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

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

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

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

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

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

1. Introduction

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

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

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

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

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

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

2. Email and Survey Measures of Social Networks

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

Most network surveys are distributed at one point (or a few points) in time to a predefined set of individuals. These individuals are asked questions (name generators) that elicit a list of names or are presented with a roster of individuals that they may have specific relations with. The name generator questions enable the researcher to define precisely the type of relationships that are of interest, by specifying the content (e.g., trust, friendship), the duration (e.g., in the last three months), or the boundaries (geographic, institutional) of the relationship of interest. Respondents are then asked to provide more information about the type of relationship that they have with each alter (e.g., frequency, emotional proximity, medium). In some cases, respondents are also asked to provide information about the relationships among the alters or even about the relationships among all the actors in the network (Krackhardt, 1987a). While there is a wider diversity in the type of questions that can be asked using a network survey, these steps represent the standard template for gathering social network survey data.

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

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

2.1. Continuous Versus Cross-sectional Data Collection

A typical e-mail dataset is composed of a series of events, each of them indicating that a message was sent from a given e-mail address to a set of e-mail addresses at a specific point in time. While the continuous nature of e-mail data opens many avenues for research, it is, in practice, difficult for a researcher to fully exploit. The first hurdle to be overcome is the sheer volume of e-mail data generated on a daily basis. Kleinbaum et al. (2008) analyze a sample of over 30,000 employees who collectively exchanged as many as 1.28 million e-mails in a single day; Bulkley and Van Alstyne (2007) recorded 125,000 emails among 71 employees in a recruiting firm over 10 months. The data management and manipulation skills to be gained in order to deal with such large datasets differ from those required to analyze survey datasets, even large ones. A second hurdle is the paucity of tools, models and algorithms that deal with continuous relational data. A few examples exist (e.g., Butts, 2008a for modeling; Moody, McFarland, & Bender-deMoll, 2005 for data visualization), but there is a substantial learning curve to master these tools and methods. Finally, there are even fewer theories and concepts that a researcher can use to analyze a continuous dataset and interpret the result based on a time dimension (Ancona, Goodman, Lawrence, & Tushman, 2001). As a result, researchers using e-mail data tend to aggregate the continuous information into one or a few cross-sectional datasets, which is a more familiar format to traditional social network research. Yet, this aggregation is not a straightforward process and decisions made during the aggregation process could potentially lead to very different networks (Butts, 2009; Grannis, 2010; cf. Kleinbaum, et al., 2008). To reflect this, we focus this paper on comparing a survey network with a collapsed e-mail network and highlight the potential differences between the two networks.

2.2. Observation Versus Recall

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

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

2.2.1. The density issue

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

2.2.2. The stable relationship issue

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

2.2.3. The social significance issue

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

2.3. Interactions Versus Relations

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

Yet, we know that an e-mail network is a communication network; in the intra-organizational setting, we assume that most communication is task-related. In that sense, we are expecting that the content of e-mail is constituted mainly of task-related information (Bulkley & Van Alstyne, 2007). The concept of a communication network is nevertheless quite broad: Monge and Contractor describe it as "the patterns of contact that are created by the flow of messages among communicators through time and space. The concept of message should be understood here in its broadest sense to refer to data, information, knowledge, images, symbols and any other symbolic forms," (Monge & Contractor, 2003, p. 3). Correspondingly, the social processes that are reflected by structural positions or configurations in an e-mail network may be interpreted very differently from a survey network. For example, in-degree centrality in an e-mail network (receiving e-mail from many different senders) might not so readily be interpreted as prominence or prestige as it would in a survey network (Knoke & Burt, 1983). Receiving many e-mails may be an artifact of the particular tasks a person performs for the organization, which may or may not be associated with prestigious positions; for example, administrative assistants have relatively high degree scores that are not necessarily related to their organizational prestige, precisely because their task is to coordinate the activities of others. Other concepts, evocative of information flow, are very applicable to e-mail data. For example, betweenness centrality is conceptualized as the extent to which an individual can control the flow of information in an organization (Freeman, 1979). As such, we argue that interpreting an e-mail network requires a careful interpretation of the type of concepts that the researcher is attempting to explore.

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

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

2.3.1. The dependence issue

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

2.3.2. The recipients issue

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

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

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

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

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

3. Data and Methods

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

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

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

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

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

3.1. Sample Selection

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

3.2. Comparing the Networks Using the Quadratic Assignment Procedure

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

3.3. Contrasting the Measures Using Exponential Random Graph Models

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

The general form of the model for multiple networks is:

$$\Pr(Y = y \mid X = x) = \frac{\exp\left[\sum \lambda_{\mathrm{A}} \mathrm{Z}_{\mathrm{A}}(y,x)\right]}{\kappa} \quad (1)$$

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

We analyze the e-mail and survey measures of the network using ERGMs with higher-order parameters for directed graphs (Robins, et al., 2006; Snijders, et al., 2006) applied to multiple networks (Pattison & Wasserman, 1999) using the XPnet software package (Wang, Robins, & Pattison, 2006). We chose to model two pairs of networks: Model 1 establishes a baseline and includes only explanatory variables related to local network structure: *Arc* indicates the overall propensity for two randomly-selected individuals to interact, controlling for the other parameters in the model. *Reciprocity* indicates the propensity for the interaction within a directed dyad to be reciprocated. *Survey–Email* reflects the propensity of observing an e-mail tie conditional on the presence of a survey citation or vice versa; unlike our other parameters, *Survey–Email* is jointly estimated across the two networks in each model. Additionally, we include effects that reflect potential endogenous local network structural mechanisms, including the propensity for various star-like forms (out, in and mixed stars) and triangle-like structures (transitive and acyclic; see Robins, et al., 2009).

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

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

4. Results

4.1. Communication Media

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

4.2. Descriptive Summary Statistics

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

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

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

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

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

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

4.3. QAP Correlation Analysis

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

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

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

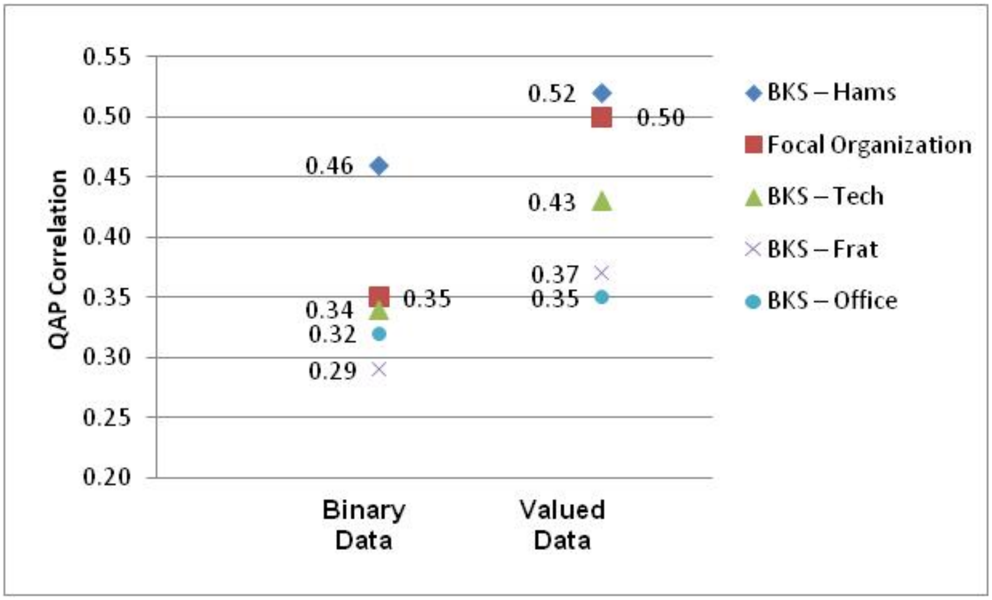

Figure 1. Between-Measure QAP Correlations
QAP correlations of valued data and binary data networks of our focal organization compared to the Bernard, Killworth and Sailor (BKS) datasets. The values represent the QAP correlation between behavioral and recall networks. BKS-Hams, BKS-Tech, BKS-Frat and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

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

4.4. Exponential Random Graph Models

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

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

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

Table 2. Exponential Random Graph Models

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

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

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

Table 3. ERGMs with Density Adjusted Email Measures

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

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

Our second core finding concerns the locus of clustering in the network. We find evidence of clustering in both the survey and the e-mail data, but the local structures that describe the clusters differ, as evidenced by the structural parameters in Model 1. In the survey measure, clustering occurs along transitive pathways: the probability of *i* nominating *j* is significantly increased when *i* nominates various third parties who also nominate *j*. In other words, triads in the survey measure tend to be closed following a balance principle: if *i* nominates others who nominate *j*, there is a strong probability that *i* will also nominate *j*. Additionally, the negative *Mix-2-Star* effect reinforces this interpretation as it suggests that transitive paths are unlikely to occur on their own (i.e. without being closed). Thus clustering in the survey data appears to be directed along transitive paths.

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

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

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

Conversely, the existence of structural effects in the e-mail measure over and above the *Email-Survey* parameter and the actor attributes means that the structural processes present in the e-mail measure are not captured well by either the survey measure or the actor attributes alone. The significant negative *Popularity Spread* parameter (-1.59) indicates relative uniformity in in-degree: the distribution of the number of senders for each recipient is homogenous across actors. By contrast, the significant positive *Activity Spread (1.36)* is a sign of heterogeneity of out-degree: some actors are observed to send e-mail to a larger number of alters than others while all actors receive email from similar numbers of alters. This result, too, is unlikely to be an artifact of data collection methods – while each actor's out-degree was limited to 10 in the survey, this constraint was rarely binding – so in practice, neither in-degree nor out-degree was constrained in either data collection method. Further, the fact that this effect only emerges as significant when we control for hierarchical level suggests that there is no main effect of level on in-degree, but that heterogeneity of in-degree occurs at each level of the hierarchy. Additionally, the significant *Mixed-2-Star* (0.08) effect suggests that some individuals do tend to play important roles in the flow of e-mail communication, by receiving and sending along messages. Finally, in terms of clustering, the introduction of actor attributes does not affect the *Popularity Closure* parameter (1.66) which is still positive and significant, but the *Transitive Closure* parameter (-0.85) is now negative and significant. This confirms the tendency shown in the survey network for hierarchical level and group affiliation to explain transitive closure.

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

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

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

4. Discussion

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

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

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

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

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

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

5.1. Implications for Research

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

Furthermore, we suggest that research that explicitly focuses on infrequent, cross-category communications or weak ties may suffer from under-reporting of those ties in surveys and should be controlled for. For example, in his classic work on the social structure of R&D labs, Allen (1977) shows that communication across intra-organizational boundaries is rare and that status places significant constraints on communication, as high-status actors rarely communicate with low-status actors. Although empirical support is generally wanting, the intuitive appeal of Allen's work has set expectations for a generation of scholars who read it. More recently, Cross and Parker (2004) report a strong hierarchy effect in network connectivity based on a survey, but they interpret the effect to be caused by information-gathering processes. Our analysis of the survey measure of our social network yields results that are consistent with prior literature, but our comparison with e-mail measures of the network suggests that these results confound – at least partially – real, socially meaningful effects with measurement error introduced by response bias.

5.2. Caveats and Limitations

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

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

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

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

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

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

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

6. Conclusions

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

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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