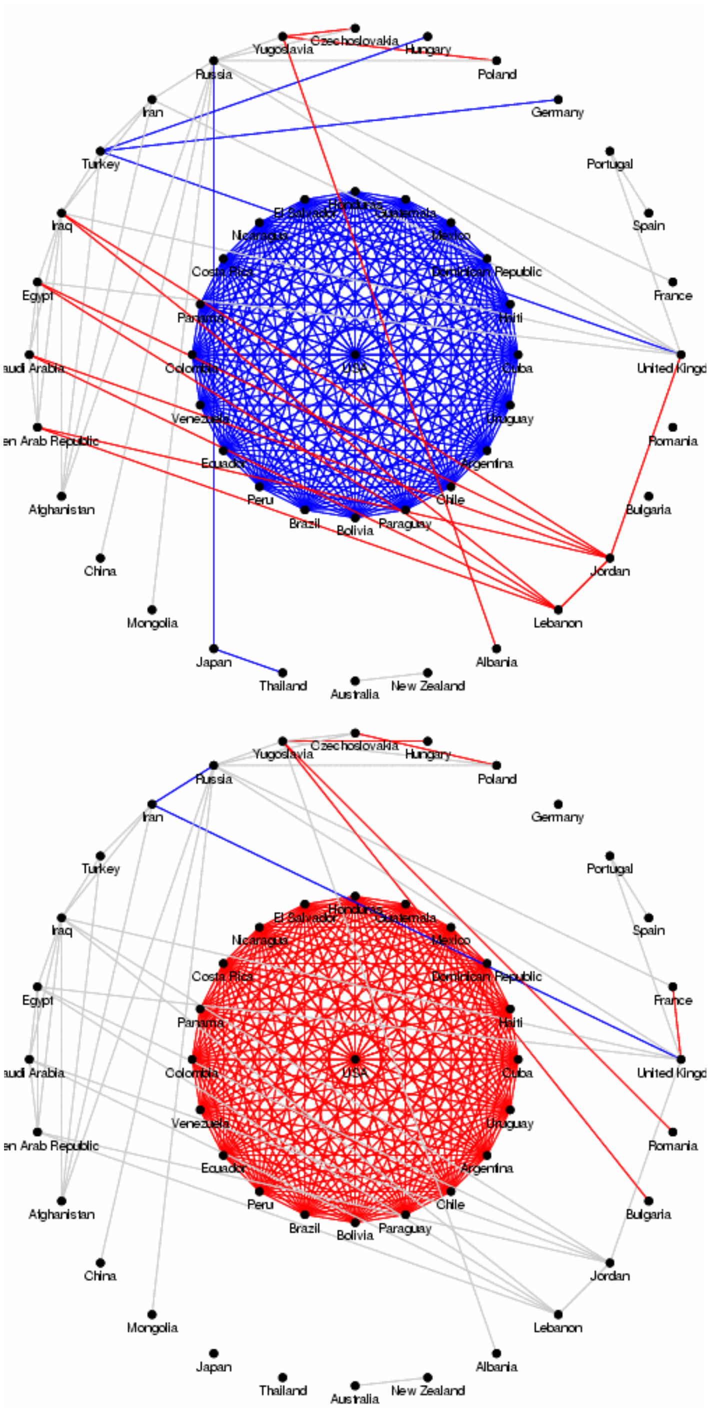
Figure 3: Graphs for the Years 1848 through 1850



The left graph depicts the changes in the alliances between 1848 and 1849, and the right the changes between 1849 and 1850. In both cases, blue edges denote edges that were removed from the first year to the second, red edges are edges that were added, and grey edges are those which stayed the same.

In Figure 3, we see the changes in alliance among the countries of Europe. These are the countries responsible for the change in the size of the graph in 1849. One hypothesis is that this is an error in the data: alliances that were in place are accidentally removed from the data in 1849. A similar effect is seen in the dip in size at 1946: the changes are displayed in Figure 4, and a reasonable hypothesis is that the data for the alliances between the United States and Central and South American countries were inadvertently dropped from the data. Alternatively, the alliances lapsed and were reinstated later. The scan statistic cannot distinguish between these two hypotheses.

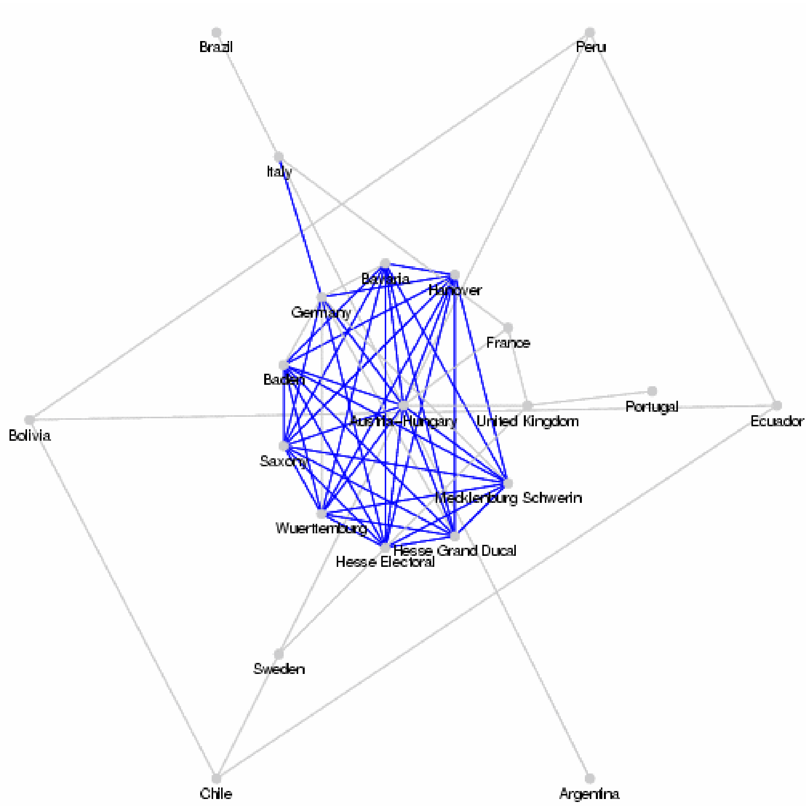
Figure 4: Graphs for the Years 1945 through 1947



The top graph depicts the changes in the alliances between 1945 and 1946, and the bottom, the changes between 1946 and 1947. In both cases, blue edges denote edges that were removed from the first year to the second, red edges are edges that were added, and grey edges are those which stayed the same. The central clique corresponds to the USA and countries in South and Central America, where the bulk of the change in 1946 occurred.

A more interesting change is the drop in size from 1866 to 1867. Figure 5 shows the changes in the alliances for this period. As can be seen, these are the result of the formation of the of Austria-Hungary empire, making the alliances with previous nation/states moot. This is a detection that could be made using a scan statistic approach, but only if one were to use the minimum instead of the maximum, or by considering the absolute value of the scaled locality statistic. This illustrates the importance of deciding a priori what types of changes one wishes to detect. In our experience, the crafting of a proper scan statistic is an iterative procedure that is used in conjunction with standard data analysis techniques.

Figure 5: Graphs for the Years 1866 and 1867



Blue edges denote edges that were removed from the first year to the second and grey edges are those which stayed the same.

In Priebe et al. (2005), an example of this is given in which a particular type of change (aliasing: a name change of an actor) was detected via the scan statistic, and the locality statistic was modified so that such a change would no longer be detected. We did not pursue these ideas on the alliance data.

Results

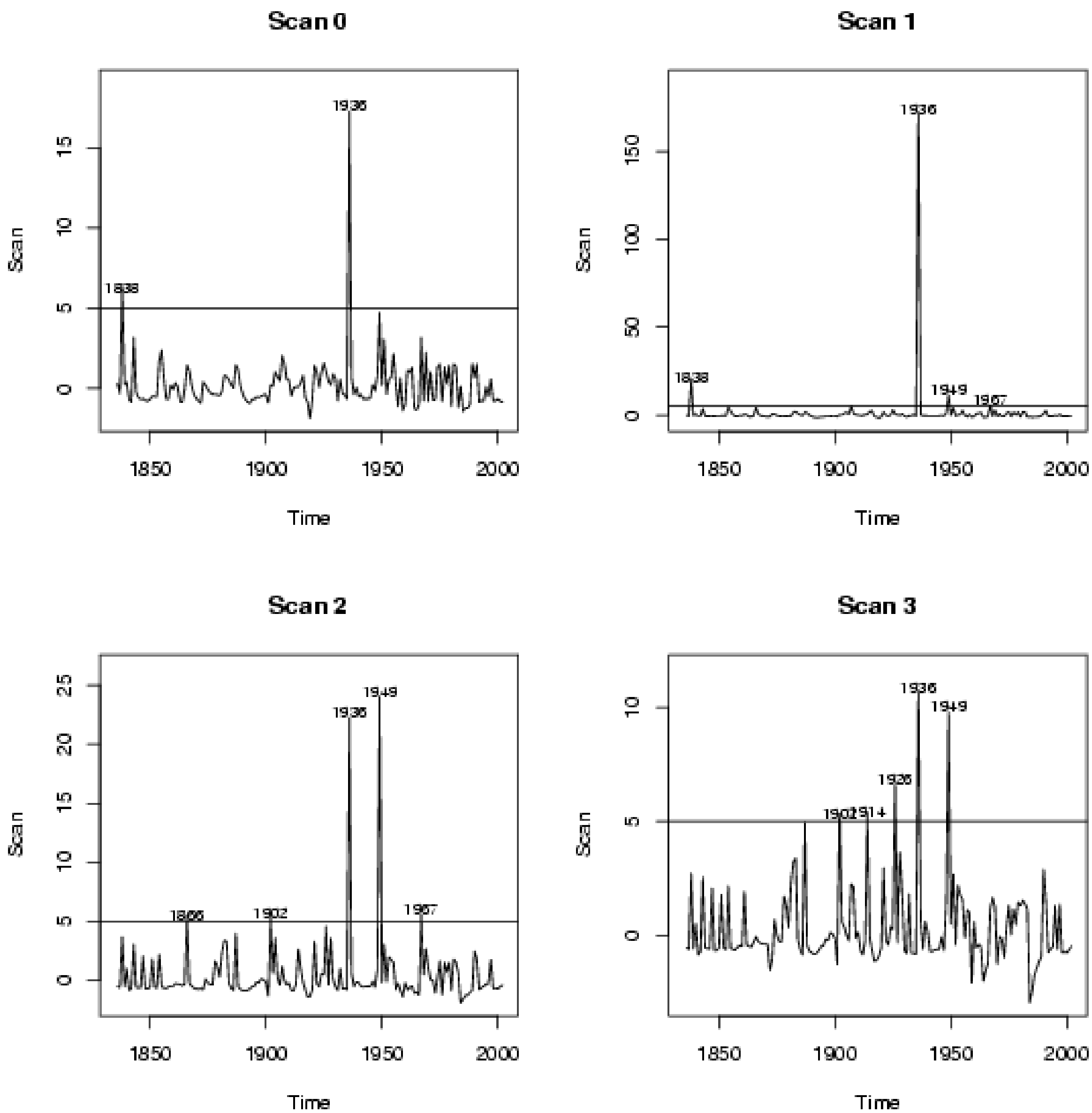
The following sections will deal with the scan statistics results in some detail. The above analysis demonstrates that there are some interesting discoveries that can be made by looking at global statistics of the time series of graphs (the size or density of the graphs).

Other graph invariants could no doubt result in other types of detections of interest. Clique number or measures of clustering, number of components, and measures of how connected the graph is (e.g. the minimum cut – how many edges are required to disconnect the graph) could all provide useful information about the structure of the graph in certain situations. One method for performing analysis on time series of graphs is to select an invariant of interest and produce a time series of this invariant. One can then fit a model to this invariant, and look for “significant” changes or other measures of interest, such as trends within the time series. This is very close to the scan statistics approach, except that instead of treating the graphs in the series individually, we use a window of the graphs to standardize the statistic of interest, as discussed above.

We now consider the results of applying the scan statistic methodology to detect unusual increases in the number of alliance among small sets of countries. In all cases, we use the window of a width of 10 years: $\tau=L=10$.

Figure 6 shows the detections (at a detection threshold of *5* standard deviations, indicated by the horizontal lines) for scan 0 (degree, $k=0$) and scans *1--3* ($k=1,2,3$) for the induced subgraph size locality statistic. The choice of *5* for the detection threshold is somewhat arbitrary. We have no reason to assume a particular model to the statistic (such as a normal distribution) however experience on several data sets has led us to the rule-of-thumb that the threshold should be set fairly high in the initial stages of analysis.

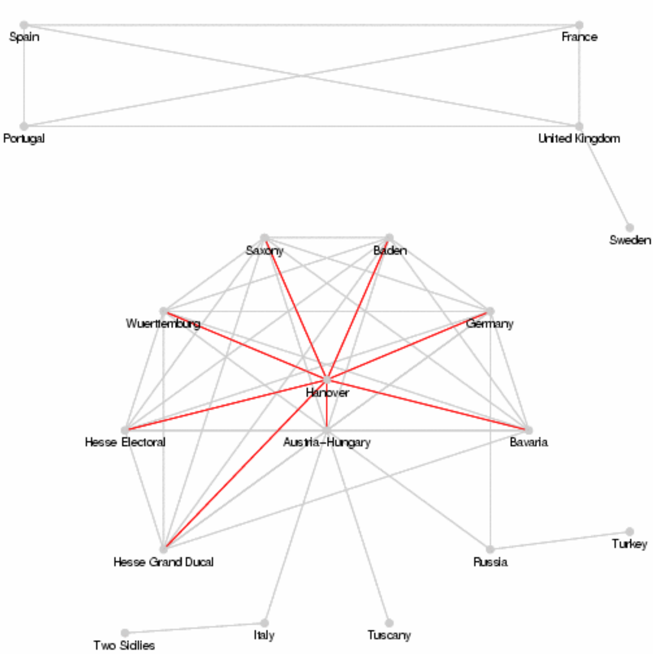
Figure 6: Scan Statistics for Scan *0* (degree) and Scans *1--3* (*k=0,1,2,3*)



The dates correspond to the detections above threshold and are indicated on the plots.

As can be seen in the figure, there are a number of detections at the different scales, and some detections at higher scales that are not detectable at lower scales. It is of interest to study these detections in particular, as they demonstrate the benefits of the multi-scale scan statistic approach.

Figure 7: Changes in the Graphs for the Years 1837 and 1838

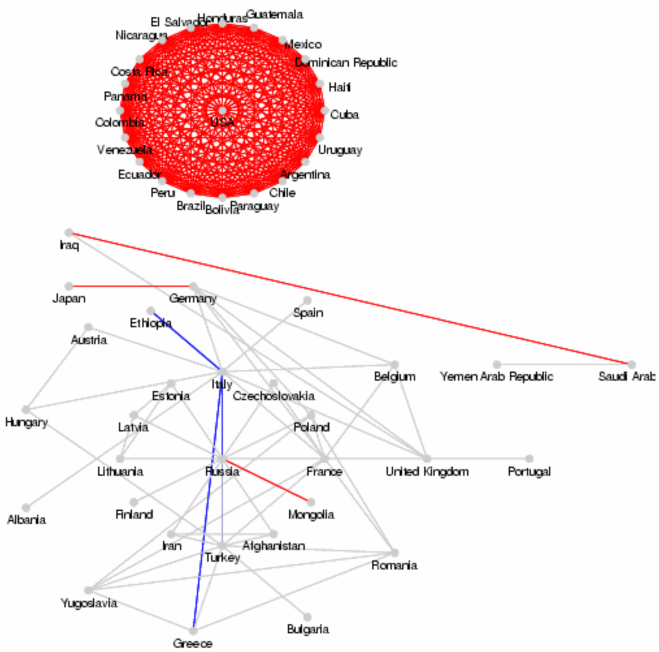


Shows the new alliances (red edges) in 1838, a result of the alliances formed between Hanover and other nations.

Figure 7 shows the first detection for *k=0*, degree.[1] This detects a new nation (or city-state), Hanover, forming alliances with eight other nations. This is an easy detection to make, based entirely on degrees, and is, in fact, the only change in the graph from 1837 to 1838.

[1] In all plots, unless otherwise noted, the entire graph, minus the isolated vertices, will be displayed.

Figure 8: Changes in the Graphs for the Years 1935 and 1936

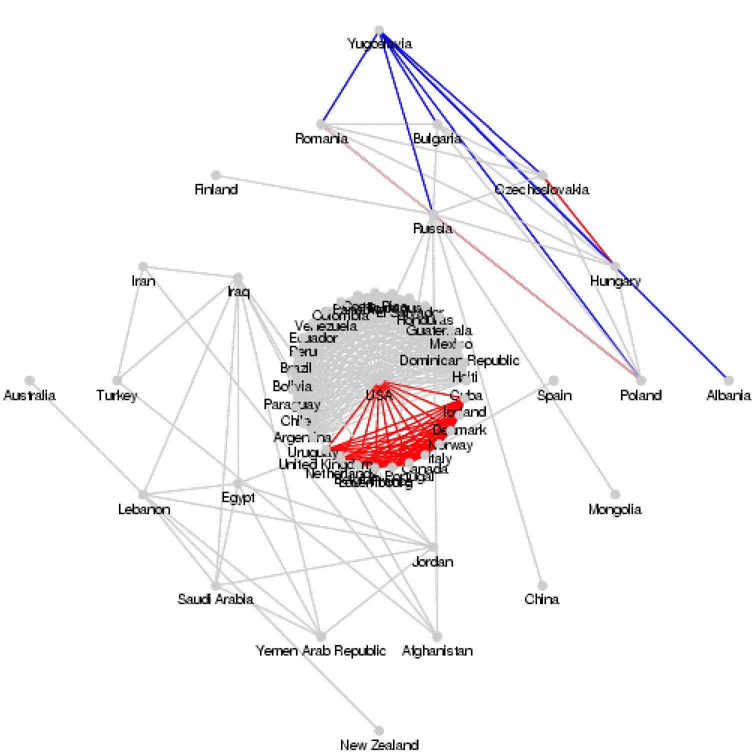


Shows the new alliances (red edges) and discarded alliances (blue edges) in 1936. The gray edges are those alliances that are in force for both years.

Figure 8 shows the detection at $k=0$ and the higher scan values in 1936 as a result of a set of alliances between the United States and the Central and South American countries. The red edges in the plot show the edges (alliances) that were put in place in 1936, and the blue edges show those that were in place in 1935 but no longer in place in 1936. The data do not support answering questions about these specific alliances; however, 1936 is the start of the Spanish Civil War, which may be the genesis of these alliances.

Figure 9 shows the $k=1$ detection for 1949. This is the result of the European partners of the United States forming alliances after the Second World War, most likely as a result of the North Atlantic Treaty, signed in April of 1949. This is the first of the detections we have seen which is a detection at $k=1$, but not $k=0$: it is not detected via the vertex degrees; rather, it is the small clique of alliances between these that produce the detection. Note further that this clique is smaller than the one represented by the US and Central and South American countries, so the detection could not easily be made via computing cliques.

Figure 9: **Changes in the Graphs for the Years 1948 and 1949**

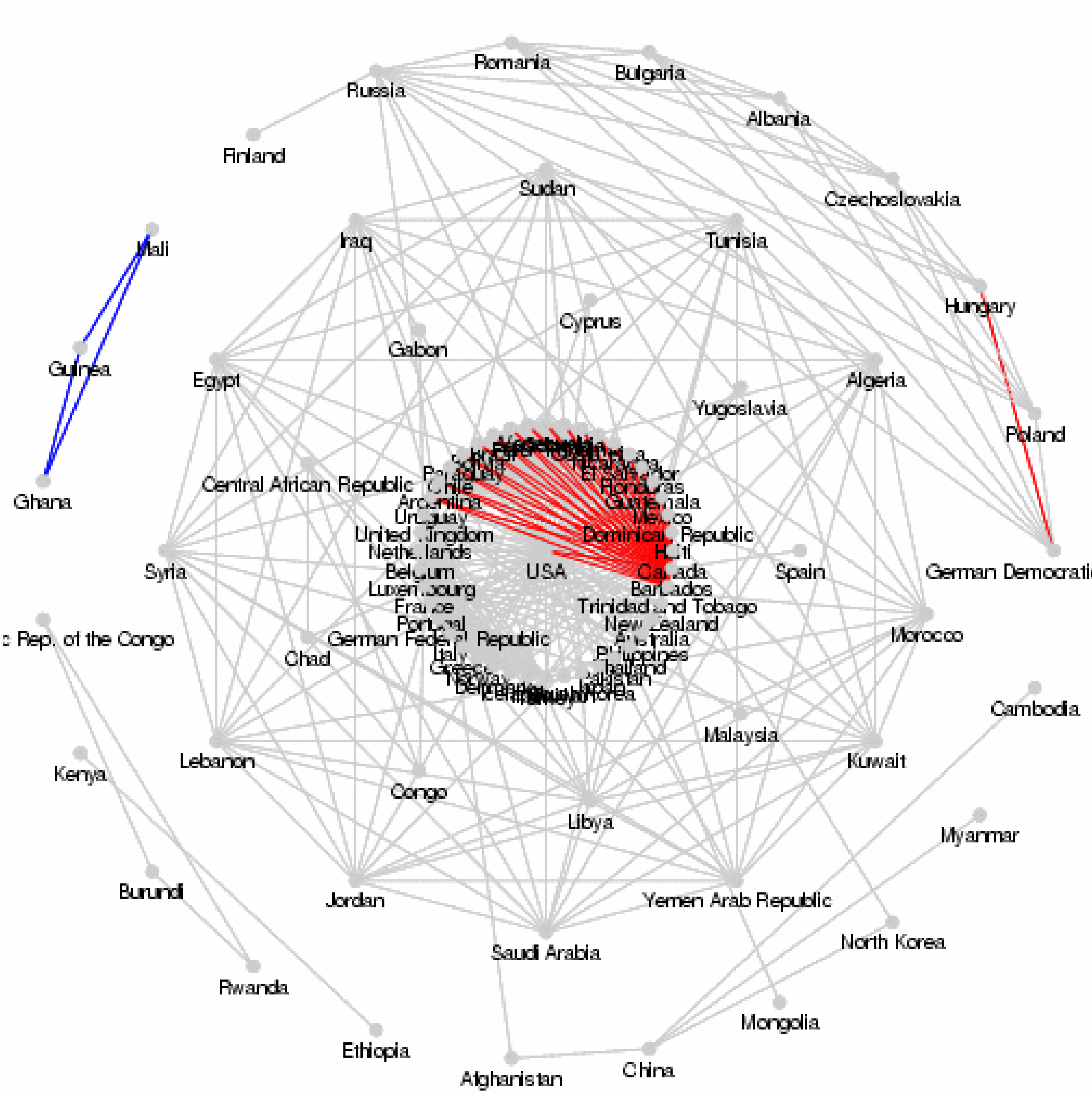


Shows the new alliances (red edges) and discarded alliances (blue edges) in 1949. The gray edges are those alliances that are in force for both years.

These two scale *0* detections are easy to detect by simply looking for an increase in degree. While the 1936 detection is also a clique, and hence detectable at higher scales, it is not necessary to go to higher scales to make the detection. Therefore, we now turn to the higher scale detections that are not detectable by simply investigating vertex degrees.

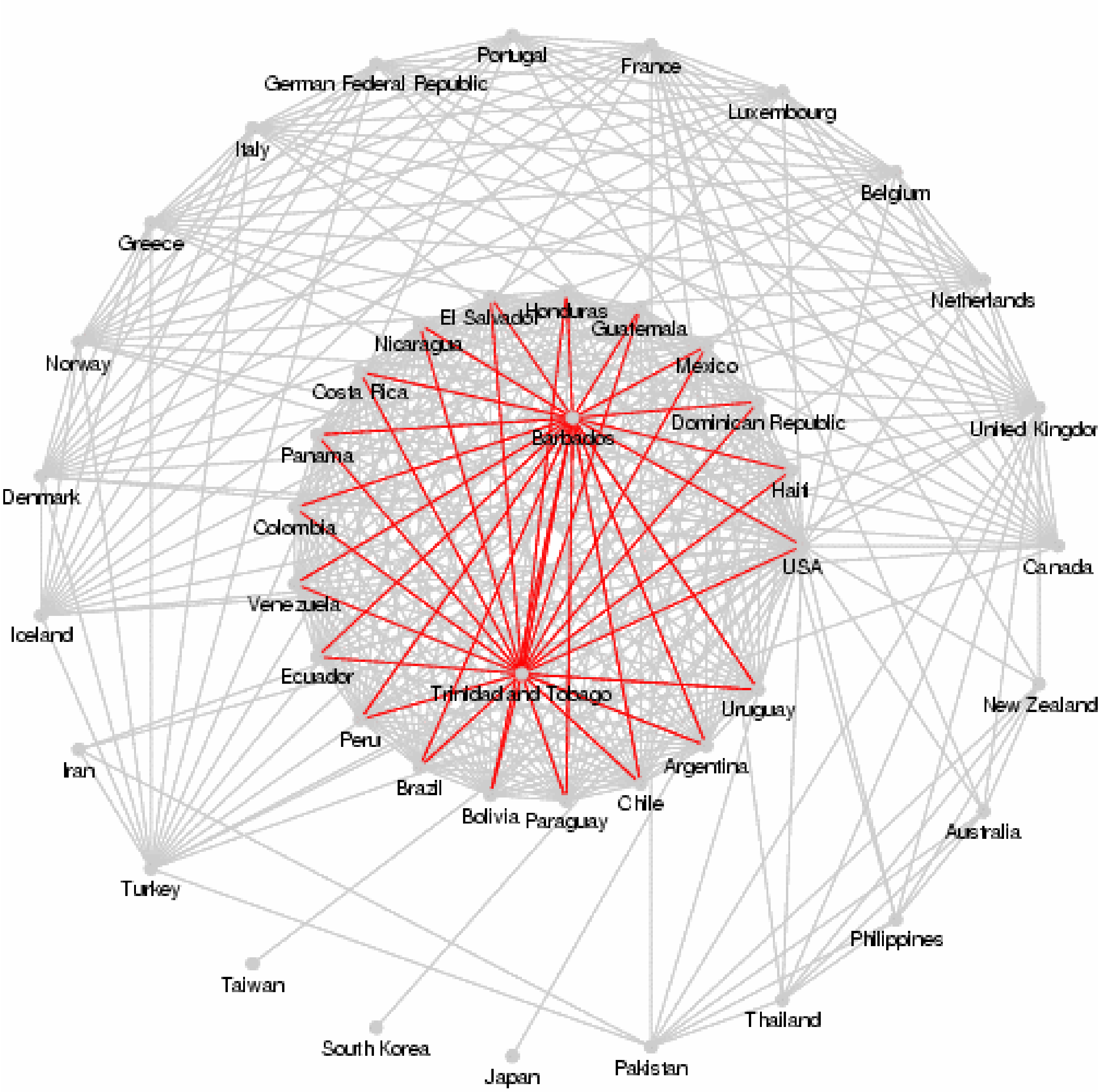
The graph for 1967, also not detected at $k=0$, is displayed in Figure 10. The change occurs in the central circle, which is represented in Figure 11 (in a slightly different layout). This is the result of new alliance between Barbados, Trinidad and Tobago, and the US and South and Central America. This detection is essentially the result of the two countries, Barbados and Trinidad and Tobago, entering into alliances with the other American countries. This detection was not above threshold in degree, although it might have been at a lower threshold.

Figure 10: **The Graph in 1967**



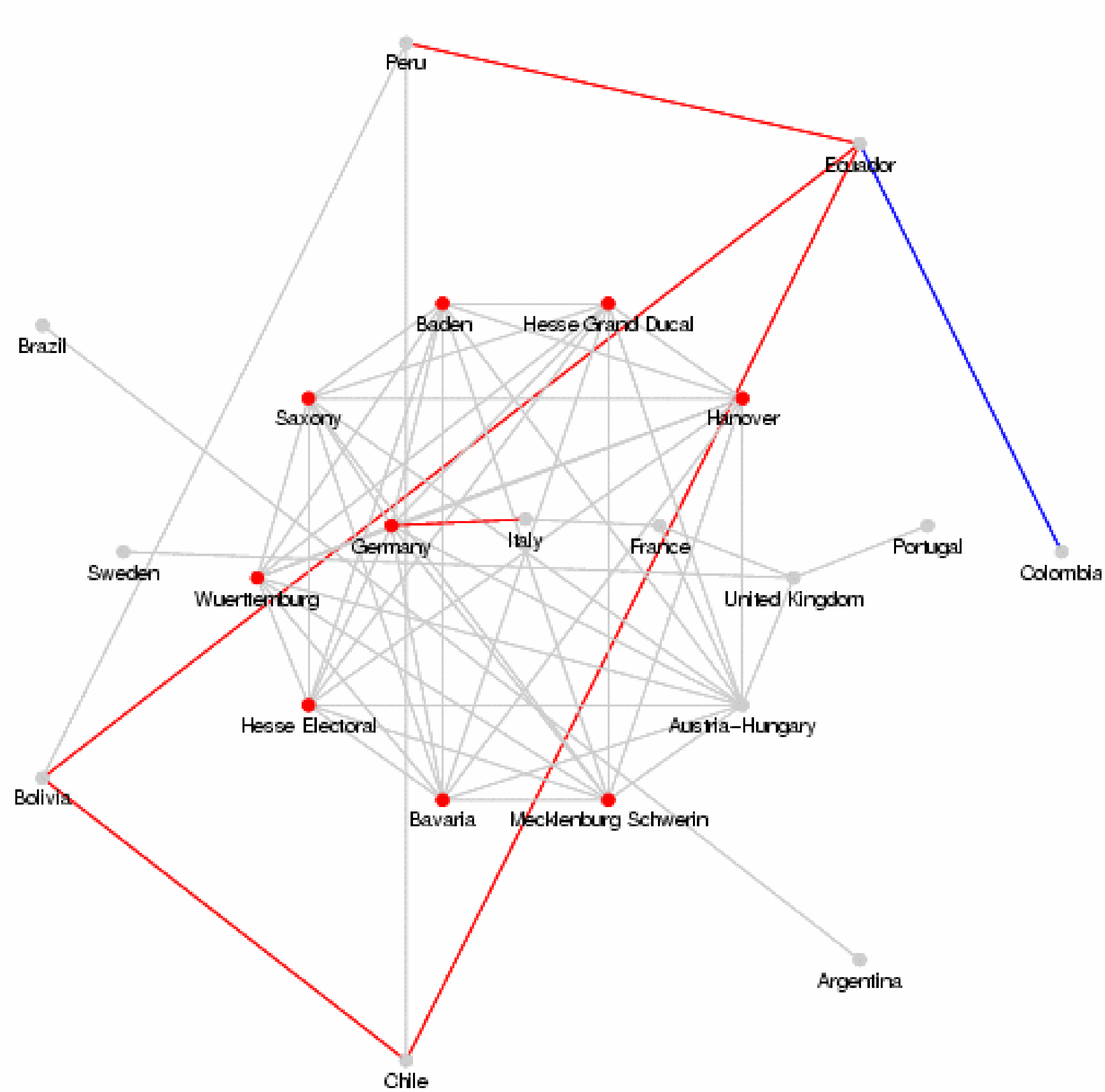
Note: Color coding of edges is as above. The graph is too large to easily display the country names.

Figure 11: **The Subgraph of 1967**



Note: When *k=1* detection occurs. The red lines correspond to the alliances added in 1967.

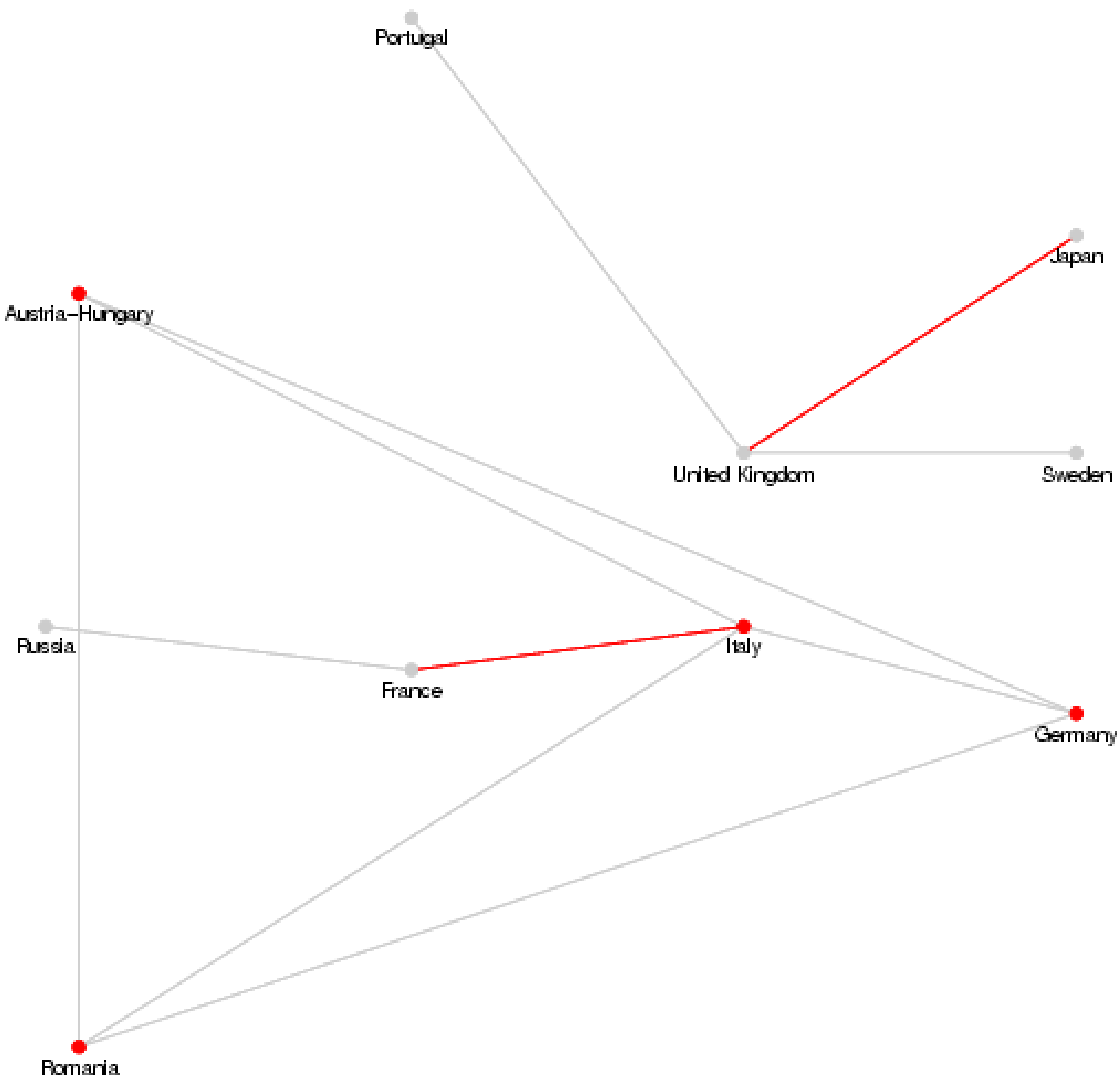
Figure 12: **The graphs of 1865 and 1866**



The color scheme for the edges is the same as above. The red vertices are the new 2-neighbors of Italy, resulting from the alliance with Germany.

We now consider the scale *2* and *3* detections. Figure 12 illustrates the maxim "the friends of my friend are my friends." Here, Italy made an alliance with Germany in 1866, thus resulting in a much larger 2-neighborhood than in the previous year. Instead of just the United Kingdom, France and Austria-Hungary in it's 2-neighborhood in 1865, now it adds the eight additional countries that the alliance with Germany brought with it. This does not necessarily mean that these alliances can now be relied upon by Italy, but to some degree it affords Italy some of the benefits of these alliances. This illustrates the fact that small changes in the graph can result in large changes in the scan statistic.

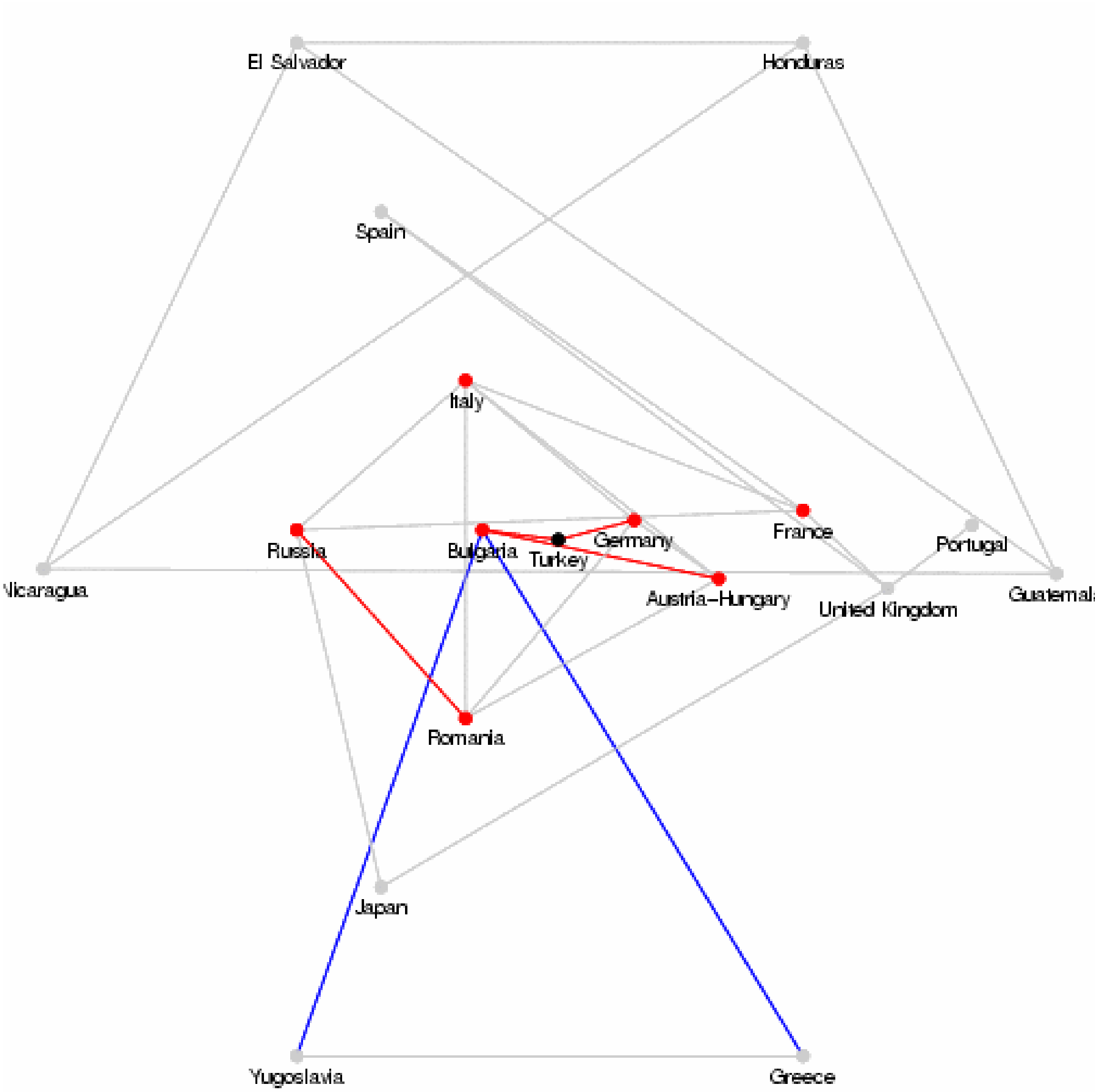
Figure 13: The Graphs of 1901 and 1902



The color scheme for the edges is the same as above. The red vertices are the new 2-neighbors of France, resulting from the alliance with Italy.

Similarly, Figure 13 shows the addition of an alliance with Italy increasing France's meager 2-neighborhood by four more countries. Similarly, Turkey made some new alliances in 1914, which, although it increased its 2-neighborhood substantially, was not enough to meet our 5 standard deviations threshold. It did, however, result in a large enough 3-neighborhood, as illustrated in Figure 14.

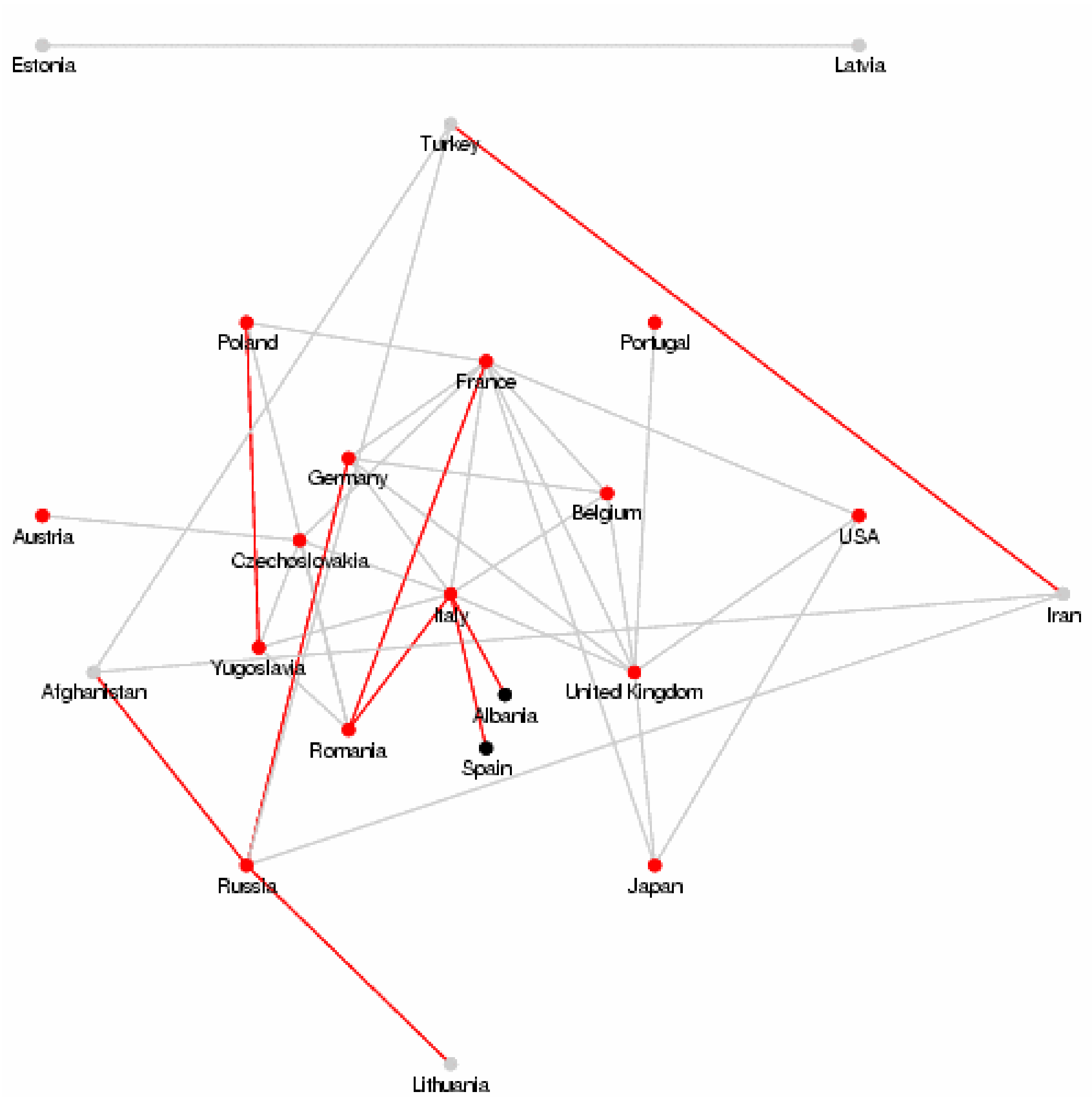
Figure 14: The Graphs of 1913 and 1914



The color scheme for the edges is the same as above. The red vertices are the new 3-neighbors of Turkey.

The graph for 1926 is shown in Figure 15, with the new 3-neighborhood shown in red. In this case, both Spain and Albania have new alliances with Italy, resulting in the same 3-neighborhood for each country.

Figure 15: **The Graphs of 1925 and 1926**



The color scheme for the edges is the same as above. The red vertices are the new 3-neighbors of both Spain and Albania, the two vertices that produce the detection.

A case can be made that the scale *2* detection for Italy is an interesting one, but it is less clear that the scale *3* detections are important. In these cases it might be more interesting to consider another locality statistic, such as density, which scales by the neighborhood size, so that detections cannot be the result of simply adding a connection to a large clique. In other applications, other statistics may be appropriate. The choice of locality statistic can determine the type of detections that can be found and the ultimate meaning of the detections. It should also be noted that these graphs tend to have a relatively small diameter, and since it makes no sense to consider scales at or above the diameter of the graph; this is another reason that the scale *3* detections might not be interesting.

Conclusions

Statistical inference on time series of graphs using scan statistics allows the detection and identification of local structural changes – a small number of vertices changing their interaction pattern over a small time scale. The methodology applied to interstate alliance graphs provides detections of numerous anomalous events - some with clear geopolitical/historical bases and some which are more subtle. The most interesting detection presented here, in our opinion, is the NATO alliance depicted in Figure 9. This shows the power of the scan statistic: it detects changes in the number of alliances among allies, and, in this case, even in the presence of a near-clique.

We have demonstrated the analysis of one type of locality statistic, the size of the induced subgraph. There are many other locality statistics that could be used on these data. The size invariant is well-suited for detecting "chatter," or the increases in the number of relationships among the actors. Density and clique number (or other measures of "cliqueness") are obvious candidates for locality statistics. A measure of the "centrality" of the node within the induced subgraph may be of interest for some applications.[1]

The locality statistic need not be a purely graph-theoretic invariant. For example, if there are covariates on either the actors or the links, the locality statistic could utilize these. For a concrete example, if data contain the amount of trade between the countries, two locality statistics of interest might include the average amount of trade within the neighborhood and the number of pairs with trade above (or below) a certain threshold. Even in this example, there are three types of alliances, and one could design a locality statistic (or a set of such statistics) which treated these different types rather than collapsing them into a single link type. For example, if one were willing to order the different types and assign them values according to increasing value of the alliance, one could compute the average weight, giving a measure of how strong the alliances are. The scan statistic methodology is quite general, and we expect that users will find clever ways to make interesting discoveries on a wide range of problems.

There are other approaches to detect changes in the structure of the graphs based on assigning a probability models to the edges and on edge structures and dependencies. The review by Snijders (2005) is a good starting place to investigate this literature. Hoff et al. (2002), Hoff (2005), and Marchette and Priebe (2008) also contain latent variable models. A different approach is taken in Shoubridge et al. (2002), in which a graph distance measure is used to compare successive graphs in the time series. The scan statistic could utilize one of these distance measures, applied to the neighborhoods, as the locality statistic.

There are several points to be considered for future work. Missing data were essentially ignored in this study, and future work will consider appropriate methods to deal with this. A second issue is that the categorical nature of the alliance relation was not used. Extensions of the scan statistic methodology to weighted edges are straightforward, but the proper extension to categorical data needs further research. Finally, tailoring the locality statistic to detect specific types of changes is somewhat of an art, and must take into account computability as well. Methods for crafting easily computable invariants (or reasonable approximations) for the detection of specific structures are of considerable interest.

[1] For example, it could determine the rise and fall of key players in a social structure.

References

Barnett, G.A. 2001. “A longitudinal analysis of the international telecommunication network.” *American Behavioral Scientist*, 44: 1638-1655.

Bollobas, B. 2001. *Random Graphs*. Cambridge University Press, Cambridge.

Gibler, D.M. and Sarkees, M. 2004. “Measuring alliances: the correlates of war formal interstate alliance data set, 1816-2000.” *Jounal of Peace Research*, 41: 211-222.

Glaz, J., Naus, J., and Wallenstein, S. 2001. *Scan Statistics*. Springer.

Hoff, P.D., Raftery, A.E., and Handcock, M.S. 2002. “Latent space approaches to social network analysis.” *Journal of the American Statistical Association*, 97: 1090-1098.

Hoff, P.D. 2005. “Bilinear mixed-effects models for dyadic data.” *Journal of the American Statistical Association*, 100: 286-295.

Mahutga, M.C. 2006. “The persistence of structural inequality? A network analysis of international trade, 1965-2000.” *Social Forces*, 84: 1863-1889.

Marchette, D.J. and Priebe, C.E. 2008. “Predicting unobserved links in incompletely observed networks.” *Computational Statistics and Data Analysis, 52: 1373-1386*.

Moaz, Z. 2004. “Pacifism and fightaholism in international politics: a structural history of national and dyadic conflict, 1816-1992.” *International Studies Review*, 6: 107-133.

Moaz, Z., Terris, L. G., Kuperman, R. D., and Talmud, I. 2004. "International Relations: A Network Approach." In New Directions for International Relations, Mintz, A. and Russett, B., eds., Lexington Books, Lanham, MD.

Priebe, C.E. 2004. “Scan statistics on graphs.” Technical Report 650, Applied Mathematics and Statistics Department, Johns Hopkins University.

Priebe, C.E., Conroy, J.M., Marchette, D.J., and Park, Y. 2005. “Scan statistics on Enron graphs.” *Computational and Mathematical Organization Theory*. 11: 229-247.

Shoubridge, P. Kraetzl, M., Wallis, W., and Bunke, H. 2002. “Detection of abnormal change in a time-series of graphs.” *Journal of Interconnection Networks*, 3: 85-101.

Smith, D.A. and White, D.R. 1992. “Structure and dynamics of the global economy: network analysis of international trade, 1965-1980.” *Social Forces*, 70: 857-893.

Snijders, T.A.B. 2005. “Models for longitudinal network data.” In *Models and Methods in Social Network Analysis*, Carrington, P.J., Scott, J., and Wasserman, S., eds., Cambridge University Press, Cambridge.

Ward, M.D., Hoff, P.D., and Lofdahl, C.L. 2003. “Identifying international networks: latent spaces and imputation.” *Dynamic Social Network and Analysis: Workshop Summary and Papers*. The National Academic Press, Washington, DC.

Ward, M.D., Siverson, R.M., and Cao, X. 2005. “Everybody out of the pool! Remodeling the democratic peace.” http://faculty.washington.edu/mdw/pdfs/eootp09112005SS.pdf.
http://www.correlatesofwar.org Correlates of War Homepage.

Preference or Propinquity? The Relative Contribution of Selection and Opportunity to Friendship Homophily in College

Jenny Godley
Department of Sociology,University of Calgary, Alberta, Canada

Abstract

This paper examines the relative importance of preference and propinquity as determinants of socio-demographic homophily in friendship choice among students at a small college in the Northeastern United States. Using unique retrospective data, the paper first assesses friendship homophily over the four years of college. Friendship is homophilous across gender and race. QAP regression is used to determine the impact of both preference and propinquity (measured by participation in joint extra-curricular activities and shared academic major) on friendship choice. While preference predicts friendship choice in the freshman year, propinquity remains the strongest determinant of friendship choice over the four years.

***Keywords**: Friendship networks; college students; homophily; preference and propinquity*

Acknowledgments: The author acknowledges the invaluable enthusiasm and research assistance of the members of the social networks class at Arbor College during the spring 2002 semester. Mark Gould, James Moody and Martin Ruef provided assistance with survey design, project implementation, and statistical analysis. Participants at the XXIV International Sunbelt Social Network Conference in Portorož, Slovenia in 2004 provided feedback on the first draft of this paper. Valerie A. Haines provided keen input and immeasurable support during the most recent rounds of analyses and writing. All mistakes are the sole responsibility of the author.

Jenny Godley is an Assistant Professor of Sociology at the University of Calgary. As well as studying the friendship patterns of adolescents and young adults, she uses social network analysis to understand the effects of social ties on demographic and health behavior across the life course and internationally. Additional research interests include the social patterning of body weight, physical activity and obesity, gender stratification cross culturally, and international education.

For correspondence, please contact Jenny Godley, PhD, Department of Sociology, University of Calgary, 2500 University Dr., NW, Calgary, AB T2N 1N4 Canada, Tel. (403) 220 7566, Fax (403) 282 9298 , Email: jgodley@ucalgary.ca.

Introduction

Friendship choice has long interested sociologists as an example of a micro-level decision that is constrained by macro-level features of social structure such as population size, composition, and distribution (Adams and Allan, 1998). Sociological studies of friendship choice tend to focus on three areas: the patterning of friendship choices within a particular environment; how individuals make friendship choices; and the impact of friendship choices on behaviors, opinions, or attitudes. This paper contributes to the literature on the patterns and determinants of friendship choice by using unique retrospective whole network data to examine the determinants of socio-demographic homophily in friendship choice among the members of the 2002 graduating class of a small liberal arts college in the Northeastern United States.

The whole network data are analyzed descriptively first to assess the socio-demographic homophily of friendship over the four years of college. Next, data on choice of academic major and comprehensive retrospective data on club and team membership are used to determine the relative importance of preference and propinquity on overall cross-year friendship homophily. Finally, the data are analyzed separately by year to examine the changing relative influence of preference and propinquity on friendship homophily over the four years.

Literature Review

One of the most resilient findings in the sociological literature on friendship is that friends tend to be similar across socio-demographic characteristics, opinions and attitudes, and even behaviors. In one of the earliest studies on homophily, reporting on a set of interviews conducted from 1976-1980, Bell concluded that adult friendship is homogeneous on age, sex, and religion (Bell, 1981). McPherson et al.'s (2001) review of the literature on homophily in several types of relationships, including friendship, concludes that in the United States, "homophily in race and ethnicity creates the strongest divides in our personal environments, with age, religion, education, occupation, and gender following in roughly that order" (p. 415). Much of the literature on adult friendship in the United States has focused on homophily across these same characteristics.

A multitude of descriptive reports show that friends tend to be homogeneous with regards to age, gender, ethnicity, as well as behaviors and ability (Hartup and Stevens, 1997; Adams and Allan, 1998). With regard to similarity across behaviors, opinions, and attitudes, there is an ongoing debate as to whether people become friends with those who are similar to them ('selection'), or whether they change to become more like their friends ('influence').[1]

In the literature on socio-demographic homophily, researchers debate whether people pick friends based on similar demographic traits ('preference,' also referred to as 'selection'), or whether they pick friends based on the opportunity for contact ('propinquity,' also referred to as 'opportunity'). Choosing friends based on propinquity will result in demographically homophilous friendship choices in a society

[1] This paper does not address behavioral or attitudinal similarity. For examples of the literature addressing this 'selection versus influence' debate, see Kandel, 1978 and more recently Crosnoe et al, 2004.

that is segregated by social characteristics such as gender and race/ethnicity. The relative contribution of preference and propinquity to the segregation of friendship groups in school and college settings has been the subject of much educational research.

Research at all levels of education, including elementary schools, high schools and post-secondary institutions, has shown a strong tendency towards homophily in friendship choices based on gender and race/ethnicity (Kupersmidt et al., 1995). Most researchers conclude that within the educational setting, both preference and propinquity probably play a role in creating homophilous friendship choices; it is often hard to tease apart the independent effects of each. For example, educational tracking may result in students of similar socio-demographic backgrounds being placed in classes together. Thus, when students in these classes befriend one another, preference and propinquity are conflated.

Responding to desegregation policies implemented during the 1960's and 70's, many US researchers analyze friendship homophily by race/ethnicity in public educational settings (Baerveldt et al., 2004). Most find that race is stronger than socio-economic class in predicting friendships, but find it hard to assess whether this is due to preference or propinquity. Using data from the 1980 High School and Beyond Study, Hallinan (1989) and Kubitscheck and Hallinan (1998) claim that preference (across race and gender) and propinquity are both determinants of friendship choice. They demonstrate that same-race friendships are far more common than interracial friendships, and argue that tracking reinforces similarity. Provocatively, they claim that in the United States, tracking can effectively re-segregate a desegregated school.

Fifteen years later, using the Add Health data, Quillian and Campbell (2002) examined multiracial friendship segregation. Arguing that cross-race friendships other than just black-white should be examined, they looked at four racial/ethnic groups: black, white, Asian and Hispanic. They found that while relative group size was important, there were racially segmented patterns of assimilation in almost all schools. In particular, students in small racial minorities tend to have own-race friends. They also conclude that both preference and propinquity play a role in racial homophily.

Several US researchers have examined racial homophily in the friendship networks and peer groups of college students (Newcomb and Wilson, 1966; Salzinger et al, 1988; Portnova, Lock, Ladd and Zimmerman, 2006). Some researchers have argued that contact between racial groups is higher in college than in other educational settings (in residence halls for example), and thus interracial friendships are more likely at the college level (Stearns et al., 2004; Levin et al., 2002).

A few educational researchers have looked at both friendship and participation in shared activities, which may provide opportunities for mixing (Feld and Carter, 1998). Clotfelter (2002) examined high school yearbooks to assess interracial contact in high school extracurricular activities, although he did not have specific data on friendship ties. He found that in most high schools, organizations were not racially balanced. He found lower rates of participation of non-whites, and evidence for selection into clubs/organizations by race. However, he does argue that these memberships may provide a way to meet people interracially, outside of friendship.

Using the nationally representative Add Health data from the 1990's, Moody (2001) found that both individual preferences and segregated activities lead to racial homophily in friendship choice in high schools. He found that both student behavior and the organization of schools influenced homophily. While there was a preference for similarity across social class, popularity, academic performance, gender, and delinquent behavior at the individual level, student mixing opportunities and the climate at the school level also affected friendship choice, leading to more homophily. Moody found that mere exposure did not promote integration, and that interracial mixing only happened in activities where people of equal status were mixing, such as extracurricular activities.

While longitudinal friendship network data is becoming more common, most longitudinal datasets are not designed to test preference versus propinquity (Hallinan and Williams, 1987; Leenders, 1996; and articles in the recent issue of Social Networks (Volume 27, 2005)).

To summarize, data limitations have prevented previous researchers from assessing the relative importance of preference and propinquity on friendship homophily. This paper employs a new dataset, with retrospective data on friendship and joint memberships, to explore the following:

A. *The socio-demographic homophily of friendship over four years.*
B. *The relative influence of preference and propinquity on friendship homophily.*
C. *The changing relative influence of preference and propinquity on friendship homophily over the four years of college.*

Setting

Arbor College[2] is a small, residential liberal arts college in the suburbs of a major city in the Northeastern United States. An academically elite institution, Arbor attracts high achieving students. While most of the students come from the Northeastern states and California, students also come from all regions of the US and from abroad. In 2002, Arbor had 1,702 students enrolled; 283 were enrolled in their senior (final) year.

Descriptive information on the population of Arbor College is available via the Common Data Set Initiative at http://www.commondataset.org/. The following information was compiled and published online for the school year 2004-2005. Figures for the graduating class of 2002 are no longer available online or from the college. The 2004-2005 figures should reflect the 2002 graduating class, as there have been no changes in admission requirements or retention rates.

Most of the students at Arbor College live on campus in co-ed dorms and apartments. In 2004-5, 100% of the freshmen students lived in on-campus housing, and 99% of all students lived in on-campus housing. Arbor has a very high retention rate, at over 90% of freshmen graduating four years later. Transfers made up 10% of the graduating class in 2005, with most students transferring at the beginning of their sophomore year.

Thus, this survey takes place in a small, exclusive, insular setting, where most students spend four years in the same graduating class. While the students do have the opportunity to take classes at other institutions of higher education in the area, most take the majority of their classes on the Arbor campus. Although the campus is not far from a major city (accessible by public

[2] Arbor College is a pseudonym, as required by the Institutional Review Board which approved this research.

transportation), many of the students do the majority of their socializing on campus (Newshel and Author, 2004); therefore, opportunities for contact with other students are extremely high. This is a very insular campus, thus the focus on internal friendship ties among students on campus is not as problematic as it would be at a college where students had (or retained) off-campus friends.

Data Collection

The friendship network survey targeted all the members of the senior (graduating) class at Arbor College in 2002. It was designed and conducted by students in an undergraduate sociology class taught by the author. Nine students assisted the author with the survey design and implementation. The survey was approved by the Institutional Review Board at Arbor College.

The survey was conducted online in February, 2002. An email was sent to all the members of the 2002 graduating class, with a link to the survey. As an incentive to take the survey, we offered a $300 prize awarded by lottery. In order to respond to the survey, students had to check their email (a requirement for most classes at Arbor College) and have access to a computer. Most students had their own personal computer in 2002; public computers were also available across campus. Once they clicked on the link to the survey, respondents were taken through the informed consent process before accessing the survey. The survey took 15-20 minutes to complete, and students had to complete the survey in one sitting.

Respondents were asked to choose their five best friends in the senior class from a list of all the members of the senior class organized alphabetically by last name and obtained from the Registrar's Office at the beginning of January, 2002. Respondents were instructed to include friends of both genders, and to include romantic partners as friends. Respondents were asked when they had met each friend (before college, freshman, sophomore, junior, or senior year).

They were also asked their academic major (chosen at the end of their sophomore year) and to list every club they had belonged to each year of their time at Arbor College. The list of clubs was obtained from the student's union and from the student's activity coordinating office. Arbor is a very active campus; there were over 125 clubs listed in the survey for each year. This club membership data were used to construct a joint membership matrix for each year and an overall joint membership matrix.

Finally, respondents were asked to provide their own demographic information. They were asked their gender, age, educational level of their parents, religion, and how they identified in terms of race/ethnicity.

While the data on club memberships are fully retrospective, the friendship network data are only partially retrospective. Respondents were asked to nominate their five best friends at a particular point in time (halfway through their senior year). They were asked when they met each of these friends, but they were not able to list all of the friends they had during each year of college. Thus only friendships which have endured through to the senior year are included in the friendship network. We have no data on friendships that were made earlier in the four years at college and then dissolved, or became less important. The benefit to asking retrospectively about friendship this way is that we gather data on strong friendships; the downside is that we do not have a full picture of the composition of respondents' friendship networks over time.

Also importantly, respondents were only allowed to nominate other members of the senior class as friends. Thus, we have no information on friends who are not members of Arbor College, or indeed friends within Arbor College who were not members of the 2002 graduating class. This may create some bias in the data, especially if certain groups are more likely to have cross-class or outside-college friendships. I speculate about the effects of bounding the network this way in the limitations section of the paper.

Sample

We had 218 respondents out of a total senior class of 283 – a response rate of 77%. Due to the nature of whole network data, where socio-demographic characteristics are gathered from individuals themselves and not from their friends, we had to eliminate all nominated friends who did not respond to the survey. Thus, the final full friendship network consists of the 218 students who responded to the survey.

On average, students who answered the survey listed 4.5 friends. Once we eliminated those friends who did not take the survey, students were left with an average of 3.7 friends. By eliminating the 65 members of the senior class who did not take the survey, we lost information on 18% of the respondents' friendship ties. Thus 82% of the data on friends was retained, providing enough information to proceed with whole network analysis. Descriptive statistics for the sample are shown in Table 1, below

Table 1. Sample Demographics

	Variable	**Percent**
Gender:	Male	44
	Female	56
Race:	White	85
	Non-white	15
Parents' Education:	Less than college	22
	College Graduate	26
	Master's Degree	24
	Doctoral Degree	28
Religion:	Catholic	14
	Other Christian	22
	Jewish	15
	Multiple / Other	10
	Spiritual, but not religious	8
	No religion	31

Just over half the respondents were female. Respondents were asked to report on the educational level of the parent(s) they lived with during high school. The highest level of parental education was taken as a proxy for social class. As Table 1 shows, these students come from highly educated families; almost a third have at least one parent with a PhD. For all analyses, sensitivity tests were conducted on the social class measure. The measure was included as a four category variable, as a dichotomous variable coded 'college and less than college' versus 'post graduate degree,' and as a dichotomous variable coded 'less than college' versus 'college graduate and above.' Across all analyses, the three different measures of social class produced almost identical results. Thus, only results using the four-category variable are presented in the paper.

Respondents self-reported on their race/ethnicity, and were allowed to pick more than one category from the following list: Hispanic/Latino; White/Caucasian; Black/African American; American Indian/Alaskan Native; Asian; Native Hawaiin/Pacific Islander; and Other. The number of respondents who chose any non-white category was low (32). Because of the risk of deductive disclosure, it was necessary to collapse all the non-white categories into one category. Any respondent who checked at least one non-white category (even if they also checked white) was counted as 'non-white." All analyses concerning race/ethnicity in this paper will focus on the difference between whites and non-whites.

The only socio-demographic data available for Arbor College as a whole is the distribution of students by gender and race/ethnicity. As mentioned above, this data is not available for the graduating class of 2002. However, data from the CDI show that in 2004-2005, students at Arbor College were 53% female, and 71% White/Caucasian (the non-white category was broken down into 6% Black, 13% Asian, 6% Hispanic, and 3% foreign-born). Thus, while the gender distribution of our sample is close to the overall distribution at Arbor, we can see that non-whites are under-represented among our survey respondents.

It is important to remember that because non-whites are under-represented as respondents, they are also under-represented as friends. One way to examine this issue is to look at the percentage of friendship ties lost when we eliminate the non-respondents as potential ties. As mentioned above, overall we lost 18% of friendship ties with this elimination. For non-whites, though, we lost 30% of friendship ties. The difference between the percentage of ties lost for whites and non-whites is statistically significant, indicating that non-whites were more likely to nominate people who did not answer the survey. Although we don't know anything about the characteristics of these non-respondents, we can speculate that they may be non-white, too. As I discuss below, given differential response rates, we may be under-estimating racial homophily for non-whites, and over-estimating racial homophily for whites.

Arbor College students are very active in clubs and teams. On average, students participate in approximately 2.5 clubs a year. Over the four years of college, club memberships for the 2002 graduating class were distributed as follows: academic clubs 12%; student government 13%; activist / political clubs 40%; theatre / music 29%; team sports 45%. Table 2, below, shows the distribution of club membership by gender, both with and without team sports (which are gender-segregated).

Table 2. Club Membership by Gender and Year

	Freshman Year	Sophomore Year	Junior Year	Senior Year	Total over four years
All Clubs					
Total	2.4	3.1	2.2	2.2	9.9
Boys	2.7	3.1	2.2	2.4	10.4
Girls	2.3	3.1	2.1	2.0	9.5
Clubs without sports teams					
Total	1.8	2.6	1.8	1.8	8.0
Boys	2.2	3.0	2.2	2.3	9.7
Girls	1.5	2.3	1.4	1.4	6.6

As Table 2 illustrates, club membership is highest in year two. There is no gender difference in club membership. Because team sports' membership is gender constrained, it will be important to examine the effects of club membership on friendship homophily by gender both with and without sports teams included. Club membership does not vary significantly by white/non-white status, social class, religion, or major.

Arbor students choose their academic major at the end of their sophomore year. In the 2002 graduating class, 28% of the students were social science majors, 26% were natural science majors, and 46% were humanities/arts or double majors. For the purposes of this paper, shared academic major is used to represent propinquity rather than preference for an academic subject. Students taking the same majors are likely to meet in classes, academic departments, laboratories, and libraries, especially during their junior and senior years. Thus, shared academic major will be tested along with joint club membership to assess the influence of propinquity on friendship selection.

Data Analysis and Results

There are 808 friendship ties reported in the data. On average, students met 64% of the friends they listed in their senior year during their freshman year at Arbor College (519 ties). An average of 2% (23 ties) were met before college, 17% (135 ties) in their sophomore year, 11% (84 ties) in their junior year, and 5% (47 ties) in their senior year (the survey was taken half way through the senior year). There were no significant differences in the percent of friends met each year by any of the socio-demographic characteristics.

The Socio-Demographic Homophily of Friendship Over Four Years

Friendships ties among the members of the 2002 senior class at Arbor College are disproportionately homophilous by gender and race, but not by social class or religion. On average, respondents nominated 67% same gender friends. Girls were more likely to nominate same gender friends than boys (74% versus 59%), but this difference was not statistically significant. While this gender homophily is not as high as reported elsewhere amongst college students, the inclusion of romantic partners in the friendship roster may have reduced gender homophily for heterosexual students.

In terms of white/non-white status, 83% of friendship nominations were to same race friends. Among non-white students, 34% of friendship ties were to other non-white students. Among white students, 91% of friendship ties were to other white students. While both whites and non-whites are disproportionately likely to choose same race friends, whites are significantly more likely to choose same race friends. However, as mentioned above, non-white students were less likely to answer the survey than white students. Thus, we may be over-estimating whites' racial homophily, and under-estimating non-whites' racial homophily.

With reference to the other demographic variables, on average 24% of the students' friends were from the same social class and 23% had the same religion. Neither of these homophily figures is disproportionate, given the distribution of social class and religion in the sample. There were no significant differences between the social classes or between religious groups in terms of homophily on these variables.

To assess the relative importance of socio-demographic homophily over time, Table 3, below, examines friendship homophily by year met. When examining Table 3, it is important to remember the nature of the retrospective friendship data. Table 3 does not indicate the socio-demographic characteristics of all the friends each respondent met each year. Instead, Table 3 indicates the socio-demographic characteristics of all friends whom the respondent still considers important friends in his/her senior year by the year they met those friends. So we can say, for example, that of all the friends met during freshman year whom respondents still consider friends during their senior year, 85% are the same race, and 69% are the same gender.

Table 3. Socio-demographic Homophily of Friendship Ties Reported in Senior Year by Year Met

	Before College	**Freshman**	**Sophomore**	**Junior**	**Senior**
% met that year	2	64	17	11	5
% same gender	75	69	63	72	67
% same social class	6	22	35	26	26
% same religion	19	21	32	20	33
% same race	81	85	83	80	58

The homophily figures are remarkably similar across the years met, with respondents disproportionately choosing same-race and same-gender friends, but not same-class or same-religion friends. There is only one statistically significant difference over time in Table 3. The percent of friends met during the senior year who are the same race (58%) is significantly lower than all other years.

__The Relative Influence of Preference and Propinquity on Friendship Homophily__

To assess the relative influence of preference and propinquity on friendship choice, this section of the paper will test whether joint club membership and shared academic major determines friendship, net of socio-demographic similarity. First, I will examine the overall cross-year friendship network (i.e. all the friends reported by respondents during their senior year, regardless of when they met). In the next section, I will look separately at the network of friendships formed each year.

Following Carley and Krackhardt (1996), Brewer and Webster (1999), and Burris (2005), I use the Quadratic Assignment Procedure (QAP) regression technique to model the independent effects of preference and propinquity on friendship choice. This procedure is implemented in UCINET 6 (Borgatti et al., 2002). There are 23,653 dyads created by multiple relations among the 218 students. These dyadic observations are not statistically independent, thus the data violates the assumptions of Ordinary Least Squares (OLS) regression.

The QAP regression procedure, which overcomes the limitations of autocorrelation, is best understood as a form of simulation (Burris, 2005). First, OLS coefficients are calculated for the independent variables in the regression. Next, the rows and columns of the dependent variable matrix are randomly permuted and the OLS regression coefficients are re-calculated. The simulation is repeated 2,000 times in UCINET 6. The initial regression coefficients are then compared with the distribution of all possible coefficients, and significance tests are based on these distributions.

Burris (2005) argues that when interpreting QAP regression results, the focus should be on the comparative magnitude of the coefficients, rather than on the overall model R2 or the level of statistical significance for each coefficient. In Tables 4 and 5 (below), I report the standardized coefficients for each independent variable, and their significance level. Discussion will focus on the comparative magnitude of those coefficients which are significant.

Examining the overall cross-year friendship network, Table 4, below, illustrates the effects of the demographic variables on friendship choice (Model I), the effects of shared academic major and joint club membership including all clubs and teams on friendship choice, controlling for the demographic variables (Model II), and the effects of shared academic major and joint club membership excluding sports teams on friendship choice, controlling for the demographic variables (Model III).

The dependent variable for all these models is the complete friendship choice network coded as a valued network where 0 represents no tie, 1 represents a non-reciprocated tie, and 2 represents a reciprocal tie (just over half of the ties in the whole network were reciprocated).

All models reported in this section were also run on the binary networks containing simple outties. Results did not differ substantially between the binary and the valued friendship matrices for any of the models. The valued networks are theoretically more interesting, as reciprocal ties are theoretically stronger than non-reciprocated ties. Thus, I present the results using the valued networks here. Results using the non-valued networks are available from the author.

The independent variables are the similarity matrices for the demographic variables (same gender, same race, same religion, same social class, and same major) and the affiliations matrix for club membership. The affiliations matrix is valued, with the value of (X,Y) as the number of shared club memberships between X and Y cumulated over the four years of college.

Table 4. Overall Cross-Year Friendship Network QAP Regression Results

	Model I	Model II	Model III
Same gender	0.047 ** (0.000)	0.032 ** (0.000)	0.040 ** (0.000)
Same race/ethnicity	0.030 ** (0.000)	0.021 ** (0.000)	0.022 ** (0.000)
Same social class	-0.002 (0.385)	-0.002 (0.403)	-0.002 (0.390)
Same religion	0.013 (0.028)	0.010 (0.051)	0.011 (0.049)
Same major	-----	0.012 (0.026)	0.012 (0.029)
Number of shared clubs including sports teams	-----	0.157 * (0.000)	-----
Number of shared clubs NOT including sports teams	-----	-----	0.121 ** (0.000)
R2	0.003	0.028	0.018

* significant at the 0.001 level

Table 4 shows standardized coefficients and proportion significance in parentheses. The values represent the effects of demographic similarity, shared academic major, and number of joint club memberships on the valued friendship choice network

Model I in Table 4 indicates that controlling for religious, social class, and major similarity, gender and racial homophily are significantly predictive of friendship ties. Models II and III demonstrate that gender and white/non-white status similarity remain significantly predictive of friendship ties, controlling for joint club and team membership. Joint club and team membership (both with and without sports teams included) are also significantly predictive of overall friendship ties, net of demographic similarity.

One complicating factor in examining the effect of joint affiliations and the effect of demographic similarity on friendship is that demographically similar individuals may join the same clubs. Comparing Models II and III in Table 4 shows the effect of excluding the sports teams from the affiliations matrix. As sports teams are gender exclusive, they are one activity that is selective on one of the demographic variables (gender). As we would expect, excluding sports teams from the affiliations matrix (Model III) increases the effect of gender similarity on friendship. However, it is important to note that gender similarity remains a significant predictor of friendship, even when sports teams are included in the affiliations matrix (Model II). Thus the effect of gender homophily on friendship operates over and above the impact of gender on participating in joint activities. There may be other types of clubs where membership is selectively based on demographic variables (for example, clubs focused on the politics of race/ethnicity may attract mostly non-white students). However, the sports teams are the only clubs which are officially segregated, and thus the only clubs where demographic variables are a known prerequisite for membership.

The Changing Relative Influence of Preference and Propinquity on Friendship Homophily Over the Four Years of College

To determine whether the relative influence of selection versus propinquity changes over the four years of college, this section of the paper examines the friendship networks separately by year. Table 5 contains the full model for each of the four years of college. Only two percent of friendship ties reported by Arbor College students in their final year were formed before college. Looking at the network of friendship ties formed before college, none of the demographic variables is a statistically significant predictor of friendship. The models for friends met before college are not included in the paper; results are available from the author.

The dependent variable for each of these models is the friendship choice network by year coded as a valued network where 0 represents no tie that year, 1 represents a tie that was not reciprocated that year (52% of the ties were reciprocated and 78% of those were reciprocated in the same year), and 2 represents a tie that was reciprocated in that year. All models reported in this section were also run on the binary networks containing simple outties. Results did not differ substantially between the binary and the valued friendship matrices for any of the models.

The independent variables are the similarity matrices for the demographic variables (same gender, same race, same religion, same social class, and same major) and the affiliations matrices for club membership separately for each year. The affiliations matrices are valued, with the value of (X,Y) as the number of shared club memberships between X and Y during that year of college.

Table 5, below, shows the QAP regression model results indicating the effect of shared demographic variables and shared club membership each year on friendships formed during that year.

Table 5. Friendship Network by Year QAP Regression Results:

	Freshman	Sophomore	Junior	Senior
Same gender	0.036 ** (0.000)	0.009 (0.075)	0.025 ** (0.000)	0.009 (0.079)
Same race/ethnicity	0.023 ** (0.000)	0.010 (0.044)	0.004 (0.260)	0.010 (0.037)
Same social class	-0.004 (0.247)	0.009 (0.073)	-0.004 (0.230)	0.004 (0.280)
Same religion	0.003 (0.312)	0.018 * (0.003)	0.004 (0.223)	0.014 (0.020)
Same major	0.007 (0.122)	0.001 (0.438)	0.003 (0.286)	0.018 * (0.003)
Number of shared clubs, not including sports teams	0.056 ** (0.000)	0.043 ** (0.000)	0.018 * (0.010)	0.017 * (0.010)
R2	0.002	0.002	0.001	0.001

* significant at the 0.001 level

Table 4 shows standardized coefficients and proportion significance in parentheses. The values represent the effects of demographic similarity, shared academic major, and number of joint club memberships each year on the valued friendship choice network

Not surprisingly, since most of the friendship ties were made during freshman year, the results for the freshman year models are very similar to the results for the overall friendship network, as reported in Table 4, above. Gender and racial similarity have a positive influence on friendships started in during freshman year, while social class and religion do not significantly impact friendship choice. The number of shared clubs in freshman year also significantly impacts friendship choice, net of demographic homophily.

Seventeen percent of friendship ties reported by students in their senior year were formed in the sophomore year at Arbor College. Results for the models run on the sophomore year friendship network indicate a shift from the freshman year. In the sophomore year, religious similarity becomes an important predictor of friendship choice. Racial and gender similarity are no longer important. Membership in shared clubs during the sophomore year also has a significant impact on friendship ties. Thus, it appears that religious preference and propinquity influence friendship choice during the sophomore year. Interestingly, the sophomore year is also the year when students are the most active in clubs and teams (see Table 2, above).

Eleven percent of friendship ties reported in the senior year were formed in the junior year of college. Results from the models run on the junior year friendship network indicate another shift in the determinants of friendship choice. Gender similarity is the only socio-demographic variable in the models which influences friendship choice. Joint club membership is also a significant predictor of friendship choice junior year. Choice of academic major still has no significant impact on friendship choice.

Only five percent of friendships reported mid way through the senior year were formed during the senior year. In the senior year, socio-demographic preference does not predict friendship choice. Instead, shared academic major and joint club and team membership during the senior year of college impacts friendships formed that year. Propinquity, particularly in regards to academic activities, matters more than preference during the senior year.

Discussion, Limitations, and Conclusions

Descriptive statistics indicate that friendship at Arbor College is remarkably stable over time. Almost two-thirds of friendships listed during the senior year of college began in freshman year. Examining univariate results, gender and racial homophily appear to be an important feature of friendships formed during the first three years of college.

By the senior year, while gender homophily is still an important factor in choosing friends, racial homophily becomes less important.

Examining the multi-variate models, friendships appear to be determined by both preference in terms of gender and white/non-white status *and* propinquity in terms of shared major and joint activity memberships. The number of shared activities has a greater impact than shared gender or race on overall friendship ties, but gender and race still matter, net of joint memberships and shared academic major, and controlling for shared religion and social class.

The models for friendships formed in the freshman year look very similar to the models for the overall friendship network. In the sophomore year, shared religion becomes important, while joint club membership continues to influence friendship choice. In the junior year, shared gender is the only socio-demographic variable tested that has a significant impact on friendship choice. Interestingly, even though almost a third of Arbor College students study abroad during their junior year (Newshel and Godley, 2004), with most of these students away from campus for at least one semester, and many away for the whole year, opportunities to participate in joint activities are significantly reduced, and yet joint activities remain a significant predictor of friendship choice during junior year. In the senior year, joint activities are again a significant predictor of friendship, and shared academic major becomes important.

Thus we can conclude that while both preference and propinquity influence friendship choice at Arbor College, the influence of preference declines over time. Preference based on gender and race is an important determinant of friendships formed during the freshman year at Arbor College. As almost two-thirds of friendships are formed during the freshman year, homophily across gender and race remain important for the whole friendship network examined in the senior year. Religious preference becomes important during sophomore year, and gender similarity is again important during junior year.

Propinquity is a determinant of friendship formation throughout the college years, controlling for demographic homophily. Across all four years of college, joint club and team membership has a stronger impact than socio-demographic preference on friendship choice. By the senior year, joint club and team membership and shared academic major are the only significant determinants of friendship choice, controlling for socio-demographic similarity. Therefore, it appears that over time the effect of preference on socio-demographic homophily in friendship choice declines, while the effect of propinquity increases.

Arbor College is a unique setting, thus it is difficult, if not impossible, to apply the findings in this paper to any other college setting. In particular, the student body at Arbor College is extremely homogenous across race and social class. Thus the findings may underestimate the impact of selection on friendship formation in other settings. As mentioned, the differential non-response rate between whites and non-whites is also potentially problematic for the analysis of the impact of race on friendship ties.

The method of data collection, where students were asked about their five closest friendships in the senior class at Arbor, may have created other biases in the data. We have no data on friendships outside of the college, or friendships with members of another class at Arbor. Certain groups of students may be more likely to make or maintain friendships outside the college (for example, students who do not feel that they are part of the college environment might rely more on friendships they made in high school), and certain groups may be more likely to have friendships with students in other years at Arbor. The data cannot capture these variations.

Students were asked to report retrospectively on when they formed their friendships, and on their club and team membership. We know that some of the data on when the friendships formed is inaccurate, as 22% of reciprocated ties were mis-matched in terms of year met. Some of the retrospective club and team membership data may be inaccurate, also. We have no data on friendships, which were formed in previous years and dissolved, or friendships formed in previous years that became less important than the top five friendships they were able to nominate.

The current project demonstrates that within a small, academically elite, residential college environment, shared activities and club memberships are a stronger and more consistent predictor of friendship choice than socio-demographic similarity across all four years of college. In the Arbor College setting, propinquity trumps preference in accounting for socio-demographic homophily in friendship choice.

References

Adams, R.G. and Allan, G. 1998. *Placing Friendship in Context*. Cambridge: Cambridge University Press.

Baerveldt, C., Van Duijn, M.A.J., Vermeij, L., and Van Hemert, D.A. 2004. Ethnic boundaries and personal choice. Assessing the influence of individual inclinations to choose intra-ethnic relationships on pupils' networks. *Social Networks*, 26: 55-74.

Bell, R.R. 1981. *Worlds of friendship*. Beverly Hills: Sage.

Borgatti, S.P., Everett, M.G., and Freeman, L.C. 2002. *Ucinet for Windows: Software for Social Network Analysis*. Harvard, MA: Analytic Technologies.

Brewer, D.D. and Webster, C.M. 1999. Forgetting of friends and its effects on measuring friendship networks. *Social Networks*, 21: 361-373.

Burris, V. 2005. Interlocking directorates and political cohesion among corporate elites. *American Journal of Sociology*, 111: 249-283.

Carley, K.M. and Krackhardt, D. 1996. Cognitive inconsistencies and non-symmetric friendship. *Social Networks*, 18: 1-27.

Clotfelter, C.T. 2002. Interracial contact in high school extracurricular activities. *The Urban Review*, 34: 25-45.

Common Data Set Initiative. 2005. http://www.commondataset.org/. Accessed 1 June 2006.

Crosnoe, R. and Needham, B. 2004. Holism, contextual variability, and the study of friendships in adolescent development. *Child Development*, 75: 264-279.

Feld, S. and Carter, W.C. 1998. Foci of activity as changing contexts for friendship. In Adams, R.G., & Allan, G., Eds. *Placing Friendship in Context*. Cambridge: Cambridge University Press.

Hallinan, M.T. and Williams, R.A. 1987. The stability of students' interracial friendships. *American Sociological Review*, 52: 653-664.

Hartup, W.W. and Stevens, N. 1997. Friendship and adaptation in the life course. *Psychological Bulletin*, 121: 355-370.

Kandel, D.B. 1978. Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, 84: 427-436.

Kubitscheck, W.N. and Hallinan, M.T. 1998. Tracking and students' friendships. *Social Psychology Quarterly*, 61: 1-15.

Kupersmidt, J.B., DeRosier, M.E. and Patterson, C.P. 1995. Similarity as the basis for children's friendships: The roles of sociometric status, aggressive and withdrawn behavior, academic achievement and demographic characteristics. *Journal of Social and Personal Relationships*, 12: 439-452.

Leenders, R.Th.A.J. 1996. Evolution of friendship and best friendship choices. *Journal of Mathematical Sociology*, 21: 133-148.

Levin, S., van Laar, C., and Sidanius, J. 2002. The effects of ingroup and outgroup friendships on ethnic attitudes in college: A longitudinal study. *Group Processes and Intergroup Relations*, 6: 76-92.

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

Moody, J. 2001. Race, school integration, and friendship segregation in America. *American Journal of Sociology*, 107: 679-716.

Newcomb, T.M. and Wilson, E.K., Eds. 1966. *College Peer Groups*. Chicago: Aldine Publishing Company.

Newshel, A. and Author. 2004. Describing Arbor: Report to the college. Presented December 2004.

Portnova, A., Lock, P.F., Ladd, B.C., and Zimmerman, C. 2006. Another Hundred Days: Social contacts in a senior class. *Connections*, 27: 49-57.

Quillian, L. and Campbell, M.E. 2002. Beyond black and white: The present and future of multiracial friendship segregation. *American Sociological Review*, 68: 540-557.

Salzinger, S., Antrobus, J., and Hammer, M., Eds. 1988. *Social Networks of Children, Adolescents, and College Students*. Hillsdale, N.J.: Lawrence Erlbaum Associates Publishers.

Stearns, E., Bonneau, K., and Buchmann, C. 2004. Interracial friendship networks in the transition from high school to college. Paper presented at American Sociological Association 2004 Annual Meeting, San Francisco.

Consent and Confidentiality:
Exploring Ethical Issues in Public Health Social Network Research

Jenine K. Harris
Department of Community Health, Saint Louis University
St. Louis, Missouri

Abstract

Current ethical regulations were necessarily developed in response to unethical treatment of human subjects by clinical and social researchers in settings ranging from Nazi concentration camps in the 1940s to U.S. government offices in the 1960s. Due to a focus on relationships, social network studies pose complex ethical dilemmas regarding consent and confidentiality that often challenge these ethical regulations. These issues have kept social network projects from receiving Institutional Review Board (IRB) approval, and, in the case of Virginia Commonwealth University, halted human subjects research university-wide. In public health, social network analysis is an effective method for understanding how diseases are transmitted, how health messages are spread, how social support impacts morbidity and mortality, and how public health organizations collaborate. A review of 50 public health articles using social network approaches showed that few authors discussed issues of consent and confidentiality. Without accessible examples of how others have addressed consent and confidentiality, these issues will continue to challenge public health social network researchers and their IRBs.

Keywords: social network analysis, public health, ethics, consent, confidentiality

Jenine Harris is an Assistant Professor of Community Health in biostatistics at the Saint Louis University School of Public Health. She is also Assistant Director of Research for the Center for Tobacco Policy Research at Washington University's George Warren Brown School of Social Work. Her research interests include the use of organizational network analysis to evaluate public health programs and systems approaches to tobacco control.

For correspondence, please contact Jenine K. Harris, Ph.D. at the Department of Community Health, Saint Louis University, School of Public Health, 3545 Lafayette Ave., Suite 300, St. Louis, MO 63104; Email: harrisjk@slu.edu.

Introduction

Social network analysis (SNA) is a set of theories and methods widely used to examine relationships in fields like sociology, business, and public health. For public health researchers, using social network analysis is a uniquely effective method for understanding how diseases are transmitted, how health messages are spread, how social support can impact morbidity and mortality, and how public health organizations collaborate (Luke & Harris, 2007). However, researchers using this tool often face complex ethical dilemmas when designing and conducting social network research with human subjects. Specifically, obtaining consent and maintaining confidentiality pose challenges in public health social network research.

Although the ethical challenges posed by social network research are well-known among social network researchers, there is little discussion in the published network literature of how these issues impact public health network research projects and how they are addressed. Two things result from this limited visibility: 1) public health social network researchers have not had the opportunity to learn from each other in developing strategies for addressing consent and confidentiality; and 2) Institutional Review Boards (IRB) have not had the opportunity to learn about ways these issues have been addressed. As a result, network researchers may not be using the most effective or efficient strategies for conducting ethically sound network research, and, even more seriously, some network research projects may never get off the ground due to lack of IRB approval.

This paper will cover three topics: 1) the origins and content of current ethical regulations for human subjects research, 2) a brief introduction to social network research methods and its applications in public health, and 3) an analysis of how ethical issues have been covered in published public health social network research.

Ethical Regulations for Human Subjects Research

The development of current human subjects regulations began just over 60 years ago in response to the unethical treatment of research participants by physicians in Nazi Germany. Following their involvement in medical experiments ranging from injecting children's eyes with chemicals to freezing people to death (Spitz, 2005), many Nazi physicians were charged with war crimes and crimes against humanity and put on trial in Nuremberg, Germany. The 1947 verdict of one of the Nuremberg Trials, the Doctors' Trial, contained ten points describing ethically sound medical research. These points are known as the Nuremberg Code and became part of international law and the basis of ethical human subjects research (Nuremberg Military Tribunal, 1996). In 1953 in response to the trials and the Nuremberg Code, the World Medical Association (WMA) began drafting the Declaration of Helsinki, another document designed as guidance for conducting ethically sound medical research. The Declaration was adopted in 1964 and remains the international standard for ethically sound medical research (Blackmer & Haddad, 2005; World Medical Association, 2007).

Even after the Code and Declaration were developed and distributed, United States physicians continued to conduct medical experiments showing little regard for their study participants. In his 1966 article, "Ethics and Clinical Research," Harvard professor Henry Beecher identified 22 studies (Beecher, 1966) with ethical deviations ranging from withholding penicillin from service men with rheumatic fever to inducing hepatitis in children at an institute for “mentally defective children.” In the opening comments of his paper Beecher suggested the problem of unethical medical research was widespread. It appears he was right, as one of the most notorious examples of unethical conduct in medical research, the Tuskegee Syphilis Study, did not appear on his list. The Tuskegee Study was conducted from 1942 to 1972 by the United States Public Health Service (USPHS). In this study, 399 African-American men were denied available syphilis treatment in order to study the natural history of syphilis in African-Americans (Gamble, 2001). The study rationale was withheld from participants, who were uneducated and low-income. The men were kept from obtaining treatment even when drafted into the military and offered penicillin (Jones, 1993). The study ended in the early 1970s after embarrassing publicity for the USPHS. Sadly, this was only after 28 of the men died of syphilis, 100 died of related causes, 40 of their wives had contracted syphilis, and 19 of their children had been born with congenital syphilis (Jones, 1993).

While the syphilis study and the 22 studies identified by Beecher were primarily clinical research, social scientists were also involved in questionable ethical treatment of human subjects during this time. Two commonly discussed social science studies that challenged ethical boundaries were the Milgram obedience experiments in 1963 (Milgram, 1977), and the Tearoom Trade sex study in the early 1960s (Warwick, 1973; Humphreys, 1970). The Milgram experiments tested how far people would go in obeying an authoritative figure when asked to administer electric shocks to another person (Milgram, 1974). In the Tearoom Trade study, sociologist Laud Humphreys posed as a lookout for men meeting other men for anonymous sex in public restrooms (dubbed ‘tearooms’) (Warwick, 1973; Humphreys, 1970) and made note of their license plate numbers, later using the information to locate and survey the men he had observed. While not inflicting the same sorts of physical harm as the clinical studies, Milgram and Humphreys, along with other social scientists, employed deceptive techniques which had the potential to inflict mental and/or social harm on unwilling participants.

These studies challenged several of the principles outlined in the Nuremberg Code and the Declaration of Helsinki. For example, the Nuremberg code begins with the statement, “The voluntary consent of the human subject is absolutely essential,” and also specifies that, “the experiment should be so conducted as to avoid all necessary physical and mental injury” (Nuremberg Military Tribunal, 1996). Neither Milgram nor Humphreys obtained consent, and, arguably, Milgram’s study may have inflicted unnecessary mental injury on participants. However, the Code and Declaration served as ethical norms, not legal documents in the United States. As such, there were no specific risks or legal consequences for the researchers if found not following the ethical norms in the Code and Declaration. Following the publicity surrounding the Tuskegee study, Congress appointed the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Five years later, in 1979, the Commission presented the Belmont Report, a document designed to guide researchers in ethical conduct.

The Belmont Report describes three fundamental principles to guide all research involving human participants (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979): 1) Respect for Persons – Individuals should be treated as autonomous agents, and persons with diminished autonomy are entitled to protection; 2) Beneficence – Make effort to secure the well-being of participants by doing no harm, maximizing benefits, and minimizing possible harms; and 3) Justice – Equals ought to be treated equally. To ensure researchers employed these principles, the Belmont Report was codified into The Code of Federal Regulations (45CFR46). This section of the Code is often referred to as the Common Rule. In the Common Rule, the three principles of respect for persons, beneficence, and justice were applied through informed consent, assessment of risk/benefit, and selection of research subjects.

As part of federal regulations, the Common Rule is subject to enforcement. The federal government polices its own research and other entities are policed by Institutional Review Boards. Both the FDA and the Office of Human Research Protection (OHRP) conduct inquiries and investigations into reports of noncompliance. If an investigator or an IRB is found noncompliant, the OHRP may take measures including: suspending or revoking approval of an institution's Assurance of Compliance,[1] suspending institutions or investigators from participating in specific projects, and/or requiring that peer groups be notified of an institution's or investigator's past noncompliance.

Many types of research pose challenges for researchers in adequately adhering to the Common Rule. For example, clinical researchers conducting studies that could potentially inflict physical harm, such as testing a new vaccine, must put many safe-guards in place and provide evidence that they are making the maximum effort to secure the well-being of participants before they are granted approval to proceed. In social science research, adhering to the Common Rule often means minimizing potential harm that could occur if sensitive information were made public. For example, failing to keep an individual's HIV status confidential may cause them harm. Because of its' unique characteristics, social network research poses unique challenges for researchers in adhering to the Common Rule. The following section discusses applications of social network analysis in public health and the specific issues that arise in addressing the Common Rule in social network research.

Social Network Analysis

Network analysis is a set of theoretical, graphical, and statistical methods for examining relationships. It has roots that are centuries old and draw on mathematics, sociology, anthropology, and a number of other fields. Recently, social network analysis has solidified its place in popular culture through New York Times articles such as the web of who-thanks-whom at the Oscars award show (Cox & Duenes, 2007), pervasive new social networking websites like Facebook, and best-selling books like Malcolm Gladwell's *The Tipping Point* (Gladwell, 2000). Charles Kadushin probably said it best in his 2005 article, "the success of social network research has led to expectations that in addition to academic research, social network research can introduce people to one another, solve organizational problems, map the

[1] The Assurance of Compliance is necessary in order to receive federal funding, so suspension or revocation is a serious matter for investigators and institutions.

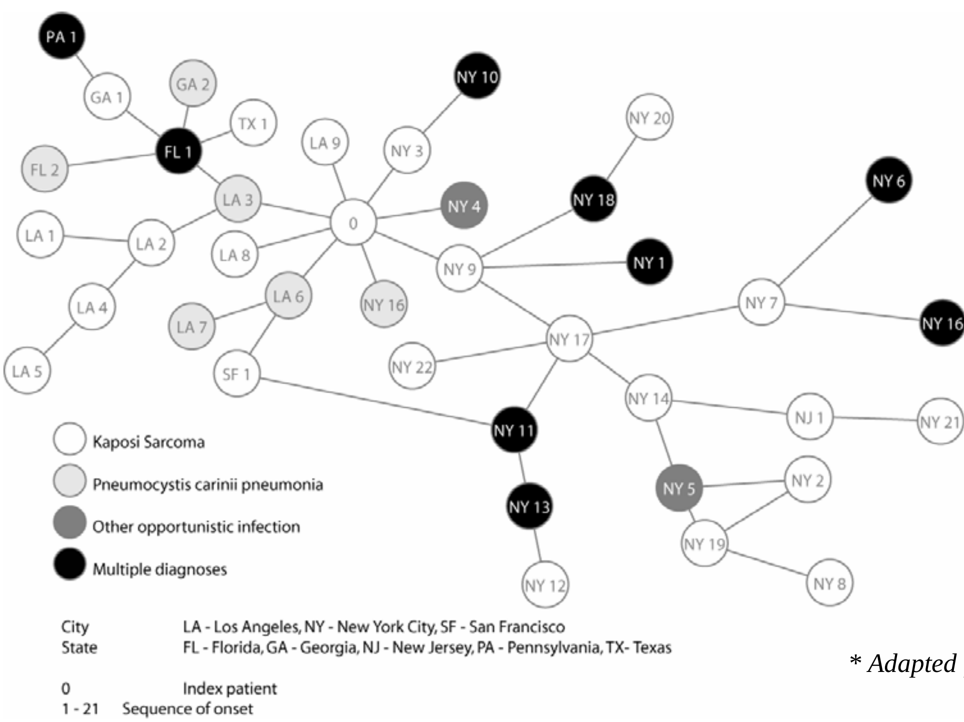
epidemiology of AIDS, and catch criminal terrorists" (p. 139). To accomplish these sorts of feats, social network analysis takes the focus off individual attributes and puts it instead on relationships such as who-talks-to-whom and who-sleeps-with-whom. Data collection, analysis, and reporting all take this unique relational perspective into consideration (Luke & Harris, 2007). It is this perspective that not only makes social network uniquely useful, but also ethically problematic.

Social Network Analysis in Public Health

In public health, SNA is used to examine three types of networks: 1) transmission networks; 2) social networks; and 3) organizational networks (Luke & Harris, 2007). Transmission networks carry a tangible entity such as infection or information. Social networks show the structure of social ties and can help determine how the ties impact health and health behaviors. Organizational networks allow researchers and practitioners to better understand agencies and organizations working on public health issues.

While disease transmission has been depicted in network formats since the 1940s (Burnet & White, 1972), they really became part of the modern arsenal of tools for understanding the spread of disease in the early days of the AIDS epidemic. Before researchers even understood exactly what AIDS was and how it was transmitted, one team of researchers collected the names of sexual partners from a number of individuals infected with this new disease (Auerbach, Darrow, Jaffe, & Curran, 1984). Through this method, they were able to connect 40 infected men in 10 cities to a single individual, patient 0 (Figure 1). The resulting network was among the first evidence that AIDS was sexually transmitted. Since then, network analysis has been used to learn more about HIV/AIDS transmission as well as Chlamydia, Gonorrhea, Syphilis, tuberculosis, and other infectious diseases.

Figure 1. Network Showing Sexual Connections Among 40 of the First AIDS Patients*

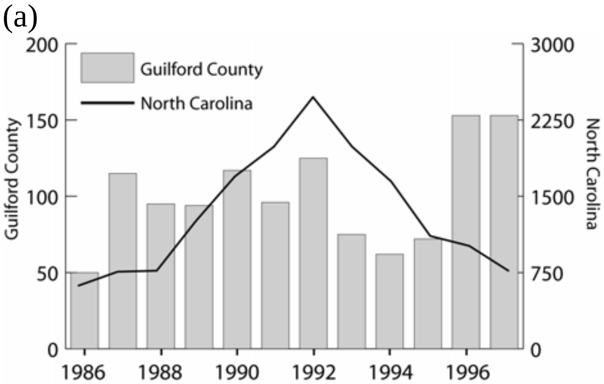


** Adapted from Auerbach et al., 1984*

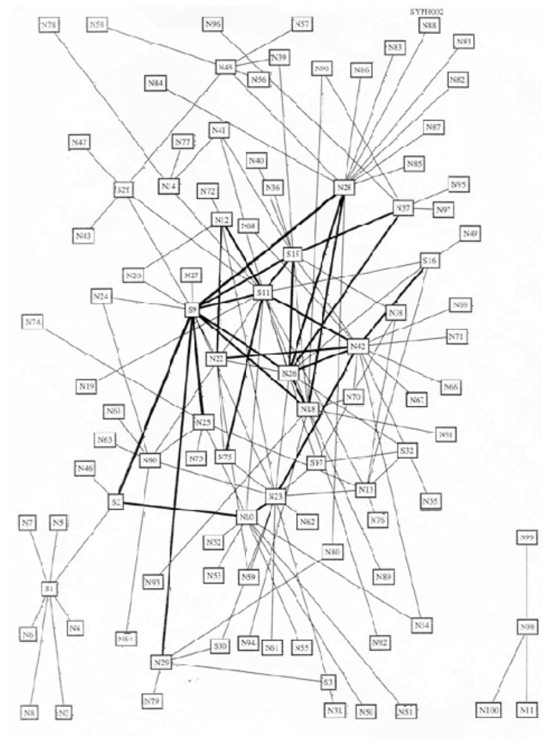
Epidemiologic studies of disease outbreaks typically follow the number of cases over time, while network models of disease transmission show how the relationships among individuals facilitate the spread of disease; Figure 2 shows this distinction (Luke & Harris, 2007). The figure shows an epidemiologic model of syphilis transmission (Figure 2a), and a network model of syphilis transmission (Figure 2b). While both contain useful information, the network model allows public health practitioners to understand exactly how syphilis is being transmitted from person-to-person in this population. This information may be useful in developing appropriate interventions.

Figure 2. A Comparison of (a) an Epidemiologic Model of a Syphilis Transmission (CDC, 1998) and (b) a Network Model of Syphilis Transmission (Frontline, 1999)*

(a)



(b)



* This figure was printed in Luke & Harris, 2007

In addition to understanding the spread of disease, public health network researchers have used network methods to understand the spread of health information and interventions. For example, network studies of how family planning and reproductive health information spread through communities revealed that both the composition of an individuals' network and having a direct link to a source of the information led to increased knowledge about family planning or increased use of contraceptives (Boulay, Storey, & Sood, 2002; Stoebenau & Valente, 2003; Valente & Saba, 2001). This type of information is useful to public health researchers and practitioners in their efforts to educate people about health issues.

Another type of public health network research utilizes information about participants' social networks to understand how the size and composition of these networks impact health and health behavior. One of the major findings in this research area is that having a large social network improves health and reduces mortality (Barber & Crisp, 1995; Bland, Krogh, Winkelstein, & Trevisan, 1991; House, Robbins, & Metzner, 1982). In addition, social network size and composition have been linked to health behaviors such as adolescent smoking (Ennett & Bauman, 1993; Valente, Unger, & Johnson, 2005), condom use (Bettinger, Adler, Curriero, & Ellen, 2004), and health screening (Allen, Sorensen, Stoddard, Peterson, & Colditz, 1999).

Finally, a relatively new area of public health network research involves public health organizations. This type of network research typically examines collaboration among public health agencies with the goal of understanding how these systems work. Public health organizational network researchers have examined systems of organizations working to address areas such as HIV/AIDS (Kwait, Valente, & Celentano, 2001), mental health services (Nakao, Milazzo-Sayre, Rosenstein, & Manderscheid, 1986), and tobacco use prevention (Krauss, Mueller, & Luke, 2004).

In all, using a social network approach to public health problems has given unique insights into disease, health behavior, and the structure of public health systems. However, along with the advances for the field of public health and benefits for the public come unique ethical considerations.

Ethical Considerations in Public Health Social Network Research

Kadushin (2005) states in the introduction to his paper on who benefits from social network analysis, "[t]he ethical issues [in social network research] are both straightforward and complex" (p. 140). In addition to all of the ethical issues that come with social science in general, social network research adds two wrinkles: 1) the collection of names is critical, and 2) the collection of names of people outside the study is common and often necessary to answer specific research questions (Kadushin, 2005; Borgatti & Molina, 2003). The first article to address these and other ethical issues specifically pertaining to social network research was published in the Journal of Applied Behavioral Science (Borgatti & Molina, 2003). In this article, Borgatti and Molina focused on *organizational* network analysis and on the ethical problems that arise in academic settings and management settings. In doing so they highlighted several differences between traditional research and network research: 1) anonymity is impossible in network data collection; 2) missing data is problematic; 3) non-participation by a subject does not mean they will be excluded from analyses; and 4) in conventional studies research participants report only on themselves, while in network studies participants report on themselves and on others.

Two years later the journal *Social Networks* published a special issue on ethical considerations in social network analysis. In his contribution to the special issue, Klovdahl (2005) described a number of assumptions implicit in most public health social network research:

1) No surgical, pharmaceutical, or other medical treatment would be provided (or withheld);
2) The research usually would be based on personal interviews with primary participants;
3) Effective means for protecting the confidentiality of the research

data – including the necessary hardware, software, and data handling protocols – would be in place and would be used;
4) Data would be ‘de-identified’ at the earliest date possible;
5) No identifying information would be shared outside the project without IRB approval for any proposed sharing; and
6) No data retained beyond the end of a project would contain information permitting the identification of any participant or network associate.

However, even if these characteristics described all public health network research, which is a tenuous assumption, the burden would still rest with investigators to develop ethically sound research methods that IRBs could confidently approve (Klovdahl, 2005). As such, Klovdahl (2005) identifies several issues that social network researchers should take into account when developing research projects, including: protecting confidentiality, identifying and applying appropriate waivers of consent, and balancing benefits against risks.

Consent in Social Network Research

The seriousness of issues of consent for network researchers was highlighted by a recent controversy at Virginia Commonwealth University (VCU) (Klovdahl, 2005). The case involved a woman in a twin study who was mailed a survey that included questions about the health of her family members (Botkin, 2001). The woman’s father read the questionnaire and was disturbed by two questions asking about abnormal genitalia and depression in male family members. He proceeded to contact National Institutes of Health OHRP, who ruled that the IRB reviewing the study did not adequately consider whether family members were also research subjects. After further review, the OHRP and the FDA suspended human subjects research entirely at VCU. Because network analytic research is based entirely on questions (sometimes on less sensitive topics) like those in the VCU study, rulings like this are problematic for network researchers.

Applying appropriate consent procedures may be more difficult in network research since determining who qualifies as a human subject may be more complicated than in most research designs (Klovdahl, 2005; Borgatti & Molina, 2003). According to the Common Rule, a human subject either has interaction with the investigator, or has private identifiable information included in the study. Since many individuals named in network studies will not interact with the investigator, and since some studies do not collect private identifiable information (meaning information that can reasonably be expected to not be observed, recorded, or made public), secondary subjects would not be considered human subjects in these studies (Klovdahl, 2005). However, many studies do collect or use information about secondary participants that might be considered private. The National Human Research Protections Advisory Committee (NHRPAC) has made recommendations to the OHRP regarding secondary participants; however, the OHRP has not currently adopted a specific policy. The NHRPAC recommends the investigator and IRB consider the following:

1. The quantity of the information collected about the secondary participant;
2. The nature of the information collected, including the sensitivity of the information and the possibility

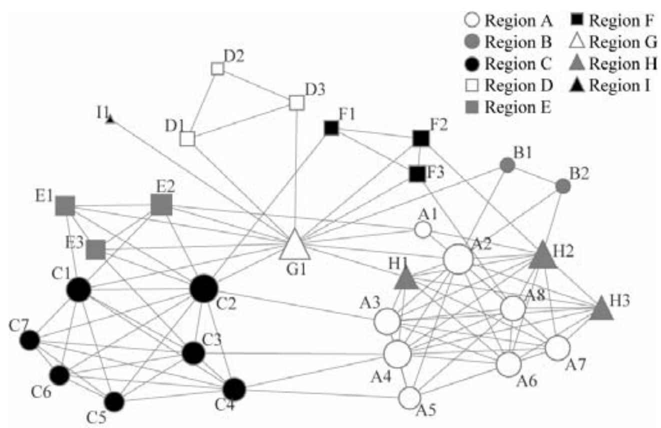
that it might cause harm to the secondary participant;
3. The ability of investigators to record information on secondary participants in a manner that protects their identity; and
4. The possibility that classification of a secondary participant as a human subject may impact the rights or welfare of the originally designated human subject requires the IRB to protect the interests of both the original human subject and the secondary subject (NHRPAC, 2002).

Should examination of these characteristics reveal that the secondary participant is identifiable (and therefore a human subject), the secondary participant would then have the rights and protections of the Common Rule, including confidentiality and consent. This could bring much of social network research to a halt since requiring consent from all named subjects would make many network studies simply infeasible (Klovdahl, 2005).

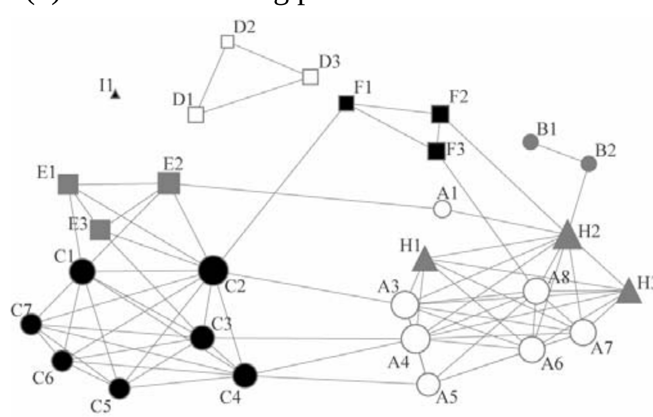
In addition to the collection of information about people who have not given consent, there is also the ethical issue of misrepresenting the "true" network should those who have not consented be removed from the data set (Borgatti & Molina, 2003). For example, consider the networks in Figure 3. The nodes in this network represent the public health emergency planners in Missouri, and the links between them represent regular communication. Network 3a (top) is complete, showing all Missouri planners; network 3b (bottom) is missing planners G1 and A2 (Harris & Clements, 2007). By removing these two nodes we could draw completely different conclusions about the communication structure of Missouri planners. In the complete network (3a), the planners in regions D and I (D1, D2, D3, I1) are part of the communication network, while in network 3b, the planners in regions D and I are not "in the loop" at all.

Figure 3. The Problem with Missing Data in Social Network Analysis*

(a) All planners represented in the network



(b) Network missing planners G1 and A2.



*Adapted from Harris & Clements, 2007

Confidentiality in Social Network Research

In protecting confidentiality, a number of things should be considered: how sensitive is the data, how practical is maintaining confidentiality, and how valuable is the data to outside individuals (Klovdahl, 2005). In addition, there are situations and topics that may be considered more sensitive than others and which may require more attention to confidentiality. First, studies about illicit activities, such as intravenous drug use, may require more attention to confidentiality than studies about everyday conversational contacts. Second, data gathered for social networks studies in organizations is often highly sensitive as the people involved may be risking their careers by giving management certain information (Borgatti & Molina, 2003). Third, information regarding secondary participants may be especially sensitive if the secondary participant is not involved in the study other than being named by a primary participant.

<u>*Weighing Risks and Benefits in Social Network Research*</u>

Borgatti and Molina (2005) describe the risk that comes with most survey research, including most social network research, as being limited to embarrassment resulting from breeches of confidentiality and discomfort from being asked sensitive questions. Borgatti and Molina (2005) and Kadushin (2005) agree that the researcher and organization typically benefit from social network research, but that the participants often do not, "…academic researchers always benefit, organizations, society and science may benefit, but individual respondents rarely do" (p. 139).

<u>*Recommendations for Social Network Research*</u>

The authors of the five articles on ethics in social network research provided a number of practical recommendations for future network research:

1) Recommendations regarding confidentiality:

- Use someone outside the research team to hold the only codebook linking names to ID numbers. This person could even be located outside the country if litigation is a potential issue (Borgatti & Molina, 2005).

- If the data being collected includes sensitive topics that could be the basis of prosecution, the researcher may obtain a Federal Certificate of Confidentiality, which states that the benefit of the research outweighs the prosecution of illegal activities by the research participants (Klovdahl, 2005).
- Segment the instruments for data collection to keep identifying information separate from other information (Klovdahl, 2005).
- Restrict the number of project personnel who have access to identifying/linking information (Klovdahl, 2005).
- Use the most secure computers available to assign network members unique non-linkable identifiers (Klovdahl, 2005).
- Do not connect the computers used for processing the raw data (with identifying information) to any network (Klovdahl, 2005).
- Never transfer files including raw data over the internet and never transport encrypted data and passwords together (Klovdahl, 2005).
- Lock-up storage media containing raw data and store back-ups securely (Klovdahl, 2005).
- Destroy any identifying information at the earliest possible date (Klovdahl, 2005).

- Train all project personnel in confidentiality protection (Klovdahl, 2005).
- Anonymize or aggregate data to the group level prior to giving the data to management in organizational studies (Borgatti & Molina, 2005).
- Write up an agreement between the researcher and management that indicates (a) what data (and in what form) management will see, and (b) how the network data and analysis will be used by the organization (Borgatti & Molina, 2005).

2) Recommendations regarding consent:

- Develop and implement a thorough and explicit consent form. This may include writing in exactly who will see what data and potentially asking management to sign a disclosure contract prior to data collection (Borgatti & Molina, 2005).
- Provide organization members with the option to exclude themselves from the study as a whole (Borgatti & Molina, 2005).
- When possible, researchers should solicit participation themselves rather than receive help from management, which can be seen as an indirect order (Borgatti & Molina, 2005).
- Offer participants *Truly Informed Consent*, meaning that participants see the management disclosure contract and are given an example of the kinds of outputs management will see. In addition, we suggest that the researcher also sign the consent form to reinforce the view that it constitutes a contract between the researcher and the respondent (Borgatti & Molina, 2005).
- Ensure that participation is truly voluntary, especially in managerial settings (Borgatti & Molina, 2003).
- Educate the IRB about name-generating questions and the value of network research (Borgatti & Molina, 2005).

3) Recommendation regarding benefits:

- After the data are collected, provide the participants with specific personalized feedback, including information they might use to improve their personal networks or that might be useful in their employment (Borgatti & Molina, 2005).

An additional overall recommendation was provided by Goolsby (2005), who said that, since codes of ethics are often outdated because they were created in response to historical events and have not been reconsidered, social scientists should work together to develop an "ethical imagination" that will move social science forward to meet the needs of the funding agencies, researchers, participants, and society.

Given the unique ethical considerations facing public health social network researchers, it appears that specific examples of effective ethically sound network studies that have gained IRB approval are needed. One way to make such examples available and accessible to both researchers and IRBs is to include information about how consent and confidentiality were addressed in published network research. The final section of this paper examines the coverage of ethical issues in published public health social network research.

Coverage of Consent and Confidentiality in Public Health Social Network Research

The ethical issues of consent and confidentiality pose dilemmas in social network research different from those faced in research not utilizing relational data. Although public health social network researchers and their IRBs are in need of examples of how these issues have been successfully addressed, published social network research, like most published social science research, does not typically include much discussion of the ethical decisions made in designing and carrying out studies. This section examines whether and how investigators conducting public health social network research included discussions of consent and confidentiality in their published research. To be clear, coverage (or lack of coverage) of these topics in an article does not imply that researchers have or have not used ethical practices in their research.

Methods

To examine how public health social network researchers have addressed consent and confidentiality, the author reviewed fifty public health social network research studies. The studies were published between 1984 and 2005 in 34 different journals and covered the three areas of network research found in public health: 1) transmission networks; 2) social networks; and 3) organizational networks.[2] The articles were selected from the bibliography of a recent review of social network analysis in public health (Luke & Harris, 2007). Articles were selected that were 1) empirical, 2) took a network approach, and 3) represented the variety of network approaches and topics that exist in public health social network research. A full list of the articles reviewed is available from the author.

To determine how and how often consent and confidentiality were discussed in public health social network research, basic information was collected on each article including: publication year, publication journal, author, and title. In addition, each article was coded for: article topic, data source, data type, vulnerable populations, discussion of consent, and discussion of confidentiality. The rationale for including variables such as data source, data type, article topic, and vulnerable populations was to assess whether articles including sensitive topics and populations were more likely to include discussion of consent or confidentiality.

Results and Discussion

Of the 50 studies reviewed, 36 (72%) used name-generation data. Name generation prompts ranged from, “Name up to six best friends” (Pearson & West, 2003) to having participants name their social network and, “[Specify] their age, HIV status, whether they were living or had died of AIDS, and whether they had ever been a sex partner” (Morris, Zavisca, & Dean, 1995). Seven studies included questions about needle-sharing or other aspects of intravenous drug use, and 18 articles were about HIV/AIDS or other sexually transmitted diseases. Over a third of the articles included members of vulnerable populations; the main subjects were children or youth in 13 articles. Sixty percent of the studies included primary data; the 40% of studies based on secondary data used regional and national data sets such as the National Longitudinal Study of Adolescent Health (National Institute of Child Health and Human Development, 2006).

[2] See above for definitions and examples of each.

Consent was discussed in 18% of the articles and confidentiality was discussed in 24% of the articles. Compared to the rates in the entire sample, studies with vulnerable populations were more likely to include discussions of consent and confidentiality, as were studies with primary data collection, studies using name data, and studies on disease transmission. Compared to the entire sample, articles on social networks (i.e., networks comprised of social relationships like social support among individuals; see above for description) were more likely to include discussions of confidentiality, but less likely to include discussions of consent.

Organizational network research articles were the least likely to discuss issues of consent and confidentiality. None of the 10 organizational network articles discussed consent and one discussed confidentiality. Table 1 shows additional characteristics of the articles.

Table 1. Characteristics of 50 Public Health Social Network Research Articles

Article Topic			n	%	Discussed Consent		Discussed Confidentiality	
Transmission Networks			**21**	**42%**	**7**	**33%**	**4**	**19%**
	Disease		20	40%	6	30%	3	15%
		HIV/AIDS	9	18%	2	22%	1	11%
		STD (non-HIV/AIDS)	9	18%	2	22%	1	11%
		Other infectious disease	2	4%	1	50%	0	0%
	Information		1	2%	1	100%	1	100%
Social Networks			**19**	**38%**	**3**	**16%**	**8**	**42%**
	Health behavior		7	14%	2	29%	6	86%
	Social support/Social capital		12	24%	1	8%	2	16%
Organizational Networks			**10**	**20%**	**0**	**0%**	**1**	**10%**
	Public health systems		10	20%	0	0%	1	10%
Other characteristics								
Vulnerable populations			**18**	**36%**	**4**	**22%**	**9**	**50%**
	Children/Youth		13	26%	4	31%	9	69%
	Low SES		3	6%	0	0%	0	0%
	Mentally ill		2	4%	0	0%	0	0%
Data Source								
	Primary data collection		30	60%	7	23%	9	30%
	Secondary data analysis*		20	40%	2	10%	3	15%
Data Type								
	Name data		36	72%	9	25%	11	31%
	Other		14	28%	0	0%	1	7%

* For the purpose of this paper secondary data analysis is defined as: *The analysis of data collected by someone else, perhaps for some purpose other than that of subsequent analyses* (Babbie, 1983).

Consent and/or confidentiality issues appeared in 17 out of the 50 articles; however, the discussions were generally brief. Statements regarding consent consisted of either the parent or the participant giving informed consent for participation, for example:

> "There were 2,002 eligible students (those with parental consent and student assent) who completed a baseline survey" (Valente et al., 2005).

> "Because of ethical concerns about participants' disclosing their drug use prior to informed consent, the screening did not include questions about individuals' own risk behaviors. …Potentially eligible individuals were asked to come to the clinic to provide informed consent, approved by the Johns Hopkins School of Public Health's Institutional Review Board, and complete a face-to-face baseline interview" (Latkin, Sherman, & Knowlton, 2003).

One study out of the 50 also discussed obtaining consent from the secondary participants named by the primary subjects: "Written, informed consent was obtained from each patient or his sexual contact before interviews were conducted" (Auerbach et al., 1984).

The 12 studies that provided information on confidentiality included either a description of the level of privacy granted to the participant during the survey/interview process, or the process of assigning identification numbers for confidentiality of participant data. For example:

> "An additional advantage of this data involves the use of laptop computers to maintain confidentiality about sensitive subjects such as delinquency. This method of data collection allowed respondents to maintain their anonymity by listening to pre-recorded questions about participation in different delinquent activities and then entering their responses directly into a computer" (Haynie, 2001).

> "The surveys were identified only by a code number, not with the students' names or any other identifying information" (Mouttapa , Valente, Gallaher, Rohrbach, & Unger, 2004).

One study provided a comprehensive description of their process for maintaining confidentiality:

A high priority was given to ensuring the confidentiality and security of the data. An encoding scheme was developed to protect the identity of all respondents. Personal computers were used for data entry and most processing. Removable hard disks were purchased for data storage and then locked away when not in use. The personal computers were not part of any network. The database design involved segmentation of information and required encrypted files to be brought together (physically) to access sensitive data. Only "sanitized" data (no identifying information) were processed on mainframe (or networked) computers (Klovdahl et al., 1994).

Overall there was limited discussion of consent and confidentiality in this set of articles. Again, that is not to say that the researchers did or did not make ethical choices and follow ethical practices, just that they did not include descriptions of these in their publications.

Conclusions

Despite expectations that social networking could be used to catch terrorists, cure HIV, and introduce you to your partner

(Kadushin, 2005), social network research typically does not claim to solve all the problems of the world or single-handedly prevent all future disease (Klovdahl, 2005). However, social network research does contribute valuable information to many fields, including public health. Along with those contributions come complex ethical decisions regarding, among other things, consent and confidentiality. Without accessible examples of how others have addressed consent and confidentiality, these decisions will continue to challenge public health social network researchers and their IRBs. Including discussions of specifically how consent and confidentiality were addressed in public health social network research publications could ease the burden on future social network researchers in designing studies and gaining IRB approval.

References

Allen, J., Sorensen, G., Stoddard, A., Peterson, K., and Colditz, G. 1999. The relationship between social network characteristics and breast cancer screening practices among employed women. *Ann. Behav. Med.*, 2(13): 193-200.

Auerbach, D., Darrow, W., Jaffe, H., and Curran, J. 1984. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. *American Journal of Medicine*, 76(3): 487-97.

Babbie, E. 1983. *The Practice of Social Research*. Belmont California: Wadsworth Publishing Company.

Barber, G. and Crisp, B. 1995. Social support and prevention of relapse following treatment for alcohol abuse. *Res. Soc. Work Pract.*, 5: 283-296.

Beecher, H. 1966. Ethics and clinical research. *New England Journal of Medicine*, 274(24): 1354-60.

Bettinger, J., Adler, N., Curriero, F., and Ellen, J. 2004. Risk perceptions, condom use, and sexually transmitted diseases among adolescent females according to social network position. *Sexually Transmitted Diseases,* 31: 575-579.

Blackmer, J. and Haddad, H. 2005. The Declaration of Helsinki: an update on paragraph 30. *CMAJ,* 173(9): 1052-1053.

Bland, S., Krogh, V., Winkelstein, W., and Trevisan, M. 1991. Social network and blood pressure: a population study. *Psychosom. Med.,* 53(6): 598-607.

Borgatti, S. and Molina, J-L. 2003. Ethical and strategic issues in organizational network analysis. *Journal of Applied Behavioral Science,* 39(3): 337-349.

Borgatti, S. and Molina, J-L. 2005. Toward ethical guidelines for network research in organizations. *Social Networks,* 27: 107-117.

Botkin, J. 2001. Protecting privacy of family members in survey and pedigree research. *Journal of the American Medical Association,* 285(2): 207-211.

Boulay, M., Storey, J., and Sood, S. 2002. Indirect exposure to a family planning mass media campaign in Nepal. *Journal of Health Communication,* 7(5): 379-399.

Burnet, F. and White, D. 1972. *Natural History of Infectious Disease*. Cambridge, UK: Cambridge University Press.

Centers for Disease Control and Prevention. 1998. Outbreak of Primary and Secondary Syphilis—Guilford County, North Carolina, 1996-1997. *MMWR*, 47(49): 1070-1073.

Cox, A, and Duenes, S. 2007. *I would like to thank...* http://graphics8.nytimes.com/images/2007/02/26/movies/0227-cul-webTHANK.jpg. Accessed 18 Mar 2007.

Ennett, S, and Bauman, K. 1993. The contribution of influence and selection to adolescent peer group homogeneity: the case of adolescent cigarette smoking. *J. Person. Soc. Psychol.*, 67(4): 653-663.

Frontline. 1999 *The Lost Children of Rockdale County.* http://www.pbs.org/wgbh/pages/frontline/shows/georgia/. Accessed 15 Mar 2007.

Gamble, V. 2001. Under the shadow of Tuskegee: African Americans and health care. In Teays, W, and Purdy, L, (Eds.): *Bioethics, justice, and health care*. Belmont, California: Wadsworth.

Gladwell, M. 2000. *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Little, Brown and Company.

Goolsby, R. 2005. Ethics and defense agency funding: some considerations. *Social Networks,* 27(2): 95-106.

Harris, J. and Clements, B. 2007. Using social network analysis to understand Missouri's system of public health emergency planners. *Public Health Reports*, 122(4): 488-498.

Haynie, D. 2001. Delinquent peers revisited: does network structure matter? *American Journal of Sociology,* 106: 1013-1057.

House, J., Robbins, C., and Metzner, H. 1982. The association of social relationships and activities with mortality: prospective evidence from the Tecumseh Community Health Study. *American Journal of Epidemiology*, 116(1): 123-140.

Humphreys, L. 1970. *Tearoom trade: Impersonal sex in public places*. Chicago: Aldine Publishing Company.

Jones, J. 1993. *Bad Blood: The Tuskegee Syphilis Experiment*. New York: Free Press.

Kadushin, C. 2005. Who benefits from network analysis: ethics of social network research. *Social Networks,* 27: 139-153.

Klovdahl, A. 2005. Social network research and human subjects protection: towards more effective infectious disease control. *Social Networks,* 27: 119-137.

Klovdahl, A., Potterat, J.J., Woodhouse, D.E., Muth, J.B., Muth, S.Q., and Darrow, W.W. 1994. Social networks and infectious disease: the Colorado Springs study. *Social Science & Medicine,* 38: 79-88.

Krauss, M., Mueller, N., and Luke, D. 2004. Interorganizational relationships within state tobacco control networks: a social network analysis. *Preventing Chronic Disease,* 1(4): A08.

Kwait, J., Valente, T., and Celentano, D. 2001. Interorganizational relationships among HIV/AIDS service organizations in Baltimore: a network analysis. *Journal of Urban Health: Bulletin of the New York Academy of Medicine,* 78: 468-487.

Latkin, C., Sherman, S., and Knowlton, A. 2003. HIV prevention among drug users: outcome of a network-oriented peer outreach intervention. *Health Psychol.*, 22(4): 332-339.

Luke, D., and Harris, J. 2007. Network analysis in public health: history, methods, and applications. *Annual Review of Public Health,* 28: 69-93.

Milgram, S. 1974. *Obedience to authority: An experimental view*. New York: Harper and Row.

Milgram, S. 1977. Subject reaction: the neglected factor in the ethics of experimentation. *Hastings Center Report,* 7(5): 19-23.

Morris, M., Zavisca, J., and Dean, L. 1995. Social and sexual networks: their role in the spread of HIV/AIDS among young gay men. *AIDS Educ. Prev.,* 7(Suppl.5): 24-35.

Mouttapa, M., Valente, T., Gallaher, P., Rohrbach, L., and Unger, J. 2004. Social network predictors of bullying and victimization. *Adolescence*, 39(154): 315-335.

Nakao, K., Milazzo-Sayre, L., Rosenstein, M., and Manderscheid, R. 1986. Referral patterns to and from inpatient psychiatric services: a social network approach. *American Journal of Public Health,* 76(7): 755-760.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. *Belmont Report.* Washington, D.C.: U.S. Government Printing Office.

National Human Research Protections Advisory Committee. 2002. *Clarification of the status of third parties when referenced by human subjects in research.* http://www.hhs.gov/ohrp/nhrpac/documents/third.pdf. Accessed 22 Mar 2007.

National Institute of Child Health and Human Development. 2006. *National Longitudinal Study of Adolescent Health*. http://www.nichd.nih.gov/health/topics/add_health_study.cfm. Accessed 18 March 2007.

Nuremberg Military Tribunal. 1996. The Nuremberg Code. *JAMA*, 276(20): 1691.

Pearson, M. and West, P. 2003. Drifting smoke rings: social network analysis and Markov processes in a lognitudinal study of friendship groups and risk taking. *Connections*, 25: 59-76.

Spitz, V. 2005. *Doctors from Hell: The Horrific Account of Nazi Experiments on Humans*. Sentient Publications.

Stoebenau, K. and Valente, T. 2003. Using network analysis to understand community-based programs: a case study from Highland Madagascar. *Int. Fam. Plan. Perspect.*, 29: 167-173.

Valente, T. and Saba, W. 2001. Campaign exposure and interpersonal communication as factors in contraceptive use in Bolivia. *Journal of Health Communication,* 6(4): 303-322.

Valente, T., Unger, J., and Johnson, A. 2005. Do popular students smoke? The association between popularity and smoking among middle school students. *J. Adolesc. Health,* 37*:* 323-329.

Warwick, D.P. 1973. Tearoom trade: means and ends in social research. *Studies/Hastings Center,* 1(1): 27-38.

World Medical Association. *WMA History: Declaration of Helsinki.* http://www.wma.net/e/history/helsinki.htm. Accessed 18 Mar 2007.