Validity Issues in the Use of Social Network Analysis for the Study of Online Communities

James Howison, Andrea Wiggins and Kevin Crowston

Under Second Round Revision for the Journal of the Association of Information Systems (JAIS)

Abstract

There is a natural match between studies of online communities and social network analysis (SNA). Information Systems research, in particular, has drawn heavily on the growing data sources available as a by-product of increased online social interaction, conducting social network analyses with this "trace data". However, this type of data has properties that are distinct from the data upon which researchers in Sociology and Anthropology have typically developed social network measures and their interpretations. This paper examines validity issues deriving from the use of trace data in IS studies of online communities, arguing that more attention needs to be paid to the chain of logic linking data collection, construct development, operationalization and measure interpretation. Four issues are elaborated and illustrated from the literature: 1) unexpected reliability concerns, 2) the need to argue and test the link from 'found data' to theoretical constructs, 3) validity problems arising from the stability of the constructs over time and 4) counter-intuitive validity issues deriving from the data being near-complete records of communities' activities. The paper concludes with recommendations for researchers and reviewers designed to increase the rigor and relevance of SNA research in Information Systems.

Keywords

Social Network Analysis, IS Scholarship, Networks and Communities, Distributed Work

Introduction

Social network analysis (SNA) techniques are seeing more use in information systems research (Borgatti and Foster, 2003; Forman et al., 2008) and represent an important opportunity for the field. For two reasons, these techniques seem particularly useful for the study of online activity, particularly online communities, which we define for the purposes of this discussion as groups that interact primarily or exclusively via information and communications technologies (ICT). We intentionally choose a broad interpretation of what comprises an online community due to our methodological focus and the variety of phenomena involving online interactions that are encompassed by IS research.

First, the nature of interactions in these communities is of interest to researchers for a number of reasons: for its fit to task demands (Ahuja & Carley, 1998) as a predictor of other measures, such as group cohesion (Hahn, Moon, & Zhang, 2008), trust (Ridings, Gefen, & Arinze, 2002), knowledge generation (Wasko, Faraj, & Teigland, 2004) or information diffusion (Hinz & Spann, 2008); and as a mediator of relations between other factors. Second, SNA techniques are data intensive, but usable data are readily available for online communities, in the form of “trace” data. Trace data are unobtrusive measures that directly capture the ICT-mediated interactions between individual members. Indeed, some researchers see the availability of trace data from such communities as a golden opportunity for research (Garton et al., 1997; Watts, 2007).

Applications of SNA have the potential to provide insight into many group processes in online settings, such as decision-making, organizing and innovation. However, we note that the underlying assumptions of traditional social network analysis methods have not often been examined in detail upon application to studies of online settings. This lack is reason for concern, as the available data and the kinds of structures they represent differ in key respects from the material of earlier social network studies. These differences lead to potential issues with validity, in particular with respect to the data and by association, to the whole chain of reasoning leading up to reported SNA results. Unfortunately, papers applying SNA techniques rarely theorize deeply about the nature of interactions that are represented by the links in the networks, often relying instead on the characterizations of interactions from earlier SNA work, which we will argue are not suitable for the data available from online communities.

The purpose of this paper is to analyze possible threats to validity for social network analyses of data derived from online communities to determine how and when SNA ought be applied to their study. The methods for acquiring data and for network construction, in addition to the nature of the data sources themselves, influence the way we understand measures. In addition, we suggest that the expected stability of the constructs of interest and the time dimension of the network data can affect the interpretation of SNA results. This paper begins by reviewing the literature on social network analysis and online communities, followed by a discussion of the validity issues surrounding these data and techniques, and concludes with a set of recommendations for improving the rigor of this genre of IS research studies.

SNA and online communities research

To provide a background for our discussion of reliability and validity concerns in the use of SNA for online communities research, we first briefly introduce the measurement process as it applies to the use of SNA. We then briefly review how measurement has been carried out in research using SNA, starting with its origins in sociology before moving to research in information systems and on online communities in particular. The goal of this review is to make the case that online communities data have different characteristics than those used in earlier studies, thus raising a different set of reliability and validity concerns that researchers must be careful to address.

Overview of SNA methods

In discussing the use of SNA, it is important to keep in mind that SNA is not a theory *per se*; rather, it is a set of mathematical techniques for analyzing networks. Various substantive theories focus attention on networks in different settings, thus motivating the use of such techniques, but these theories and the analysis approach are conceptually distinct. As a result, it is not correct to speak of SNA findings, any more than it would be to speak of regression findings. To assess reliability and validity, researchers need to examine the fit of this methodology to the theoretical constructs of interest and the particular kinds of data available.

Using SNA in a research study is a data reduction method that has parallels to using other quantitative techniques. For analysis, a set of relationships is represented as a mathematical structure called a graph composed of nodes and links, often encoded as a matrix. In other words, the use of SNA assumes that the network nodes and links have been measured, just as the use of conventional statistical techniques assumes that constructs of interest have been measured as series of numbers. Given a graph or interaction matrix, calculations can be made of individual-level scores for the structural position of individual nodes, such as various individual measures of network centrality, as well as of overall summaries of structural characteristics for the overall network, such as network density or centralization. The application of these techniques is conceptually similar to the statistical computation of an individual score, such as a Z-score to show an individual's relative position in a distribution, or a summary statistic such as a mean or standard deviation to summarize an entire sample.

Just as descriptive statistical analysis techniques like averaging and finding standard deviations can be applied to data representing a wide diversity of constructs, techniques of SNA can be applied to networks representing diverse kinds of nodes and links. The choice of which real-world phenomena to measure reflects the researcher's questions and theoretical approach. Nodes in the network represent the actors of interest, which are often individuals but may also be organizations, countries, web pages or even molecules, depending on the unit of analysis. Links represent a theoretically interesting connection between actors. For example, links between individuals might represent friendship, advice seeking or a transaction such as a loan; for countries, alliances or international trade; for web pages, a hyperlink; for molecules, a chemical reaction.

Networks can also include more than one kind of node; common examples are networks that include people and the organizations to which they belong. In this case, links are constructed from individuals' memberships in the organizations, forming what is called an association network. The individuals can be dropped from the resulting network to study connections between organizations due to joint membership (e.g., interlocking boards of directors Brandeis (1914); Davis et al. (2003)) or vice versa.

Finally, diverse data sources can be used to identify actors and their links. A matter of continual concern in network studies is the ability to harvest or construct complete, authoritative network structures while dealing with the internal validity issues that arise for each of the primary methods for collecting network data: questionnaires, interviews, observation and archival records, among others (Wasserman & Faust, 1994). Thus the use of SNA raises questions of measurement reliability and validity parallel to those for the use of other quantitative analysis techniques. While these questions have been answered for the constructs and data examined in earlier studies, we argue that the nature of the data from online communities is sufficiently novel to require their re-examination. It is not possible to simply cite prior work as warrant, as reliability and validity are dependent on the match between particular data elicitation approaches and theoretical constructs. Similarly, researchers cannot simply rely on findings from prior work that uses different theoretical constructs or different kinds of data.

In the remainder of this section, we present brief review of SNA research focusing on the methodological issues regarding measurement. The goal is not to summarize all social network research or even all SNA applications in information systems research, as these tasks are well beyond the scope of a single paper, but rather to outline the general trends in past research in order to highlight where online communities research using SNA differs. Subsequent sections will examine the implications of these differences in particular for reliability and validity of the measures. These implications form the basis for a set of recommendations for authors and reviewers intended to improve the quality of SNA research in online communities.

SNA in Sociology and Anthropology

Social network analysis emerged as a useful means of formalizing social properties and processes, providing precise and consistent definitions that allow for logical, testable models of social concepts such as "group" and "social role" (Wasserman & Faust, 1994). The origination stories of social network analysis pull together interdisciplinary threads of research in sociology and anthropology, in particular the work of sociometric analysts who made significant contributions to graph theory. These streams of research converged around the middle of the twentieth century to generate the foundations of social network analysis as it is generally practiced today (Scott, 2000). The earliest examples from sociometry are psychology studies based on self-reported relationships elicited in surveys or interviews (Moreno, 1943; Toeman, 1950), although the terminology for "social network" is usually ascribed to Barnes (1954). The use of archival data for board interlock analysis dates back to 1914 (Brandeis, 1914; Davis et al., 2003), however, providing a long history of archival data mining for indirect evidence of social network structures.

The primary methods for SNA data elicitation in the sociological tradition are surveys and interviews. Though some network researchers prefer to construct networks based upon behavioral observations or archival records, due to the higher likelihood of reliably representing the relationship of interest (de Nooy, Mrvar, & Batagelj, 2005), the vast majority of SNA has been conducted using data derived from questionnaires. Sampling methods vary widely, but are often purposive; for example, snowball sampling of nodes is a frequent choice in social network studies, though its use brings its own problems for boundary specification and implications for interpretation (Laumann, Marsden, & Prensky, 1989). This stream of research has prompted the development of many survey types specific to SNA, such as name generators and roster-based rankings, in which each member of the group being studied is asked to report on relations with every other member.

Traditional sociological surveys inquire of the individuals themselves to describe their relationships, rather than asking about specific interactions; direct responses are typically preferred for measuring the existence of relationship ties between people. This use of this approach acknowledges that people are notoriously poor at reliably reporting discrete interactions but generally good at recalling long-term social structures (Marsden, 1990). Further, direct elicitation of network is often a preferred approach for sociologists and anthropologists, because any given set of observed interactions may not be truly representative of underlying social structure, for a number of reasons that we discuss later in this paper.

At the same time, interaction-based social networks have been the subject of sociological research that employ data from observations, archival records, experiments, and diary studies. Notably, however, the use of elicitation methods other than questionnaires and interviews has typically been an adaptation to conditions that make it otherwise impossible to collect SNA data through the preferred methods (Wasserman & Faust, 1994). This is perhaps a reason that a review of the key SNA journal, *Social Networks*, quickly shows that there are almost no articles that make use of archival data alone (with Adamic and Adar, 2005 a recent exception).

The mode of data collection will impact the assumptions that can be made about the nature of the social network representation. Surveys can be crafted to yield reliable, valid measures of network links representing relatively stable relationships such as those based on friendship or advice (Marsden, 1990). Depending on the type of survey employed, e.g. roster versus free recall response, the existence of links also makes an often implicit statement about the absence of links. If rosters are used, then the absence of a reported link implies the absence of a relationship due to the exhaustive nature of the elicitation method, but in a network based on free recall responses, the absence of a link cannot be assumed to have the same meaning. The conceptualization of absence of links is often neglected, but can be very important for some interpretations, such as studies that examine structural holes (Burt, 1992) or otherwise exploit the constraining nature of a network for possible flow paths, such as studies that draw on indirect centrality measurements (e.g. betweeness or eigenvector centrality).

Borgatti et al. (2009), writing in the journal *Science*, examine the differences between SNA research as carried out in the social sciences and the burgeoning work using similar techniques in the natural sciences, physics in particular. They make the point that social scientists using SNA have been interested in multiplex ties and their interrelationship, meaning multiple ties of different kinds between individuals, stating "social scientists typically distinguish among different kinds of dyadic links both analytically and theoretically" (p. 893). These different types of links include similarities (such as location or membership), social relations (such as kinship), Interactions (such as communication or sex) and flows (such as flow of beliefs). Survey elicitation, sometimes combined with archival data, can be crafted to measure such multiplex ties at each node. By contrast networks research in physics has focused on massive networks derived from trace data and then comparing those structures to randomness to build a typology of overall network structure.

SNA in Information Systems research on online communities

SNA has also seen growing popularity in IS research. It has found many different applications but is particularly prevalent in three areas: diffusion and adoption of IT (e.g., Bruque et al., 2008; Kane and Alavi, 2008), studies of virtual and hybrid organizations (including IS research itself) where it is primarily used to represent organizational structure (e.g., Oh et al., 2005) or social capital (e.g. McLure Wasko and Faraj, 2005) and studies of online markets (e.g., Hinz and Spann, 2008; Xu et al., 2008). This section briefly discusses a few exemplary studies, focusing on how networks were measured in online community research in order to illustrate the points that follow in this paper. It is organized around three aspects of network studies: data sources, the treatment of temporal information and whether networks have considered multiple relationship and node types. The review concludes with a brief discussion of the use of previous results from the sociological literature in the IS literature.

From the earliest work, such as Robey et al. (1989) there has been a tendency to draw on archival data produced during Information System use. Rice (1990) lays out the justification explicitly, "The fact that CMC systems can unobtrusively collect data on usage, flows, and content from a full census of users provides researchers with new opportunities for understanding the application, management, and consequences of such systems. A theoretically appropriate analytical approach is network analysis of CMC system data." (p. 643). System data has been used to construct networks in three key ways: 1) interaction data, 2) association data and 3) survey data, particularly drawing on systems in which users specifically nominate contacts with specific relationships.

Interaction data draws on records of computer-mediated communication, such as email and different types of message boards or work repositories. Email has been a frequent and persistent source of data from which networks have been constructed. For example Ahuja and Carley (1998) studied email interactions in a scientific collaboration, examining the relationship between a structure based on senders and receivers and task performance measured by publications. Two related studies (Aral et al., 2006; Brynjolfsson et al., 2007) utilized email networks, in addition to other data, inside an HR recruiting organization to examine the effect of network position on worker productivity. The Enron data set, which includes internal corporate emails from the Enron corporation made public during a court case has been used for a number of studies (Falkowski et al., 2008; Murshed et al., 2007).

In addition to email, studies have drawn on interaction data from message boards and forums. For example, McLure Wasko and Faraj (2005) studied knowledge contribution and used a network to represent structural social capital as one of their independent variables. They measured a link between two participants by the order in which they posted in a message board, and calculated degree centrality statistics on this network. Studies of open source software development have also drawn heavily on this data source, including data from online bug trackers as well as forums (e.g., Crowston and Howison, 2005; Howison et al., 2006; Long and Siau, 2008; Wu et al., 2007). For example Wu et al. (2007) derived a network structure from interactions in the SourceForge bug-tracking system and used summary measures of these to predict the success of open source software projects.

A second source of network structures come from association networks, often based on data about team or project membership or coincidence in activities. Xu et al. (2005) studied association networks based on activities in open source software projects. Grewal et al. (2006) used association networks based on project membership to identify different types of "network embeddeness", which have a complex relationship to project success, suggesting a number of different types of communities where different structures play different roles. Barbagallo et al. (2008) draw on similar data, suggesting a link between association network structures and software quality. Daniel and Diamant (2008) examined knowledge transfers through networks formed from project membership.

There are only a few studies which have used survey data to measure social network variables, as would be more common in the sociological literature. One example is Kane and Alavi (2008) who administered a survey to healthcare providers where they rated their intensity of interaction with each other and with systems deployed in their organizations. Another is Zhang et al. (2008), which used different rosters to measure online and offline interactions.

More common, however, are papers that utilize pre-existing data sources intentionally produced by the communities themselves as the description of relationships (rather than inferred from traces of their interactions). Wagstrom et al. (2005), in addition to email interaction data, included ties based on "certifications" of open source software participants’ skills levels. Xu et al. (2008) studied an online community where participants made lists of friends and provided reviews of products, assessing the impact of the ‘friend’ nomination on product adoption and diffusion (indicated by a friend providing a review). The IS community has not yet made substantial use of such data from social networking sites, such as LinkedIn and Facebook, although judging by their popularity in cognate disciplines, such as Communications and Computer Science, these will rise rapidly in the next few years and SNA will be an obvious tool.

In general the IS literature has constructed networks which represent only a single kind of relationship, such a "replied to" interaction (e.g., McLure Wasko and Faraj, 2005). Some studies do utilize multiple sources to draw their networks (e.g., Wagstrom et al., 2005) but nonetheless eventually draw their networks with only a single relationship. A rare exception is the work of Kazienko et al. (2008) who studied the photo sharing site Flickr, using different kinds of activity, such as tagging other's photos, or having applied the same tag to a photo, as well as contact lists, eventually outlining "nine separate layers in one multirelational social network", going on to compare structures in different layers. They do not, however, make strong theoretical arguments regarding the possible separate constructs being measured by different layers, as is more common in sociological applications of SNA (Borgatti et al., 2009). Multiple relationships are not the only way to expand the representativeness of networks: Kane and Alavi (2008) argues that SNA research in IS would benefit from a multi-modal approach, meaning different kinds of nodes. Their perspective specifically includes systems as actors, demonstrating their approach through a study of system use in a healthcare setting that draws on the idea of "indirect system use" through interaction of non-system users with system users.

It is common in IS to create networks from longitudinal trace data by collapsing time, sometimes years of interactions, although more recent work has begun to investigate "dynamics" by drawing networks for consecutive time periods, thereby producing time-series of network statistics and analyzing the trends (e.g., Christley and Madey, 2007; Falkowski et al., 2008; Howison et al., 2006; Long and Siau, 2007). Some studies have exploited the temporal nature of trace data by examining the effect of previous interactions on later decisions, such as Hahn et al. (2008) who studied the effect of previous working relationships on later decisions about which open source software project to join. Recent publications have also explored visualization techniques for handling the fine-grained temporality of online discussion data (Trier, 2008).

In summary, then, IS research has tended to construct networks of a single tie type based primarily on trace data collected over time and then aggregated and dichotomized to derive a network for analysis. This contrasts sharply with traditional SNA methods that tend to utilize surveys and interviews, together with some observation, and collect multiplex relationships. In this sense IS research drawing on SNA is closer to the network research undertaken in physics (Ebel and Mielsch, 2002; Kossinets and Watts, 2006), than to network analysis in sociology (Stephen P Borgatti et al., 2009). This is true even though the research questions considered in IS bear greater similarity to those in sociology than to physicists' interest in the topological classification of massive networks and their variation from randomness.

Another regrettable tendency in the IS literature using SNA is that authors make arguments by transferring results from earlier, sociologically focused, SNA studies into the study of virtual organizations. Early work, such as Ahuja and Carley (1998), makes this move explicitly, outlining findings from offline environments and providing some reasoning as to their applicability in online environments, specifically questioning whether the concepts and measures will be appropriate to the new environment. Other works, such as Wu et al. (2007), have been less careful to problematize their adoption of interpretations based on earlier work, instead making claims such as "Past research in social networks has shown that centrality is an important indicator of group performance" and citing an "SNA classic" such as Freeman et al. (1980), without considering that the environment in which the data was generated has an impact on the interpretation of the network measure. Even work that is somewhat cautious about this move, such as Ahuja and Carley (1998) can fall into this pattern. Indeed overall, Ahuja and Carley were unable to replicate findings of objective performance links with a network structure based on email exchange but suggest that "New theories may need to be developed to explain objective performance in virtual organizations", rather than questioning whether measurement issues involved in inferring social structure from trace data might be part of the problem.

Unique Properties of Online Communities Data

SNA seems particularly appropriate for studies of online communities in particular, as it allows analysis of the connections between individuals that result from the interactions. As well, these groups' reliance on electronic interactions provides excellent opportunities for data collection from the ICT directly (Watts, 2007). Garton et al. (1997) go as far as to say that "gathering data electronically replaces issues of accuracy and reliability with issues of data management, interpretation, and privacy", although not all researchers would go this far. We refer to data that can be harvested from the ICT system of the online community that generates them as "online communities data". Writing in an editorial introduction to a recent special edition of Information Systems Research, Agarwal et al. (2008) put it thus, "Most transactions and conversations in these online groups leave a digital trace ... this research data makes visible social processes that are much more difficult to study in conventional organizational settings." Visibility and availability notwithstanding, however, these data bring considerable challenges along with such opportunities.

In some cases, online communities data are similar to what would be collected in a sociological survey. For example, data about friendship links harvested from a system like Facebook might be considered comparable to the results of a survey of group members about their ties. Similarly, data about individuals' membership in a set of online groups maybe comparable to a survey of affiliations and useful to measure an association network. In these cases, harvested online communities data may even be superior to data from a survey, as this form of unobtrusive data collection can eliminate demand biases and missing responses.

Often though, the available data represent actual interactions (e.g., sending or receiving an email). Following Agarwal et al. (2008), we refer to this kind of data specifically as "trace data", as they are stored by the ICT as a trace of a user interaction. These data can be quite useful for research, but we note that they are significantly different from data used in prior studies: nearly all such data are system-generated, longitudinal, and provide complete records of interaction. While these properties seem and may in fact be beneficial for research, we argue that these differences raise significant concerns about the reliability and validity of the measurement of networks for online communities. In the remainder of this section, we discuss the nature of online communities data before turning to a discussion of particular reliability and validity concerns they expose.

System-generated data. The primary difference between online communities data and the data typically used in earlier SNA studies is that the online community data are generally captured by ICT as a byproduct of some underlying social activity, rather than being generated for scientific research. In contrast, instruments like surveys are specifically designed to provide measurements of the theoretical constructs of interest. For system-generated online communities data to be useful for research, researchers must establish construct validity by describing a connection between the captured data and the theoretical construct, similar to using of other archival data. In addition, the system and its interaction with the social processes occurring through it, may play an important role in structuring interaction, a particular point of interest for IS researchers (Orlikowski & Iacono, 2001).

Longitudinal data. A second difference is that in prior studies, networks were often measured as a snapshot of the network structure at a single point in time, e.g., through a survey of the ties of network participants or by observation during a short time period. These approaches are appropriate to measure the relatively stable links being studied. Indeed, many sociologists prefer survey data for exactly this reason: it captures the participants' impressions of the network in general rather than the specific interactions at a particular moment (Marsden, 1990). On the other hand, online communities data, especially trace data, are more typically a time series of events (e.g., email messages sent a different points in time), collected across some period of time. While having longitudinal data can be very valuable for testing certain kinds of theories, such data typically have to be aggregated to reveal some overall network structure (Trier, 2008). As detailed below, the extended period of data collection and the aggregation process have implications for both reliability and validity of the measures (Howison et al., 2006).

Complete data. A final difference is that many online communities are purely virtual groups that interact only via ICT, rather than groups that interact in multiple modes. The nature of the interaction may also be different, as many online communities rely on broadcast rather than person-to-person communications (e.g., interacting through public mailing lists or discussion fora). If the group is indeed purely virtual, then the data from the ICT provides a complete description of the groups' interactions—a different kind of data than a sample of interactions or impression gained from observation or a survey. On the other hand, even apparently pure virtual groups may have "back channels" of communication, e.g., private emails among group members who otherwise interact mostly through public fora.

Overview of the remainder of this paper

The differences in both the phenomenological environment and the kinds of data being employed in SNA studies of online communities suggests a need to carefully consider the validity of both our data as measures of the networks of interest, and the analysis techniques applied. Just as in the application of statistical techniques to data (Straub, Boudreau, & Gefen, 2004), there are a series of questions that must be addressed about the validity of the data and analysis as measures of the theoretical constructs of interest. We follow the discussion of these issues and their implications for research with a set of recommendations to help guide researchers in developing and evaluating this type of research. Our goal is to provide guidance to IS researchers interested in online communities in their use of this technique. We expect that adoption of these recommendations by the IS research community will result in improved rigor and relevance of research results from SNA studies.

Issues in the measurement of networks in online communities

In this section, we discuss reliability and validity issues stemming from the unique nature of online communities data and the interaction of these characteristics with the assumptions of SNA. We start by addressing the reliability of the data before turning questions of validity of network measurements.

Reliability

We first consider questions of measurement reliability, which ask how error prone the measurement of the network is. On the surface, relying on a system to automatically collect data, as with online communities data, would seem to ensure its reliability. However, as noted above, online community data are typically harvested from the ICT that support the community, and with few exceptions, data retrieved from these systems exist to support the operation of the community, rather than being created for research. Issues such as time zone management, server outages, and incomplete or inconsistent event logging can significantly threaten the reliability of network measurement based on these data. For example, in a system that records email messages, times on the messages may be local time for the sender, local time for the server, GMT or (in the worst case) some undecipherable combination. Resolving this question is difficult but necessary to reliably determine the order of messages or aggregate them temporally. Data harvested from a system may be incomplete in unknown ways. For example, when studying communication patterns, it is important to know if community members use back channels of communication, and if so, how often and for what kinds of messages. It would be difficult to study norm enforcement in online communities, for example, if most violations of norms are sanctioned initially via private emails, a plausible scenario. Unfortunately, the contents of these back channels are unlikely to be accessible to researchers, even when the researcher is aware of their usage.

The issues noted above are exacerbated in the case of trace data, for which data collection is extended over time. Online communities and the systems that support them are likely to be constantly evolving, increasing the chances for incompatible changes in data generation or storage. The owners of these systems are generally not concerned with maintaining instrumentation consistency for the benefit of researchers who take advantage of their largesse. Establishing the reliability of the collected data requires careful examination, though it is too often taken for granted. For example, Wiggins et al. (2008), who analyzed interactions on an open source bug tracking system, came across one project in which hundreds of bugs had apparently been resolved within a few minutes. Detailed examination of this case revealed that the project had transferred bug reports from an old system to a new one via a bulk import, with the result that these bug reports were stored with the same open and close time. Including the unreliable data from this project in the analysis would have biased their results, demonstrating that automated data collection and analysis does not obviate the need for close attention to anomalies in the data.

Similar issues exist even with data that researchers themselves have not collected such as database dumps that record community data over time. For example, the SourceForge dumps provided to the Notre Dame SourceForge Research Data Archive (SRDA, http://zerlot.cse.nd.edu/) provide a convenient source of nearly complete data about SourceForge projects. However, the tables in the system are periodically purged to ensure a manageable size for running the website. This process results in database dumps with apparently extensive history that are actually truncated at an arbitrary date with no explicit record of such truncations. The complete history may be available from earlier dumps, but merging these disparate sources is quite difficult, particularly given the system changes made over time that result in incompatible database schema. Similarly, systems that make usage-reporting data available may change their data sources or methods of calculation without notice, and almost undoubtedly without recalculating historical usage reports according to the new method.

In summary, system generated data is subject to numerous threats to the reliability of the data for scientific research. Intimate knowledge of the online community system and its quirks is required to establish the reliability of data harvested from the system. Unfortunately, the system details needed to assess instrumentation reliability are rarely public, and hard to obtain even for participants in the community who often are not privy to system administration details. Researchers with personal connections or who are otherwise in a position to acquire this information have an advantage in establishing the reliability of their measurements, though such particularism is difficult to demonstrate in ways that match the expectations of the open values of science.

Construct validity

We turn next to questions of construct validity, which address the question of "the extent to which a given test/instrumentation is an effective measure of a theoretical construct." (Straub et al., 2004; p. 424). As with any operationalization, the selection of data has to fit the chosen conceptual framework. We will consider in turn validity issues created by using data that are system-generated, longitudinal, and complete.

Validity and system-generated data

Our first observation is that with online community data, data are largely given by what the system stores and often further constrained to data that are publicly available, making the theoretical fitting *post hoc* and often *ad hoc*. Online communities data provide many choices for identifying links between individuals, but many studies are surprisingly vague about the theoretical rationale for the choice of a particular construct and its connection to the data. As a result, there is often a significant mismatch between the data and the intended construct. When connecting data to theory, it is important to be conscious of the underlying technical and social processes that led to the data's creation and storage. For example, research might use Facebook "friends" as evidence of a social link, but such an argument is difficult to support without knowing how individuals decide whom to "friend" and the consistency of their decisions over time.

In addition, the researcher must consider the meaning of the absence of a link in the data. Network analysis research typically assumes that the absence of a link means that there is no connection between those individuals. While this may be a reasonable assumption in some cases, as previously discussed, it is not always justifiable. It is therefore incumbent upon the researcher to delineate the assumptions about the ontological nature of the absence of a link. In the example of a network based on Facebook friends, without knowing how individuals decide whom to friend, it is hard to ascribe a useful interpretation to the absence of a friend link on the system.

Example: Flow of information from communications links

To make the discussion of validity concerns more concrete, we will use a specific example of a theoretical construct of interest and a network that might operationalize it, namely the construct of information sharing, which is important in innovation, diffusion and contribution studies (e.g., Brynjolfsson et al., 2007; McLure Wasko and Faraj, 2005). (Similar arguments could and should be made for other constructs and data sources, but this example will be sufficient to illustrate our methodological argument.)

Information sharing can be studied from a network perspective by measuring the network of individuals linked through their communication activities. Given an information-sharing network, SNA can provide insight into the processes of information sharing by identifying key individuals and providing measures for comparison of different groups. For example, high betweenness centrality indicates which individuals are on the shortest path between many others, and therefore positioned to affect the flow of information. Likewise, network diameter indicates the maximum number of links through which information must travel in order to be transmitted between an average pair of individuals, suggesting how responsive a group may be to new information.

However, the validity of these measurements of the communications network depends on the assumption that information in fact flows along the measured links, as would typically be the case in a face-to-face network. This assumption about information flow may also be valid for interaction via ICT, as when emails are exchanged directly from senders to a short list of recipients listed in the message (e.g., private uses of email). On the other hand, many online communities employ listservs for which all emails are archived and made publicly available to all community members (or even to the general public). When email communications occur via a listserv, whether archived publicly or not, we have little or no direct evidence of information flow and control. Email listserv messages may be read by only the people who are replying in a given thread, by every member of the email list, or more likely, by some unknown proportion of the email list subscribers (Howison et al., 2006) and possibly even non-community members accessing the archive. Very little work has directly examined readership, since it usually leaves no trace data; notable exceptions are Lakhani and von Hippel (2003) and Yeow et al. (2006).

A common strategy with such data is to examine the structure of message responses, i.e., using "reply to" message threading structures to define a link (e.g. Crowston and Howison, 2005; McLure Wasko and Faraj, 2005; Wu et al., 2007). Unfortunately, response structure is not a valid measure of information flow. While those who reply to a message have most likely read it, non-response does not indicate that other members have not. The private versus public nature of some trace data with which networks may be constructed therefore changes the way we can understand and interpret measures of information flow, control, and brokerage. For example, if messages are publicly posted, it is simply impossible to argue the meaningfulness of indirect measures, such as betweenness or closeness, as measures of importance based on information control, because in this case there is no such mediating control (e.g., Bird et al., 2006; Wu et al., 2007). Similarly, calculations of the diameter of a reply-to network are meaningless for understanding information flow if information is broadcast on a mailing list, potentially reaching all group members at once.

Example: Intensity of relations

Intensity of relationships introduces another challenge for the application of SNA in communication networks based on trace data (Garton et al., 1997) and this challenge has, on the whole, not been well met. The intensity issue turns on the argument that the strength of ties affects the nature of interactions between individuals (Granovetter, 1973). Offline SNA work has approached this through survey questions about different types of relationships based on strength, allowing participants to translate their memory and interpretation of patterns of past interactions into measures. Direct interaction data from online communities would seem to provide useful data, since a count of multiple messages exchanged over time (or other quantifiable link characteristics, like the rate of message exchange or the volume of text in the messages) can be used to indicate varying intensities of interaction between actors by creating weighted networks. Without more contextual information to guide the selection and interpretation of measures of intensity, however, the researcher must choose an operationalization themselves.

There are a number of techniques for taking this information into account. One approach is unit weighting, which increases the weight of each edge value by a fixed unit for each message between a pair in the network sample. Node strength is also an option for evaluating centrality with this edge weighting method (Valverde, Theraulaz, Gautrais, Fourcassie, & Sole, 2006), indicating the volume of activity in dyadic pairs. Other approaches include applying a time-based decay (Wiggins et al., 2008) that gives more weight to more recent interactions.

Unfortunately, relatively few SNA techniques are intended for use with weighted networks. Most assume dichotomous relationships because they were designed to evaluate networks of abstract relationships assumed to be of roughly equal strength, as opposed to highly variable interaction-based links from trace data. Indeed, the assumption of dichotomous inputs is so prevalent that some SNA software packages do not even safeguard against misuse, for example summing link values for degree instead of counting links. Misapplication of techniques intended for use in dichotomous networks will generate incorrect results when applied to weighted networks.

As few robust techniques utilize edge weights, the usual analysis approach calls for dichotomizing the networks based on threshold criteria (e.g., only counting links with more than 5 interactions). Such an approach seems unsatisfactory for several reasons. First, dichotomization involves throwing away much of the available source data. Second, dichotomization requires selecting threshold criteria, which can be sensitive to such factors as the size of the data sample. As a result, careful analysis is also needed to determine appropriate selection criteria for setting thresholds. Finally, dichotomization assumes that the construct is in fact binary, as opposed to continuous. Alternately, in some cases, it may be more appropriate to treat high and low levels of interaction frequency as indicative of different types of relationships, as in Granovetter's (1973) theory of weak and strong ties.

Unfortunately decisions about dichotomization are usually acknowledged only in passing or mentioned as a limitation at the end of papers (e.g., Ahuja and Carley, 1998; Crowston and Howison, 2005; Wagstrom et al., 2005). When the interpretations of participants' own understandings of the importance and meaning of past patterns of interactions is not available, the threshold point at which a pattern of interactions (such as count, recency, multiple channels or even content) becomes representative of the strength or quality of a relationship is a key methodological decision with clear construct validity implications.

Validity and longitudinal data

We next examine issues of validity created by the temporal nature of the data. This discussion is specific to trace data that is longitudinal and episodic, in contrast to more common cross-sectional data samples. As noted above, measuring a network based on trace data requires aggregating the interactions over some period of time. Aggregating interactions over a brief period of time produces a snapshot data set, while longitudinal data are aggregated from interactions occurring over a longer period of time. While the research context should inform the decision as to how much time constitutes a snapshot versus a longitudinal data set, longitudinal data can usually be repackaged as a series of snapshots, and occasionally snapshot data sets can be assembled into a longitudinal form. As a result, consideration of the dynamics of networks over time is important for meaningful interpretations of SNA measures.

Aggregation creates particular problems when the links are directed. For example, consider a study of information sharing using point-to-point communications links, in which A sends a message to B and B sends a message to C. If the messages are sent in this order, it is possible for A's information to make it to C, but not if the messages occur in the opposite order (see Figure 1). Similarly, in the case of an association network, if two individuals are members of a group at the same time, there is a possibility of some kind of influence process, but if their memberships do not overlap, the influence can be in one direction at best. Unfortunately, aggregating links across time to form a single network will likely suppress these nuances, possibly leading to invalid conclusions. It is possible, although very rare, to avoid this type of issue by representing the "network" as a set of actual sequential paths through nodes, as is done by Brynjolfsson et al. (2007).

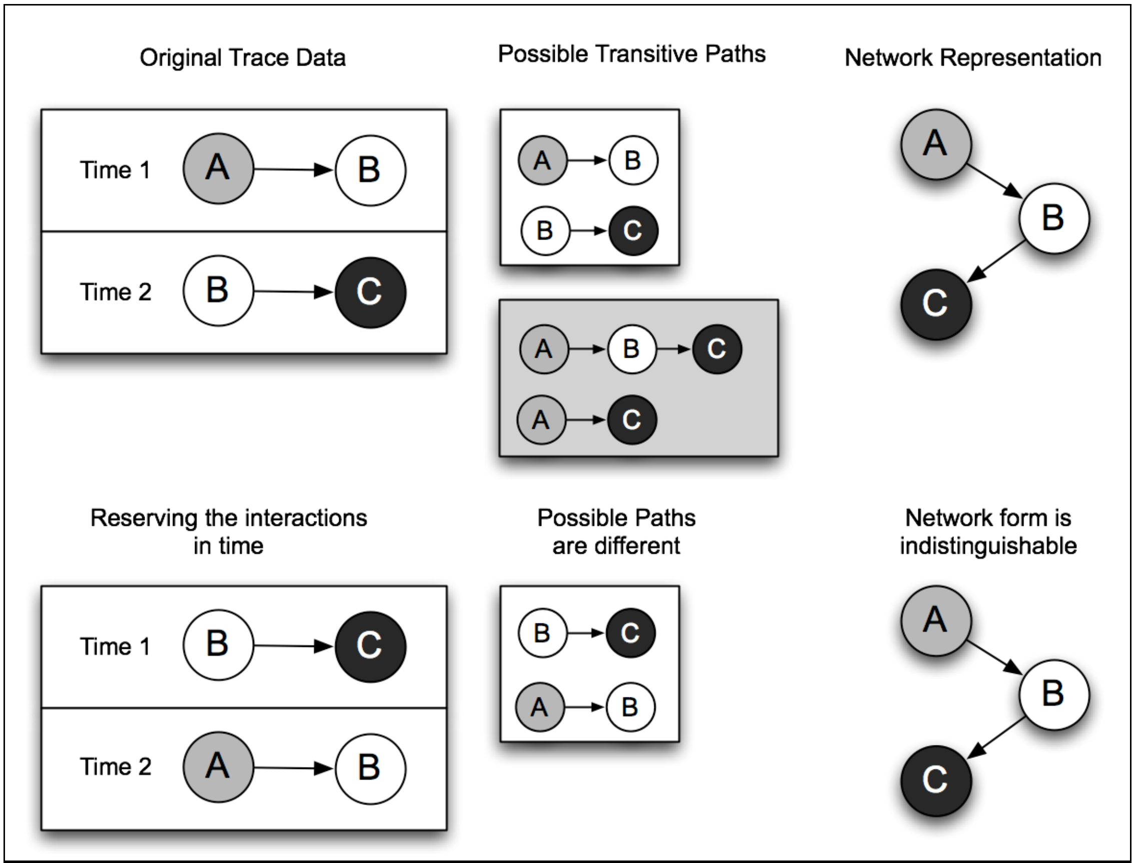
Original Trace Data
Possible Transitive Paths
Network Representation
Time 1
A
B
Time 2
B
C
A
B
B
C
A
B
C
A
C
A
B
C
Reserving the interactions in time
Possible Paths are different
Network form is indistinguishable
Time 1
B
C
Time 2
A
B
B
C
A
B
A
B
C

Figure 1: The implications of collapsing networks built from trace data over time, note the indistinguishable network representations but the quite different set of possible paths.

Even with undirected networks, a key question is the stability of the construct of interest, compared to the nature of the data collected. The particular construct measured as a network link may be conceptualized as being stable (e.g., long-term friendship ties) or dynamic, meaning that the network structure changes and evolves over time (see Huisman and Snijders, 2003; Leskovec et al., 2005). These characteristics of network data and constructs, shown in Table 1, have varying degrees of impact on the internal validity of conclusions drawn from SNA.

Table 1: Time characteristics of data and stability of the construct of interest determine appropriateness of aggregate or dynamic analysis approaches

Network data characteristics	Stable Construct	Unstable Construct
Snapshot data	Aggregate	Aggregate
Longitudinal data	Aggregate	Dynamic

We note that for three of the four pairings of data type and construct stability, aggregating the network structure data does not significantly impact the validity of analysis and interpretations. If the constructs of interest are stable, then aggregation of interactions in the form of snapshots and longitudinal representations should yield similar overall results. For example, networks of familial relationships will show more-or-less the same links in both snapshot and longitudinal representations, with the exceptions of the addition or subtraction of actors over time due to birth, death, marriage and divorce. If the constructs are not stable, then a snapshot will reveal the configuration at that point in time, and although the network structure in this snapshot cannot be generalized to other points in time, it may still provide useful insights into social processes.

However, aggregating longitudinal data about unstable constructs can lead to misleading results. For example, Wiggins et al. (2008) reports examining centrality in open source development teams initially aggregated interaction data across the life of projects and were surprised to discover that some projects had many apparently central actors, suggesting a relatively decentralized structure, while other projects had only a few or just one. However, when they examined the data dynamically, they discovered that all of the projects in fact exhibited a high degree of centralization at any point in time, but in some, the most central actor would change from time to time. It was only when a succession of centralized networks were aggregated that the resulting network appeared to have multiple central nodes and thus to be decentralized, leading to invalid conclusions if corrective action was not taken.

Unfortunately, communication networks constructed from trace data quite often fall into the final category of longitudinal data and dynamic phenomena. Because they are based upon evidence of communications over time, the data are longitudinal. As well, the construct being measured may be more or less dynamic in the network structure of interactions; private communications (e.g., email networks constructed from email server logs, see Kossinets and Watts, 2006; Tyler et al., 2003; Wu et al., 2007) may demonstrate more consistent structures than public communications with high social transparency as in common in online communities. Thorough investigation would suggest the need for a preliminary analysis to verify that a longitudinal network data set was in fact stable with respect to the construct of interest before applying aggregate analysis measures.

Validity and complete data

Finally, we consider validity concerns raised by the completeness of recorded data about online community interactions. In traditional SNA data collection, researchers expect that they cannot observe most of the interactions between individuals; there are understood to be more interactions than periodic or partial observation will measure. Therefore, the observation of a specific interaction is assumed to indicate the presence of many similar unobserved interactions, particularly if more than one actual interaction is observed. In other words, the universe of all interactions between members of a community can be thought of as a population from which the particular observations (or reported links) are somehow sampled, allowing the application of inferential logic to make claims about all interactions, and from there to relationships. In many face-to-face groups, it might further be assumed that the intensity of interactions is roughly comparable (i.e., an assumption about the likely distribution in the population), which again facilitates inferences about the population from the sampled interactions.

However, because the ICT that provides a platform for online community interaction may also archive every interaction, it can provides persistent evidence that is very different from available data about face-to-face interactions that leave no observable records. As a result, in the case of trace data from online communities, particularly those with little or no opportunity for offline interaction, researchers may in fact be observing all interactions in the community space, thus generating a census rather than a sample of links.

The completeness of the data on the one hand is a good thing, as it allows more definite conclusions to be drawn based upon the observed dynamics. Researchers using these data have a rare and enviable degree of certainty that the data are comprehensive, and little if anything has been omitted, barring reliability issues previously discussed. On the other hand, researchers using such data must thus be wary of the human tendency to infer structure from interactions and assume that evidence based on a set of events is representative of deeper meaning. In the case of trace data, what you see is what there is. In this situation of rich observation, there is no reason to assume that the observed interactions represent a partially hidden pattern of interactions; the pattern is, in fact, quite explicit. The result is an otherwise unusual situation in social science research, which requires thinking differently about the analysis. Researchers can readily acquire sufficiently complete data that inferential statistics or thinking are no longer necessary or appropriate. Inappropriate use of inferential logic poses a validity risk in a wide range of analyses, particularly those using interaction data (e.g., Aral et al., 2006; McLure Wasko and Faraj, 2005, and many others).

The distinction between sample and census mirrors the structuralist versus connectionist approach to creating networks, which discriminates between the way ties and their functions are treated (S. P Borgatti & Foster, 2003). The structuralist view focuses on ties as a topology, while the connectionist perspective sees ties as flows of resources. In the first case, the measured interactions are seen as describing a topology on which or through the phenomena of interest are assumed to occur (e.g., knowledge sharing happening through a network of communications). From the latter perspective, the links do not form the topology, but instead represent the actual flows of interest (e.g., actual instance of knowledge sharing in action).

These two views require different ways of interpreting the data, but are often confused.

This difference is particularly important when it comes to reasoning about network processes: do these occur over the network or form the network? If a set of observations form a sample from a much larger distribution, then it is reasonable to infer many interactions beyond those observed. By taking the observed interactions and collapsing them over time the analyst obtains measures of persistent relationships that are generative of interactions. For example, in studying knowledge-sharing, the analyst can interpret the meaning of the interactions from "spoke to" to "knows" and use this new relationship to infer other, unobserved, "spoke to" interactions. These other interactions could be channels for information transmission, influence or other network processes.

As a concrete example of this approach, the observation that Person A spoke with Person B in Week 1 of the study might be taken as evidence of a link down which information could travel. The analyst might then infer the likely existence of other specific communication events. Their network transformation relies on a reasonable expectation that Person A would speak with Person B at a later date, and perhaps that Person B would speak back to Person A at some point. Indeed, this logic is behind the validity of observing only the second set of interactions shown in Figure 1, collapsing to the network form, while assuming that the additional paths of the first set are likely to have occurred at some unobserved point in time and thus including them in interpretation.

On the other hand, if the researcher is reasonably confident of having observed all interactions, this conclusion is invalid. Regardless of any relationship that may be indicated by Person A speaking to Person B in Week 1, if the data do not indicate that they speak again, this link does not provide an information channel. As a result, interpretations that tacitly or explicitly assume processes taking place over the network topology should be considered suspect when it is likely that the data show close to the totality of interactions. Researchers should exercise great caution not to overstep the bounds of what is reasonable to assume based on the class of events represented by the data set. Unfortunately, as demonstrated in Figure 1, making this assumption can occur in the very act of drawing the network.

Inappropriate use of inferential logic poses a risk also to studies of association network data (e.g., Daniel and Diamant, 2008; Grewal et al., 2006). In the case of association network data the association is often interpreted as though it were a proxy for interactions, even when detailed interaction data is available that circumscribes the possible paths or when temporal overlap data regarding membership is dropped (Christley & Madey, 2007). Very few online community SNA studies avoid this issue, with Brynjolfsson et al. (2007) and Hahn et al. (2008) being notable and commendable exceptions.

Of course, the reverse caution is also relevant. Even if there are complete archives of interactions, interaction data alone may not represent the complexity of social reality in the way that traditional social network elicitation methods attempt to do. For example, a record of every interaction in any online community cannot satisfactorily reflect relationships that extend beyond the boundaries of that online community (should they exist).

Recommendations

There are many challenges to address as researchers work to improve the quality of research using social networks analysis with trace data from online communities. As the variety of research in this domain readily makes apparent, the scope of these issues are beyond a single prescriptive approach. We can recommend, however, a three-point review process to help researchers assess the strength of their own work as well as that of others. By checking the quality of data, providing adequate theory and validation and evaluating the appropriateness of dynamic analyses, a significant gain can be realized in the rigor of information systems research on online communities using SNA. In addition, these points are sensible for authors to include as an appropriate level of detail in the research methods section for papers using either network analysis techniques or trace data; both reviewers and editors should feel justified in requesting that authors address these three review points when evaluating works that describe SNA results. Finally, we recommend a more careful consideration of research ethics in these studies; the potential mismatch of organizations and participants who may experience benefit or harm from SNA studies of online communities merits thoughtful evaluation.

1. Verify the quality and reliability of system-generated data

The first step for any analysis using interaction data should be to verify the quality of the data for reliability, which affects its suitability for answering the research question. Reliability issues are particularly important when the research design compares different networks with data generated from different systems; in such cases, great care must be taken to ensure that the measures for comparison are appropriately congruent.

Relevant techniques for evaluating data reliability include: 1) validation checking through community members with intimate system knowledge, e.g. asking administrators, “Have you ever changed the way you calculate these statistics?”, “Can you remember episodes of data loss or system outages?”, etc.; 2) face validity checks on time series of activity, which can apply whole-network measures to look specifically for indicators of anomalies such as unreasonable truncation, steps or spikes, phase changes, or direction reversal in cumulative processes; 3) validating the source of any such anomalies that cannot be dismissed as measurement problems; and 4) cautiously applying corrective analytic strategies for normalizing longitudinal data that are affected by changes in measurement.

2. Be explicit about operationalizations and their validation

The second point for evaluation is whether the data and analysis metrics have a sound theoretical basis. SNA studies very rarely discuss the validation of their network measures, quite unlike survey-based research which has a standard methodology for validate items of scales. Far too often the SNA work in IS relies on demonstrating expected theoretical correlations between measures, but such correlations cannot simultaneously establish the validity of independent measures and the validity of hypothesized relationships between them.

Validation of metrics by methods triangulation is particularly desirable for establishing the internal validity of the specific approach. Face validity is more readily demonstrated when authors illustrate findings with a narrative of how the data provide a reasonable operationalization of theory. A good story from the data strongly improves the ability of the reader to understand, but also to critique the interpretations made of the network data throughout a paper.

In IS studies, in particular, researchers ought to consider whether their analysis has adequately considered the role of the IT artifact in communication, recognizing that collapsing networks to person-to-person links may inappropriately disguise the role of the now absent IT artifact (Orlikowski & Iacono, 2001). This is especially of concern when links are combined from interactions over different types of media or different types of data, such as intentional links as with ‘friending’.

As is always the case for research on multi-level phenomena, SNA researchers must ensure that level of analysis is carefully considered to avoid ecological fallacy and generalization errors. Fortunately, every network measure is crafted with this consideration in mind, and investigators may select measures specifically designed for analysis at the individual level or group level. Thus, our recommendation is to verify that the selected measure is appropriately aligned with the level of analysis for the research question. If the measure and level are not properly aligned, researchers should consult the literature for alternate choices of measures that are at the appropriate level; network measures are complex calculations, and it is nearly always inappropriate to perform otherwise commonplace analytic transformations, such as taking the average of individual-level measures to obtain a network-level measure or log-transforming a distribution of individual or group level scores to achieve normality.

When constructing weighted networks, the theoretical basis for weighting links should be well-reasoned and explicit, as weighting introduces a great deal of additional complexity. For application of dichotomous measures to a weighted network, choosing a threshold at which to dichotomize the network can have a large impact on the results. Therefore, the chosen dichotomization threshold ought to be argued from theory, or its impact tested empirically through a sensitivity analysis; preferably both. In studies of email networks, multiple thresholds may be required; for example, researchers have applied thresholds that dictate individuals must exchange a total of at least 6 emails over a period of three months (Adamic & Adar, 2005), and messages must have 10 or fewer recipients so as to avoid conflating broadcast-style email messages with more direct communications (Tyler et al., 2003).

When choosing appropriate measures of network phenomena, such as centrality/centralization, it is vital to consider their relationship with the nature of the collected data. For example some measures and their interpretations rely heavily on the drawn network as a constraining topology, over which multiple processes play out. The interpretation of indirect centrality as brokerage power is one such interpretation. On the other hand, if the network does not constrain communications, such as the pattern of replies or thread participation in public venues, then relying on interpretations built on constraining networks is inappropriate. Of particular assistance in this endeavor is the recent work of Borgatti and colleagues on building a taxonomy of tie types and relevant network mechanisms. Borgatti et al. (2009)

identifies four types of network processes: transmission, similarity, binding and exclusion. Borgatti (2005) provides detail on transmission mechanisms, which are the most common type of mechanism considered in IS. These involve the transmission of something between network nodes and can be classified according to whether that thing is thought to move by a copy mechanism (such as ideas) or a move mechanism (such as money), as well as whether the type of path through the network that the thing follows (e.g. shortest path, random path, parallel paths). They argue that getting this understanding correct is key to choosing the appropriate measures, as "different measures make implicit assumptions about the manner in which things flow in a network". Getting this matching wrong means that "we lose the ability to interpret the measure ... or we get poor answers" (p. 56). The assumed network mechanisms have implications for appropriate analyses, such as using linear regression when the process is non-linear, or using a grouping algorithm that has its logic in similarity or binding mechanisms when the data were constructed from a flow logic. Researchers ought to carefully assess—and state in publications—the logical connection between the type of nodes, links and network processes they examine and the set of measures they calculate as well as the appropriate analysis of such measures. Finally, researchers should be wary of interpretations of complete interaction data that rely on inferential arguments.

3. Determine and design for the temporal stability of constructs

Networks constructed from aggregated data over time would benefit from examining the empirical stability of the network metrics and thus the appropriate operationalizations, perhaps drawing on conceptual work regarding time in organizations (e.g., Bluedorn and Standifer, 2006; Van de Ven and Poole, 2005; Zaheer et al., 1999). Methods for evaluating construct stability include examining the variability of key measures in a series of snapshots of the network, varying the aggregation periods, employing sliding windows, and periodically recalculating the network measures of interest. Methods of examining binning effects for constructing histograms or time-series may also be appropriate, as many social networks demonstrate scale-free properties that make logarithmic scales preferable to linear cumulative binning. A punctuated equilibrium view of time, in which time periods are not arbitrarily based on convenient calendar units but instead on important events in the community, can also be a productive perspective, though it can prove analytically challenging.

It is possible to avoid some of the issues of temporality in networks by understanding a network not as a timeless set of nodes and links, but as a collection of actual sequential paths taken through the network. Brynjolfsson et al. (2007) demonstrate this type of analysis in their study of email networks within an organization. This alternative network representation is finding traction in social network studies outside of IS as well, where it is often called "walk" or "trail" data and it is a way to avoid the validity issue shown in Figure 1.

Research ethics in online communities

Finally, we wish to draw attention to possible ethical issues in the application of network methods to online communities data. Social networks are a summary of data that provide insight into social structures; the complex ethics of collecting, analyzing, and reporting social networks data lies, in part, with the consideration of who will benefit from the research (Rice, 1990). According to Kadushin (2005), the benefit of social networks research rarely accrues to the individual, who bears greatest risk for potential harm. As well, the data necessary to conduct SNA-based studies is quite detailed, such as roster-based reports of ties or logs of interactions. As a result, Institutional Review Boards (IRBs) often reject social network research proposals that are based on traditional elicitation measures due to issues with participant anonymity and presumed privacy, unless the IRB has a network research review protocol.

On the other hand, network data for online communities are based on materials that are intentionally public and therefore may be considered a form of publication or mass communication, exempt from the need for informed consent (Herring, 2004). However, the public nature of this data does not relieve researchers of the obligation to consider the ethical implications of its use in their research. Even when the data are publicly available, aggregating them has the potential to reveal potentially damaging associations for the individuals in the network. Common ethical standards for social network analysis generally require anonymizing nodes for confidentiality purposes (Klovdahl, 2005), but in online communities, this is not sufficient to prevent discovery of individual identities. Despite the loss of information incurred in creating a network, the anonymized network representation of the real structure can be transparent to those who are embedded within it. A typical social network will contain one or two highly connected nodes, and any participant in the network will likely know who these highly connected individuals are simply by virtue of their "popularity," and by inference, may be able to identify other nodes in the network.

In addition, as Narayanan and Shmatikov (2009) demonstrate, social networks can be de-anonymized algorithmically due to the uniqueness of each individual's social network topology. Therefore, use of SNA in online communities research presents the difficult ethical question: when anonymization can be rendered essentially ineffective, how can researchers adequately protect the individuals represented in the network from potential harm? This issue is particularly pressing when the network represents professional or business networks (Borgatti & Foster, 2003), or reveals an informal social network that may run contrary to expected formal network structures.

Unfortunately, we are unable to provide definitive solutions to this quandary; it seems that anonymization of the networks is not enough, but perhaps further anonymization of the context of the network, beyond that which is customary, may assist in obfuscating individual identities. This ethical issue of protection of subjects makes the otherwise sensible, and increasingly commonplace, provision of exemption by the IRBs for research using public data gathered from online sources a questionable practice. Aggregating otherwise disparate data can be surprisingly revealing in ways that the original sources were not, and we suggest that serious consideration of potential harm from public data aggregation in SNA is appropriate, particularly as researchers prepare IRB applications and report results.

Conclusions

Social network analysis provides a set of powerful techniques that is particularly well suited to the study of the use of Information Systems and particularly online communities, both inside, outside and across organizational boundaries. It provides a way of harnessing the data contained in archives and of reducing it to operationalize theoretically interesting concepts. Further, it provides a much-needed way to join the individual and group levels of analysis and so allows researchers to employ novel constructs in their analyses.

However, this paper sounds a strong note of caution about the manner in which SNA concepts have been translated to the IS field and to research on online communities in particular. Through an analysis based in a detailed consideration of the type of data available and widely used, the paper has demonstrated that online communities data is of a different sort than used in earlier studies using SNA, raising a set of pernicious reliability and validity concerns which extend from the initial transformation and reduction to networks and the chain of logic from construct, operationalization and analysis of those networks. We have attempted to highlight studies that deal well with these challenges as well as a few that illustrate the concerns. Finally a set of recommendations for both researchers and reviewers are provided; it is hoped that these can improve the quality of SNA based research in IS, especially in terms of theoretical correspondence, and position the field to make important contributions to the "Twenty-first century Science" (Watts, 2007) of network analysis of online activity.

Bibliography

Adamic, L., & Adar, E. (2005). How to search a social network. *Social Networks*, *27*(3), 187–203.

Agarwal, R., Gupta, A. K., & Kraut, R. (2008). Editorial Overview–The Interplay Between Digital and Social Networks. *Information Systems Research*, *19*(3), 243-252. Retrieved from http://isr.journal.informs.org/cgi/content/abstract/19/3/243.

Ahuja, M. K., & Carley, K. (1998). Network structure in virtual organizations. *Journal of Computer-Mediated Communication*, *3*(4).

Aral, S., Brynjolfsson, E., & van Alstyne, M. (2006). Information, Technology and Information Worker Productivity: Task Level Evidence. In *ICIS 2006 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2006/21.

Barbagallo, D., Francalenei, C., & Merlo, F. (2008). The Impact of Social Netowrking on Software Design Quality and Development Effort in Open Source Projects. In *ICIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2008/201.

Barnes, J. A. (1954). Class and committees in a Norwegian island parish. *Human relations*, *7*(1), 39.

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006). Mining email social networks. In *Proceedings of the 2006 international workshop on Mining software repositories* (pp. 137–143).

Bluedorn, A. C., & Standifer, R. L. (2006). Time and the Temporal Imagination. *Academy of Management Learning and Education*, *5*, 196-206.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, *27*(1), 55-71.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of management*, *29*(6), 991.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network Analysis in the Social Sciences. *Science*, *323*(5916), 892-895. doi: 10.1126/science.1165821.

Brandeis, L. D. (1914). *Other Peoples' Money, and How the Bankers Use It*. Stokes.

Bruque, S., Moyano, J., & Eisenberg, J. (2008). Individual Adaptation to IT-Induced Change: The Role of Social Networks. *Journal of Management Information Systems*, *25*(3), p177 - 206. Retrieved from http://search.ebscohost.com.libezproxy2.syr.edu/login.aspx?direct=true&db=bsh&AN=36456538&site=ehost-live.

Brynjolfsson, E., van Alstyne, M., & Aral, S. (2007). Productivity Effects of Information Diffusion in E-Mail Networks. In *ICIS 2007 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2007/17.

Burt, R. S. (1992). *Structural Holes*. Cambridge, MA: Harvard University Press.

Christley, S., & Madey, G. (2007). Global and Temporal Analysis of Social Positions at SourceForge.net. In *The Third International Conference on Open Source Systems (OSS 2007), IFIP WG 2.13*. Limerick, Ireland.

Crowston, K., & Howison, J. (2005). The social structure of Open Source Software development teams. *First Monday*, *10*(2). Retrieved from http://firstmonday.org/issues/issue10_2/crowston/index.html.

Daniel, S. L., & Diamant, E. I. (2008). Network Effects in OSS Development: The Impact of Users and Developers on Project Performance. In *ICIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2008/122.

Davis, G. F., Yoo, M., & Baker, W. E. (2003). The Small World of the American Corporate Elite, 1982-2001. *Strategic Organization*, *1*(3), 301.

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

Falkowski, T., Barth, A., & and Spiliopoulou, M. (2008). Studying Community Dynamics with an Incremental Graph Mining Algorithm. In *AMCIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/amcis2008/29.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291.

Freeman, L. C., Roeder, D., & Mulholland, R. R. (1980). Centrality in social networks: II. Experimental results. *Social Networks*, *2*, 119–141.

Garton, L., Haythornthwaite, C., & Wellman, B. (1997). Studying online social networks. *Journal of Computer-Mediated Communication*, *3*. Retrieved from http://jcmc.indiana.edu/vol3/issue1/garton.html.

Granovetter, M. (1973). The strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360-1380.

Grewal, R., Lilien, G. L., & Mallapragada, G. (2006). Location, Location, Location: How Network Embeddedness Affects Project Success in Open Source Systems. *MANAGEMENT SCIENCE*, *52*(7), 1043-1056. doi: 10.1287/mnsc.1060.0550.

Hahn, J., Moon, J. Y., & Zhang, C. (2008). Emergence of New Project Teams from Open Source Software Developer Networks: Impact of Prior Collaboration Ties. *Information Systems Research*, *19*(3), 369.

Herring, S. C. (2004). Content analysis for new media: rethinking the paradigm. In *New Research for New Media: Innovative Research Methodologies Symposium Working Papers and Readings* (pp. 47–66).

Hinz, O., & Spann, M. (2008). The Impact of Information Diffusion on Bidding Behavior in Secret Reserve Price Auctions. *Information Systems Research*, *19*(3), 351.

Howison, J., Inoue, K., & Crowston, K. (2006). Social dynamics of free and open source team communications. In E. Damiani, B. Fitzgerald, W. Scacchi, & M. Scotto (Eds.), *Proceedings of the IFIP 2nd International Conference on Open Source Software (Lake Como, Italy)*, IFIP International Federation for Information Processing (Vol. 203, pp. 319-330). Boston, USA: Springer. Retrieved from http://floss.syr.edu/publications/howison_dynamic_sna_intoss_ifip_short.pdf.

Huisman, M., & Snijders, T. A. B. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods & Research*, *32*(2), 253–287.

Kadushin, C. (2005). Who benefits from network analysis: ethics of social network research. *Social Networks*, *27*(2), 139–153.

Kane, G. C., & Alavi, M. (2008). Casting the Net: A Multimodal Network Perspective on User-System Interactions. *INFORMATION SYSTEMS RESEARCH*, *19*(3), 253-272. doi: 10.1287/isre.1070.0158.

Kazienko, P., Musial, K., & Kajdanowicz, T. (2008). Profile of the Social Network in Photo Sharing Systems. In *AMCIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/amcis2008/173.

Klovdahl, A. S. (2005). Social network research and human subjects protection: Towards more effective infectious disease control. *Social Networks*, *27*(2), 119–137.

Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, *311*, 88–90.

Lakhani, K., & von Hippel, E. (2003). How open source software works: “free” user-to-user assistance. *Research Policy*, *32*, 923–943.

Laumann, E. O., Marsden, P. V., & Prensky, D. (1989). The boundary specification problem in network analysis. *Research methods in social network analysis*, 18–34.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, New York, NY, USA* (pp. 177–187).

Long, Y., & Siau, K. (2007). Social Network Structures in Open Source Software Development Teams. *Journal of Database Management*, *18*(2), 2–40.

Long, Y., & Siau, K. (2008). Impacts of Social Network Structure on Knowledge Sharing in Open Source Software Development Teams. In *AMCIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/amcis2008/43.

Marsden, P. V. (1990). Network data and measurement. *Annual review of sociology*, *16*(1), 435–463.

McLure Wasko, M., & Faraj, S. (2005). Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice. *MIS Quarterly*, *29*(1), p35 - 57. Retrieved from http://search.ebscohost.com.libezproxy2.syr.edu/login.aspx?direct=true&db=bsh&AN=16313358&site=ehost-live.

Moreno, J. L. (1943). Sociometry and the cultural order. *Sociometry*, *6*(3), 299–344.

Murshed, T., Davis, J., & Hossain, L. (2007). Social Network Analysis and Organizational Disintegration: The Case of Enron Corporation. In *International Conference on Information Systems (ICIS2007)*.

Narayanan, A., & Shmatikov, V. (2009). De-anonymizing Social Networks. In *IEEE Security & Privacy '09*.

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with*

Pajek. New York: Cambridge University Press.

Oh, W., Choi, J. N., & Kim, K. (2005). Coauthorship Dynamics and Knowledge Capital: The Patterns of Cross-Disciplinary Collaboration in Information Systems Research. *Journal of Management Information Systems*, *22*(3), p265 - 292. Retrieved from http://search.ebscohost.com.libezproxy2.syr.edu/login.aspx?direct=true&db=bsh&AN=19675859&site=ehost-live.

Orlikowski, W. J., & Iacono, C. S. (2001). Research Commentary: Desperately Seeking the 'IT' in IT Research: A call to theorizing the IT Artifact. *Information Systems Research*, *12*(2), 121-145.

Rice, R. E. (1990). Computer-mediated communication system network data: Theoretical concerns and empirical examples. *International Journal of Man-Machine Studies*, *32*(2), 627-647.

Ridings, C. M., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems*, *11*(3-4), 271–295.

Robey, D., Vaverek, A., & Saunders, C. S. (1989). Social Structure and Electronic Communication: A Study of Computer Conferencing. In *Paper presented at the Hawaii International Conference on Information Systems, Hawaii.*

Scott, J. (2000). *Social network analysis: A handbook*. Sage.

Straub, D., Boudreau, M., & Gefen, D. (2004). Validation guidelines for IS positivist research. *Communications of the Association for Information Systems*, *13*, 380-427.

Toeman, Z. (1950). History of the Sociometric Movement in Headlines. In *Sociometry in France and the United States: A Symposium*.

Trier, M. (2008). Research Note–Towards Dynamic Visualization for Understanding Evolution of Digital Communication Networks. *Information Systems Research*, *19*(3), 335.

Tyler, J. R., Wilkinson, D. M., & Huberman, B. A. (2003). Email as spectroscopy: Automated discovery of community structure within organizations. In *Communities and Technologies: Proceedings of the First International Conference on Communities and Technologies, C&T 2003* (p. 81).

Valverde, S., Theraulaz, G., Gautrais, J., Fourcassie, V., & Sole, R. V. (2006). Self-organization patterns in wasp and open source communities. *IEEE Intelligent Systems*, *21*(2), 36–40.

Van de Ven, A. H., & Poole, M. S. (2005). Alternative approaches for studying organizational change. *Organization Studies*, *26*(9), 1377.

Wagstrom, P. A., Herbsleb, J. D., & Carley, K. M. (2005). A Social network approach to free/open source software simulation. In *Proceedings, First International Conference on Open Source Systems (IFIP 2.13)*. Genoa, Italy.

Wasko, M., Faraj, S., & Teigland, R. (2004). Collective action and knowledge contribution in electronic networks of practice. *Journal of the Association for Information Systems*, *5*(11-12), 493–513.

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

Watts, D. J. (2007). A twenty-first century science. *Nature*, *445*, 489.

Wiggins, A., Howison, J., & Crowston, K. (2008). Social dynamics of FLOSS team communication across channels. In *Proceedings of the Fourth International Conference on Open Source Software (IFIP 2.13)*. Milan, Italy.

Wu, J., Goh, K., & Tang, Q. (2007). Investigating Success of Open Source Software Projects: A Social Network Perspective. In *ICIS 2007. Proceedings of International Conference on Information Systems 2007*.

Xu, J., Gao, Y., Christley, S., & Madey, G. (2005). A Topological Analysis of the Open Souce Software Development Community. In *HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005.* (p. 198a). Retrieved from http://csdl.computer.org/comp/proceedings/hicss/2005/2268/07/22680198a.pdf.

Xu, Y. (., Zhang, C., Xue, L., & Yeo, L. L. (2008). Product Adoption in Online Social Network. In *ICIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2008/200.

Yeow, A., Johnson, S., & Faraj, S. (2006). Lurking: Legitimate or Illegitimate Peripheral Participation? In *ICIS 2006 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2006/62.

Zaheer, S., Albert, S., & Zaheer, A. (1999). Time scales and organizational theory. *Academy of Management Review*, *24*(4), 725-741.

Zhang, X., Venkatesh, V., & Huang, B. (2008). Students Interactions and Course Performance: Impacts of Online and Offline Networks. In *ICIS 2008 Proceedings*. Retrieved from http://aisel.aisnet.org/icis2008/215.