Analysing Social Networks Via the Internet

Bernie Hogan

I. Introduction

The purpose of this article is to introduce the reader to the history, concepts, measures and methods of social network analysis as applied to online information spaces. This is done through description as well as a sustained example using the online social news site Digg.com. Social network analysis is a rapidly expanding interdisciplinary paradigm , much of which is taking place with online data. As such, some concepts will only be addressed superficially, while others (such as positions, p* models and multilevel analysis) will be excluded entirely. The goal is to facilitate enough network literacy to begin a research project rather than provide a complete end-to-end solution. Social network analysis has emerged in the past half-century as a compelling complement to the standard toolkit of social science researchers. At its foundation is a belief that explanations for social organization are not to be found in innate drives or abstract forces. Instead we can look to the structure of relationships that constrain and enable interaction (Wellman, 1988) alongside the behaviors of agents that reproduce and alter these structures (Emirbayer & Mische, 1998). While this paradigm has been applied to fields as diverse as sexual contacts among adolescents (Bearman, Moody, & Stovel, 2004) and intravenous drug users (Koester, Glanz, & Baron, 2005), social network analysis is particularly well suited to understanding online interaction. There are two key facts about online interaction that make it particularly amenable to social network analysis - the nature of online interaction and the nature of digital information.

Online interaction is almost always social network-oriented. At its simplest, social networks refer to a series of nodes (such as people, organizations or web pages) and the specific links between two of these nodes. Hypertext (such as the World Wide Web) is an unstructured series of pages and links between pages. Communication online can be represented as a network of senders and recipients. Finally, relationships on social software sites constitute an obvious series of nodes (profiles) and links (friends). As Barry Wellman muses, "when computer networks link people as well as machines they become social networks" (1996, p. 214).

While digital information does not have to be network-oriented, this certainly facilitates the capture of network data. Granted, communication patterns and relationships were studied as networks long before the internet. However, collecting in-person data is time consuming and difficult; people are sometimes unclear of who is in their personal network (or how strong the tie is), and it is important to gather high response rates. These problems can be minimized online because information is digital and encoded merely through the act of sending a message or adding a friend to one's page. Also, there is virtually no marginal cost in making a perfect replica of the messages for analysis.

II. The fundamentals of social networks

A. Social networks in historical context

The roots of social network analysis are found in the mathematical study of graph theory (such as the work of Erdos, Harary and Rappaport) and empirical studies of social psychology (such Bott, Heider and Moreno)[1]. While the former group were charting various axioms between abstract nodes and lines, the latter found nodes and lines to be a sensible way to map concrete relationships between individuals. As the field matured in the latter half of the twentieth century these two groups converged on a series of metrics and methods to tease out underlying structures from complex empirical phenomena.

As a paradigm, network analysis began to mature in the 1970s. In 1969, Stanley Milgram published his Small World experiment, demonstrating the now colloquial "six degrees of separation" (Travers & Milgram, 1969). In 1973, Mark Granovetter's published the landmark "The Strength of Weak Ties" which showed empirically and theoretically how the logic of relationship formation led to clusters of individuals with common knowledge and important 'weak tie' links between these clusters (Granovetter, 1973). This decade also saw the first major personal network studies (Fischer, 1982; Wellman, 1979), an early, but definitive, statement on network metrics (Freeman, 1979), and the formation of two journals (*Social Networks* and *Connections*) and an academic society (The International Network of Social Network Analysts). The following two decades saw explosive growth in the number of studies that either alluded to or directly employed network analysis. This includes work on the interconnectedness of corporate boards (Mizruchi, 1982), the core discussion networks of Americans (McPherson, Smith-Lovin, & Brashears, 2006), the logic of diffusion (Rogers, 1995)) and even the social structure of nation states (Wallerstein, 1997).

Increasing computational power and the dawn of the Internet ushered in the second major shift in network thinking. By this point, physicists, biologists, and information scientists began contributing to a larger paradigm of 'network science'. Massive datasets could be gathered and analyzed in reasonable time frames. This led to maps and insights not only about a schoolyard or a few hundred personal networks, but about the billions of nodes on the World Wide Web. During this time, Watts and Strogatz showed that Milgram's small worlds could be found in movie actor networks and neural structures alike (Watts, 2002). Through an analysis of virtually the entire World Wide Web, Barabasi and Albert illustrated a major class of networks known as "scale-free networks" (1999), which

[1] See Freeman (2004) for a comprehensive review of the field from its inception to the present day

have been subsequently found in traffic patterns, DNA and online participation (Barabasi, 2003). Meanwhile, statisticians and social scientists have been busy working on a class of computationally expensive but extremely promising p* models that can decompose a messy and seemingly random social network into its simple and non-random underlying parts (Wasserman & Pattison, 1996).

This new era of network science is coming full circle with the advent of social software like MySpace and increased online participation generally. Social scientists can now analyze millions of email messages for general properties of communication or thousands of web log (or blog) links to understand the differing cultures of liberals and conservatives. Yet all of this analysis begins with the basic concept of the network.

B. What do we mean by a network?

Simply put, a network is a set of nodes (such as people, organizations, webpages, or nation states) and a set of relations (or ties) between these nodes. Each relation connects two of the nodes.[2] If the relation is directed, it is referred to as an *arc*, if it is undirected it is referred to as an *edge*. An email network, for example, is a directed network of senders and receivers. A social software network, on the other hand, is usually an undirected network of 'friends'. The premise behind this concept is that networks represent real structures that can constrain or enable social action. For example, if there is only one node connecting two groups, that node is particularly important in information transfer - the node can even manipulate information as it passes from one side to the other (Burt, 1992). Moreover, networks also represent intrinsically interesting structures - showing the overall connectivity of an email network can make the pattern of relationships far more intelligible to the owner of the inbox (Fisher, 2004).

Contrary to postmodern understandings of networks, such as Latour's "Actor Network Theory" (Callon & Law, 1997) or Deleuze and Guttari's "rhizome" (Deleuze & Guattari, 1987), social network analysis works best when all nodes are the same class of object. For example, since blogs can have more than one author, one would perform an analysis of blogs by only looking at blogs, and not blog authors or non-blog websites. In order to examine more than one type of object (such as bloggers and commenters), one can employ "two-mode analysis", which comes with its own set of considerations. Relations should also be of the same type. If one is linking email addresses, it is not advised to build a network where one relation can stand for "is in A's address book" and another relation stands for "sends email to A". While these assumptions simplify social relations to single types of nodes and relations, multiple networks can be superimposed to provide a more holistic picture of the social relationships between individuals.

Depending on the research question, one might require either a very large but superficial social network or a series of small but rich networks. The following section highlights three kinds of networks, and illustrates how they can be employed to address varying social issues. Sociological insights, both online and off, have come from all three.

[2]Or connects n nodes in the case of a hypergraph, although hypergraphs are rarely used in practice.

III. Network Types

A. Whole Networks

Whole networks describe the relationships within a clearly demarcated population. Online examples include an email distribution list, an entire social software community (such as all the users of MySpace), or all the people who work at a specific office, and their online communications. Whole networks are the most commonly used networks in social network analysis, but this is changing based on the practical demands of the researcher. Gathering all ties in an office is not particularly difficult, but getting a valid list of all ties on MySpace is practically impossible, as the list changes so rapidly during the process of data collection. Within a whole network, one asks questions of group structure, specific network member types and examines the networks for particularly prominent individuals.

Online records allow one to collect unobtrusive data on whole networks, such as all the postings in a newsgroup Webb2001. Work by Smith and colleagues at Microsoft research have illustrated that some newsgroups have particularly prominent individuals who answer questions altruistically, while other groups have a structure that that looks like a free-for-all discussion (Smith, 1999; Fisher, Smith, & Welser, 2006).

Whole networks can also be gathered actively. Traditionally, this is done with the use of a roster. One can then approach each member of the population and ask about his or her ties to everyone else on the roster. Each list is then a row in a matrix (often in a spreadsheet) which can be used to plot arcs from respondents to everyone else. Active data collection is useful when assessing subjective states and how individuals perceive the overall network, whereas unobtrusive data collection is useful when examining behavioral networks.

B. Personal networks

In whole network analysis, the goal is often to describe the characteristics of the network, and ask why certain individuals occupy a particular location in the network. (E.g., why do people always reply to him? Are there multiple subgroups in this network?) By contrast, personal network analysis is comparative in nature. One examines the differences in the size, shape and quality of a number of personal networks. These networks are commonly captured by sampling from a population. In this regard they are akin to traditional surveys as one would similarly want a representative (even stratified) random sample from a population. Each sampled case in this context is referred to as "ego", and the nodes connected to ego are referred to as "alters". One can either capture a star network (which is merely the ties to ego) or a full personal network (which includes the ties between alters).

One can unobtrusively collect personal networks in social software sites, communication and web pages. In each case

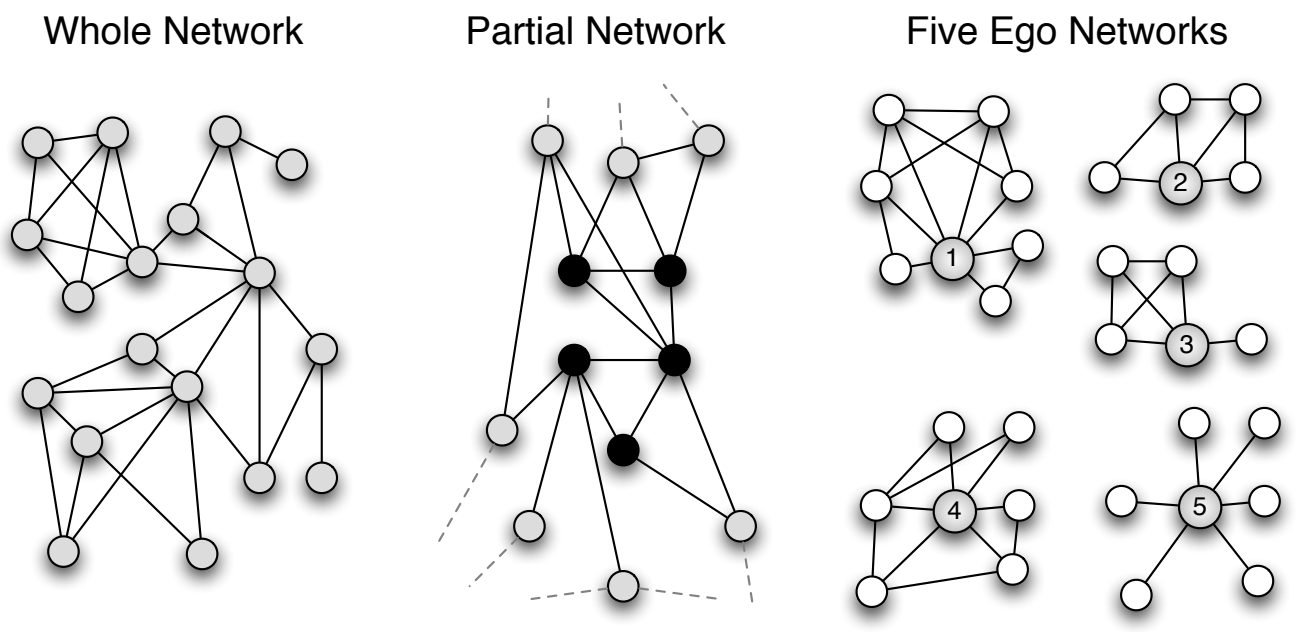

Fig. 1. Three network types.

one captures a list (such as a friend list) and then checks to see who on this list is also tied to each other.

Active collection of personal networks can make use of a number of pre-existing interview and survey techniques. The most prominent are the name generator (Hogan, Carrasco, & Wellman, 2007; Burt, 1984) and the position generator (Lin, Fu, & Hsung, 2001). Other techniques include the resource generator (Van Der Gaag & Snijders, 2005) and summation method (McCarty, Killworth, Bernard, Johnsen, & Shelley, 2000). With the exception of the name generator, these techniques are not designed to gather links between alters.

C. Partial networks

Partial networks are essentially the application of snowball sampling to relational data. These networks represent a compromise between the desire to capture a single large network and the fact that some networks are simply too massive to interpret meaningfully. One may start with a single web page or set of pages (known as the 'seed set') and look at the pages linked to the set, and then all the pages on each of these links. The sampling process stops when one has gathered a sufficient number of pages, when one has run out of new links, or when a certain criteria is met (such no more pages with more than 400 words).

Partial networks are a realistic solution for a great deal of network data collection on the web. One might not be able to gather data on all blogs, or on all individuals on MySpace, but one can build a network of relations that links together the personal networks of many individuals. Since it is easier to perform such a snowball technique on the web than it is in person, we can expect to see an increased number of researchers using partial networks to answer questions about social behaviour online. At present this is an active research domain often referred to as 'link analysis' (Thelwall, 2004; Park, 2003).

Because one is working outwards from a seed set, partial networks introduce concerns about generalizability. As Rothenberg notes snowball sampling in social networks, "[i]n the absence of a probability sample, the statistical superstructure collapses and, in principle, desirable statistical properties are not available to the investigator" (1995, p. 106). This constrains statistical generalizations but it does not inhibit descriptive analysis and inferences of this sample. Thus, generalizability may take place on a theoretical level, if not a statistical level. Moreover, one may capture most of the entire desired population through a well chosen seed set and follow all of the links that meet certain conditions (such as the presence of a particular set of keywords).

IV. Sources of online social networks

A. Email logs

There are myriad uses to email logs as a means to social network analysis. In the past they have been used to demonstrate differences between organizational structure and social structure (Adamic & Adar, 2005), differences in communication patterns in online and offline communications (Loch, Tyler, & Lukose, 2003; Haythornthwaite, 2005), and to help explain email overflow and the home work-boundary (Hogan & Fisher, 2006).

Unfortunately for the researcher, email is an overloaded technology (Whittiker & Sidner, 1996), which is to say the uses of email outnumber those for which the system is designed. It is a system of communication, a means for sharing files, a to-do list, a mass mailing outlet and a contact manager. All of these uses find their way into the same inbox. Before the researcher can analyze email as a social network many of these concerns have to be dealt with.

1) Email data capture: There are a number of ways to capture email data. These generally fall into 'server-side' and 'client-side' strategies.

Server-side: If one captures the entire email spool for a university domain (such as @utoronto.ca), one is assuming that this is the primary email for these individuals. This is more plausible in a workplace than for educational institutions. However, strict policies about deleting email have the potential to drive individuals away from their corporate accounts for anything other than official correspondence. That said, one can still gather a massive database and derive interesting results. For example, Kossinets and Watts (2006) analyzed millions of messages in a year long email spool. *Client-side:* Client-side data-capture involves the use either of email monitoring software or parsing scripts. The data is taken from a specific mail store and then parsed into a specific database base. Client-side data-capture is well suited to personal network analysis as one can capture the network on the client's computer and compare it to similarly captured networks. It is less than ideal for whole network analysis as one only has the mail that is seen by a particular address. The strategies below are weighted towards client-side strategies.

2) Building the network: Email networks are generally weighted directed networks. Arcs go from the sender to each of the receivers. Since messages are often sent to more than one person, and the recipients reply to everyone, there are often ties between the various email addresses in the mail store, and not just ties between ego (the owner of the mail store), and those people that send ego mail. The networks are weighted since people can send more than one message.

3) Email thresholds: When one is working from a server side mail spool, one may also have a complete list of all addresses associated with a particular domain. Thus, one can focus on messages between these individuals. However, if one does not limit the analysis to communication between specific addresses, one still has to differentiate relevant correspondence from spam and mailing lists. This can be accomplished through the use of structural metrics, whereby the network is trimmed down to specific messages and the network is created from these.

To trim the network down to meainingful correspondence, one can employ thresholds. One can threshold to 4 nested zones. Figure 2 provides a graphic representation of these thresholds (with levels 3 and 4 collapsed into one zone).

Zone 1: All messages in a mail store - This includes spam, distribution lists, broadcast announcements, etc... *Zone 2*: Ego's neighbourhood - Authors who have sent messages directly to ego, or received messages directly from ego. This eliminates messages to distribution lists that are forwarded to ego. It also eliminates messages bcc'd to ego and any distribution lists which ego has never sent a message. In practice, the loss of bcc'd messages is minimal as one can include such bcc'd mail if the sender also sends regular correspondence to ego. *Zone 3*: Ego's symmetric neighbourhood - There has to be a message from ego to alter and from alter to ego. This will eliminate all remaining distribution lists as they do not send to ego. It will also eliminate spam / junk mail / receipts and all other senders to which ego never replies.

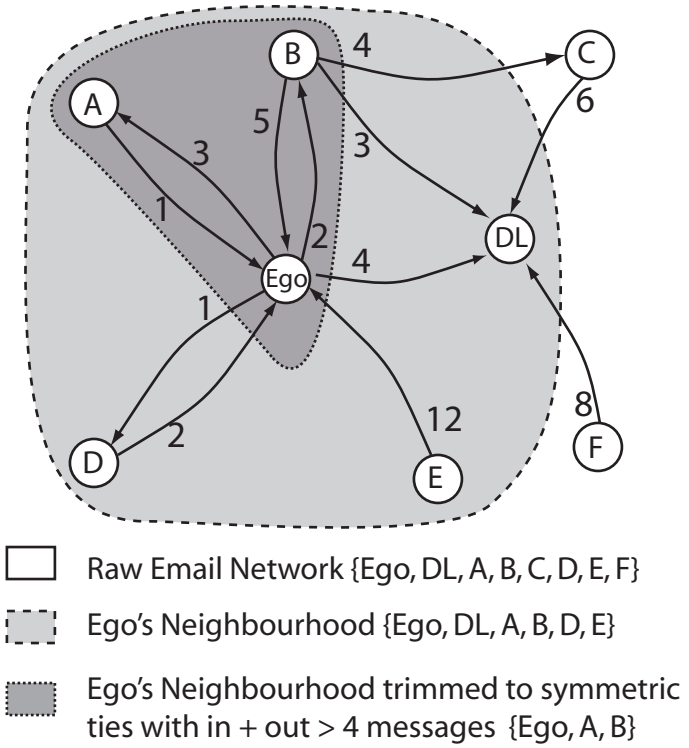

Fig. 2. The three zones of email. The outermost zone includes all email, such as DL distribution lists and spammers. The second zone includes only mail directly addressed to the respondent. The third zone is mail that is reciprocated, thus removing forwards, junk mail, spammers, etc...

Zone 4: Ego's thresholded neighbourhood - There has to be at least n messages from ego and (or) n messages from alter. This differentiates 'significant contacts' from fleeting / isolated correspondence. Adamic and Adar (2005) use 6 messages from and to ego. This author has used a more minimal approach in previous (unpublished) work at least one message from and to ego, and the sum of messages from and to must be 4 or greater. The actual amount to use varies by project, but should be justified substantively as presently there are few heuristics for an appropriate threshold.

4) Privacy issues with email stores: There are numerous potential strategies for safeguarding the privacy of email inboxes. However, these strategies can constrain the possible analyses done by a researcher, and so one must account for the trade-offs between user privacy and research questions. Collecting all information from an inbox may be ideal for a researcher but scare off potential respondents. Also, large studies of inboxes produce copious data that may be hard to manage. The follow strategies are available:

Removing message bodies: This will inhibit a textual analysis of mail, but it can cut down the size of the dataset dramatically. It is also very reassuring to the respondent. *Performing all text processing on the client-side:* If the research question must include a textual analysis of the message bodies, this can be done on the client's computer. What is

saved to the researcher's dataset is the outcome (such as the number of words, frequency of keywords, use of pronouns like 'he/she') rather than full message bodies. *Masking the addresses:* Technically speaking, email addresses are masked using a "hash" which encrypts the address so that it is represented by a string of unique characters, but cannot be decrypted. There are three levels of hash security. The first is a two-way hash, meaning the address is encoded but can be decoded with the appropriate key. This is important if the researcher wishes to attach additional attribute data to the email addresses (such as position in the company). The second is a one-way hash. This means the researcher or anyone else cannot determine the address once it has been hashed. The addresses can be hashed in the same way across email stores, thereby enabling the researcher to build a meta-network of many email inboxes but still maintain the confidentiality of any email address. The third is a salted one-way hash. Again the address cannot be decrypted, but the salting ensures that addresses are given mail-store specific hashes so the same address looks different if it comes from different mail stores. This means one can only do comparative ego-network analysis, but it is the most secure.

B. Blogs and other webpages

As the web is one giant network, it makes sense to approach it from a network perspective. In fact, doing so has led to captivating insights both for the web itself and for other areas of network science. One example is the now-famous scale-free distribution of Internet sites mentioned above (Barabasi & Albert, 1999). Another insight closer to conventional sociology comes from the linking patterns of liberal and conservative American blogs. Three separate studies have found that conservative blogs are denser and less centralized than liberal bloggers, and that liberals and conservatives online form two distinct sub-groups (Adamic & Glance, 2005; Ackland, 2005; Hargittai, Zehnder, & Gallo, 2006). The difference between these two subgroups can affect how fast ideas move through these blogs, how easy it is to achieve consensus of opinion and how easy it is to mobilize resources and people.

1) Methods of data capture and processing: To gather network data on the web, one can either use a pre-existing archive or gather new data using scrapers and spiders. Scrapers are automated computer scripts that take a web page and parse its content so it is useful as data. Spiders are a special class of scrapers that that follow links and collect information along the way. Data for spiders often comes from a "seed set" or a purposively selected set of pages and return a set of node-node pairs between this set and the pages they are linked to. One can then repeat this exercise from the newly gather pages until one runs out of links or fulfills a particular criteria (such as 2 steps out from the seed set). These pairs can then be assembled into a network dataset. Spidering is a common practice for search engines and for hypertext analysis. However, one must be careful only to follow appropriate links (rather than advertisements), to respect the site's spidering policy (usually contained in a robots.txt file such as www.google.com/robots.txt) and/or get explicit approval of the site maintainer. Schrenk (2007) offers extensive tutorials both on the practice and the pitfalls of spidering.

Datasets of the web also exist, and can be employed in the service of gathering network data. The most comprehensive is the Internet Archive from Alexa, which as of writing, is in the process of making its massive data archive available to researchers through Cornell University. In the meantime, researchers are encouraged to visit the Archive's "wayback machine" for an analysis of webpages at any given time dating back to 1996. Alexa also provides current metrics of the most popular sites. Nielsen Netratings also has a private database of web traffic, and its sister company Nielsen BuzzMetrics offers a publicly available database of blog traffic.

C. Social software

Social software programs are currently the most explicit representation of social networks on the Internet. People using these sites are encouraged to forge specific links, often titled 'friend', 'buddy' or 'associate'. The seminal social software site is Friendster, but its popularity has waned in favor of numerous others such as Facebook, MySpace and YouTube (Bausch & Han, 2006). The fact that these sites enable explicit dichotomous links between people will likely entice researchers to examine the structure of these online spaces. That said, early work in this area has been dogged by the fact that a social software friend is a qualitatively different character than an offline one (boyd, 2006).

In the world of social software, the term friend is synonymous with 'tie' or 'edge' in social network analysis. It denotes a relationship between two actors. However, when an individual has hundreds of friends in these spaces, the common emotional component of the term is hollowed out, and what remains is something much more insignificant and instrumental. As boyd notes, people become friends online:

> "[b]ecause they are actual friends, to be nice to people that you barely know... to look cool because that link has status, to keep up with someone's blog posts, bulletins or other such bits, to circumnavigate the "private" problem that you were forced to use [because] of your parents, as a substitute for bookmarking or favoriting [and because] it's easier to say yes than no if you're not sure." (boyd, 2006, p. 3).

Thus reasons for friendship are not merely different gradations of the same concept (as is the case with "closeness", a common subjective tie in personal network studies; Hogan et al., 2007; Burt, 1984; Granovetter, 1973). But these links actually stand for fundamentally different sorts of relations.

Links on social software sites can be scraped in much the same manner as links on other sites. However, the core difference is that for some of these sites one can only see the links between people up to four degrees away while on other sites one cannot view profiles and links without individual permission thereby leading to gaps in the network.

V. Analyzing networks through visualization and statistics - A primer

Once one has captured a network, one can ask specific questions about the network structure. This can either be

done within the confines of standard regression, qualitatively through the use of mapping, or within network analysis proper through the use of custom metrics. All three approaches are valid and used regularly. This paper will give an overview of the specific metrics developed for network analysis proper.

A. First steps: Mapping the network

A common first step in network analysis is visualization. These diagrams are an excellent tool for rapid pattern recognition. They can tell the viewer which nodes are proximate, for what reason and where to find dense clusters of activity. In addition to the examples found herein, the site Visual Complexity contains a massive array of network diagrams from the social sciences and beyond.[3]

Visualizations are common in social networks papers and *de rigueur* in presentations. However, it is possible to oversell the utility of these diagrams. They are interpretive tools, not unambiguous facts. In many cases the visuals have to be carefully massaged to accentuate the aspect of the graph that the researcher finds noteworthy, which is then reinforced by tabular data. As with the adage, "an unexamined life is not worth living", an uninterpreted sociogram is not worth presenting. Moreover, the conventional layouts can play into cognitive biases such as considering nodes placed in the center to be more prominent regardless of their real importance (McGrath, Blythe, & Krackhardt, 1997).

B. Considering the network as a whole: Density and clustering

Density is a measure of the number of edges within a graph divided by the maximum number of edges possible. It is a common measure and a useful first measure when comparing graphs of similar size or the same graph over time. That said, it can be misleading when comparing graphs of substantially different sizes. This leads to the perennial problem of how to say if a graph is sparse or dense. One solution is to calculate the density of a fictional network with nodes of an average degree, and compare that to the actual measure. Another is to only discuss a network's density in relation to the density of similar networks. However, in many other cases, researchers are not interested in density *per se*, but in how clustered the graph is.

Clustering coefficient is a measure that scales much more efficiently than density, and its use is increasing in the social sciences (Watts, 1999; Newman, 2003b; Kossinets, 2006). The local clustering coefficient is a measure of how well connected are the nodes around a given node. The clustering coefficient is the mean of the local clustering coefficient for all nodes in the graph. When the clustering coefficient is large it implies that a graph is highly clustered around a few nodes, when it is low it implies that the links in the graph are relatively evenly spread among all the nodes. Applying the clustering coefficient, Kossinets and Watts (2006) showed that the email network at a large American university did not get more clustered as the school year progressed. Individual networks got more or less clustered as people added new individuals or deleted old ties, but the overall clustering of the graph remained very consistent.

[3]http://www.visualcomplexity.com/

C. Considering the key players in the network: Centrality

Centrality scores describe the relative prominence of a given node in comparison to others. The average centrality score is also known as a centralization score, and indicates how strongly weighted the graph is towards a single node. There are three standard centrality measures: Degree centrality, closeness centrality and betweenness centrality. The reader is encouraged to consult Freeman (1979) for additional details and formulae.

Degree centrality expresses the number of links into and out of a given node divided by the total number of other nodes. A score of 1 indicates a node is connected to all others, while 0 indicates the node is an isolate. As many Internet networks are directed, there is also merit to looking at in- and out-degree centrality. High out-degree centrality indicates that a node is an "authority", they are the sort of site or person that can rapidly diffuse information to many individuals. High in-degree centrality indicates that a node is "celebrity" - they are the sort of site or person that many people will watch. Google.com has billions links out towards other sites. It is an authority. YouTube.com has relatively few links out towards other sites. However, many people link to Youtube or embed YouTube content in their own pages. It is a celebrity.

Closeness centrality expresses how close a node is to all other nodes in the network. As Freeman points out, it is a measure of efficiency. This is because a node that is closest to all nodes in the graph is best poised to receive a new innovation or infection. It is expressed formally as the number of other nodes divided by the sum of the distances between a node and all others in the graph. A score of one means that the node is connected to all others. It is likely that blog media sites such as Gizmodo.com and DailyKOS.com have very high closeness as they link to many sites, while many others link to them.

Betweenness centrality expresses how many shortest paths between all the members of a network include a given node. It is a measure of control. If a particular node has a high betweenness score that might suggest that it is the only link between many different parts of the network.

D. Considering the groups in the network: Cohesive subgroups and community detection

Halfway between overall network metrics and measures of individual prominence are community detection and cohesive subgroups methods. Cohesive subgroups metrics seek to find particularly dense pockets of links within an overall network whereas community detection algorithms seek to partition the network into sets that are themselves particularly dense relative to the overall network.

Common cohesive subgroup methods: The most typical measure is the clique which is a maximally complete subgroup (i.e. all nodes are connected). The clique concept can be relaxed as a k-plex whereby *most* of the nodes in a subgroup are connected (Seidman & Foster, 1978). While k-plexes work

well in theory, it is rarely seen in practice. Moody and White (2003) is a notable exception, which used a variant of k-plexes to assess the embeddedness of individuals in a network. Another measure is components, which are the number of connected subgraphs in a network. After removing ego from a personal network this measure shows how fragmented the network is from ego's point of view. *Community detection algorithms:* More common in the information sciences are community detection algorithms. The most popular is presently the Girvan-Newman algorithm (Girvan & Newman, 2002). Using this method one iteratively deletes edges of highest betweenness under the assumption that if there are two dense clusters any edge linking them would be the highest betweenness. However, there is a certain arbitrariness to this measure, and it does not work well under all conditions. Newman has come up with subsequent measures that have the potential to illustrate dense pockets in a graph, with greater reliability (2006). This area is still being actively explored and interested researchers are encouraged to examine the most recent literature.

E. Considering the attributes of network members: Homophily and assortativity

The above measures treated all nodes equally. Yet nodes, be they authors or pages have different attributes. In many cases one would like to know if nodes of like type link to each other - and do they link more frequently than by chance? Linking to similar nodes is referred to as homophily. For example, are bloggers of high-status likely to link to other high status bloggers or to low-degree blogs of their friends? McPherson, Smith-Lovin, and Cook (2001) offer an excellent overview of homophily and explain many of its subtleties. As they note, homophily is such a sure concept in social network analysis that it is not enough to ask if homophily exists in a social network, but to ponder what sort of homophily provides the logic for organizing the network.

Assortative mixing is a slightly different variant on homophily. Originally developed in the epidemiology literature (Gupta, Anderson, & May, 1989), this measure looks at whether individuals are likely to link to others who are similar, dissimilar or both. Newman (Newman, 2003a) gives a clear overview of the use of assortative mixing online. Interestingly, he shows that social networks are highly assorted in terms of degree. This means that people of high degree frequently link to people of high degree and low degree to those of low degree. This can be contrasted with networks such as the Internet infrastructure where servers of high degree link to computers of low degree.

F. Special notes for personal networks

All of the above mentioned network measures are designed for whole networks. That said, many will be informative measures for personal networks as well. The only thing to bear in mind is that some measures require the inclusion of ego, while others require ego's exclusion. Most specifically, closeness centrality and betweenness centrality rely on *geodesics* (shortest paths). Because ego usually connects everyone in the network it is best to exclude ego for these measures. McCarty (2002) gives an excellent overview of the specific application of many of these measures to personal networks alongside common best practices.

G. Advanced Network measures

More advanced techniques are outside the scope of this paper. The reader is encouraged to examine the recent volume on advances in network analysis by Carrington, Scott, and Wasserman (2005), the *Journal of Mathematical Sociology* and the journal *Social Networks* for additional techniques and information. Additionally, one may consult the recent compendium of papers from the physical and information sciences edited by Newman, Barabasi, and Watts (2006).

VI. Digg.com: An example social software site

The following example illustrates how to analyze Digg.com, a popular social news site. On Digg, users submit stories while others vote on these stories. The most popular stories of the day make it to the front page and receive upwards of millions of hits. Like many of these sites, Digg.com enables users to select friends. Stories that are voted on by friends are aggregated for the user.

One of the complaints of Digg.com is that the system is dominated by a particular group of individuals who set the agenda by reinforcing each others stories. This analysis suggests that this happens, but it is primarily benevolent social participation and diffusion rather than contrived manipulation. This claim is addressed below through a short analysis of Digg.com's top submitters.[4]

A. Capturing online data through scraping

Gathering a social network (or networks) online is quite a technical affair. Presently, only a few software packages exist to enable non-technical researchers to gather these links efficiently, and these packages are domain specific. As such, it is difficult to capture the desired data and one really has to collect the data through some automated means. There are two general strategies and both involve scripting.

The first is to use a domain-specific Application Program Interface (API). APIs are high-level interfaces to the database that renders html code. Through the use of an API, a user does not need to deal with potentially messy html, but can instead query a site for links. Publicly accessible APIs are available but not ubiquitous. Touchgraph, Inc. have released programs that interact with three major APIs - Amazon, Google and Facebook. However, Touchgraph only presents visualizations and not data. Recently, Digg.com released an API, although this example was produced beforehand.

In lieu of an API one can 'scrape' a page directly (as is done in this example). Here, the researcher downloads a page as html and then extracts the links from this page. The advantage

[4]Up until January 2007 Digg published a list of the top 1000 diggers, thereby creating an incentive for people to post (as they would move up in the rankings). This list was later removed, but it was still calculated by Christopher Frinke up until the time of writing. Special thanks to him for providing the sampling frame

to scraping is that users can also capture additional data on the pages which might be useful attribute data or explanatory variables, plus it works for any html page (but not for flash).

For this particular sample, I have chosen the top 910 diggers as of February 27, 2007. These individuals are the only ones to have 7 or more stories reach the front page of Digg. To access the friend page of these users one can go to http://digg.com/users/[user]/friends/list. These are the links out from the user. To access the links into the user, one can go to http://digg.com/users/[user]/friends/befriended. This is the sampling frame such that we can consider the whole network of these 910 submitters, but in order to create a complete list of their ties for analysis we have to build a network that is one-degree outwards from these ties. As such, this list does not generalize to all of Digg, but can be used as a theory building exercise to compare Digg's core network to the core network of other social news/bookmarking sites such as del.icio.us, Stumbleupon, Slashdot and Reddit.

To create a simple list of friends, one can count or mark down the friends listed on each page. However, this is tedious and prone to error. As such it makes sense to use a computer language to capture the page, parse it and store it as a datafile. This author's preference is to use python. This language has been called 'executable pseudo-code' because of its reputation as clear and concise. The following snippets illustrate some of the basic processes involved.[5]

If one has a list of names (in this case the top 910 diggers), they can be stored in an array:

```
namelist = [``top1",``top2",``top910"]
```

Then one can iterate through the array, and parse each page in turn:

```
site = ``http://www.digg.com/users/"
for i in namelist:
        p = urlretrieve(site + i + ``/friends/list")
        pagetext = p.read()
```

By viewing the sourcecode for a page one can see that all of the friend names are preceded by: ahref="users/ . Thus one can search the page for a 'regular expression' which includes the aforementioned text followed by characters, followed by "> . The following is a regular expression written in python:

```
fregex = re.compile("href=\"/users/\w*")
flist = fregex.findall(pagetext)
```

After cleaning up the list of names so that it excludes the user (which also fits the regular expression) and removes the surrounding characters (href, etc...), one has a list of friends. As a network this is like a star with the user at the center and points radiating outwards. To capture the links between those friends, one must repeat the above process and check each friend's page to see who is also a friend of the user. If one considers all of the user's friends as one set, then one must take the intersection of this set and the set of each friend's friends.

[5]The full code can be obtained from the author.

```
fset = set(friendlist)
for i in friendlist:
        #find all friends on i's page.
        #Just like above - call it flist_2
        fset_2 = set(flist_2)
        flinks.append((i, intersection(fset, fset_2))
```

There are a number of ways to scale up the process of collecting this information so that one does not need to scrape user pages multiple times. For example, one does not need to get the friend's friends for every user. One can combine the friend lists of all the users first, and then go find the links, this way, each friend page is visited only once rather than every time the friend is mentioned by a user. Other ways might be apparent to the researcher. In any case, the researcher should take pains to minimize the number of calls to a webpage as it might either arouse suspicion or unnecessarily slow down the site's server.

B. Analyzing this data

As mentioned above, one of the first steps in analysis is visualization. For the network of the top 910 users of digg, 433 are not connected to a giant component, whereas 477 are. Of the 433, less than 20 have any ties to other top submitters and most have no ties. Layout was done using GUESS (Adar, 2006). Figure 3 shows the giant component. The node size is the log of stories made popular, whereas the tint represents betweenness. Only the symmetric lines are shown. This diagram is laid out so that the number one contributor to Digg ('digitalgopher') is in the center. Each ring around digitalgopher is one step away from him. First, one can notice the intense linking around this top submitter, and second that those in the center have larger nodes (i.e. more popular stories) than those on the periphery.

As mentioned above, there have been suggestions that Digg is dominated by a few posters. Underlying this simple assertion is a host of network-oriented questions. How many posters? Are there factions / subgroups? Do the top posters reinforce each other? Does friending even make a difference? Using the scraped data from Digg, I have performed a series of nested linear regressions predicting to the distribution of the number of popular stories and the ratio of stories submitted to stories made popular. The raw number of stories is a power curve, as seen in figure 4. This necessitated a linear transformation of the variable as seen in the inset of figure 4. Because the models predict to the transformed variable the coefficients are not easily interpreted. One should pay greatest attention to the relative magnitude and significance rather than the value.

The models include eight variables, six of which are related to social network characteristics and the other two are measures of social participation.

- For both other top 910 users and non-top users:
  - Number of Symmetric ties (both friends and befriended)
  - Number of fans (befriended but not reciprocated)
  - Number of submitters watched (friended but not reciprocated)
- Profile data:
  - Number of stories submitted
  - Number of page views

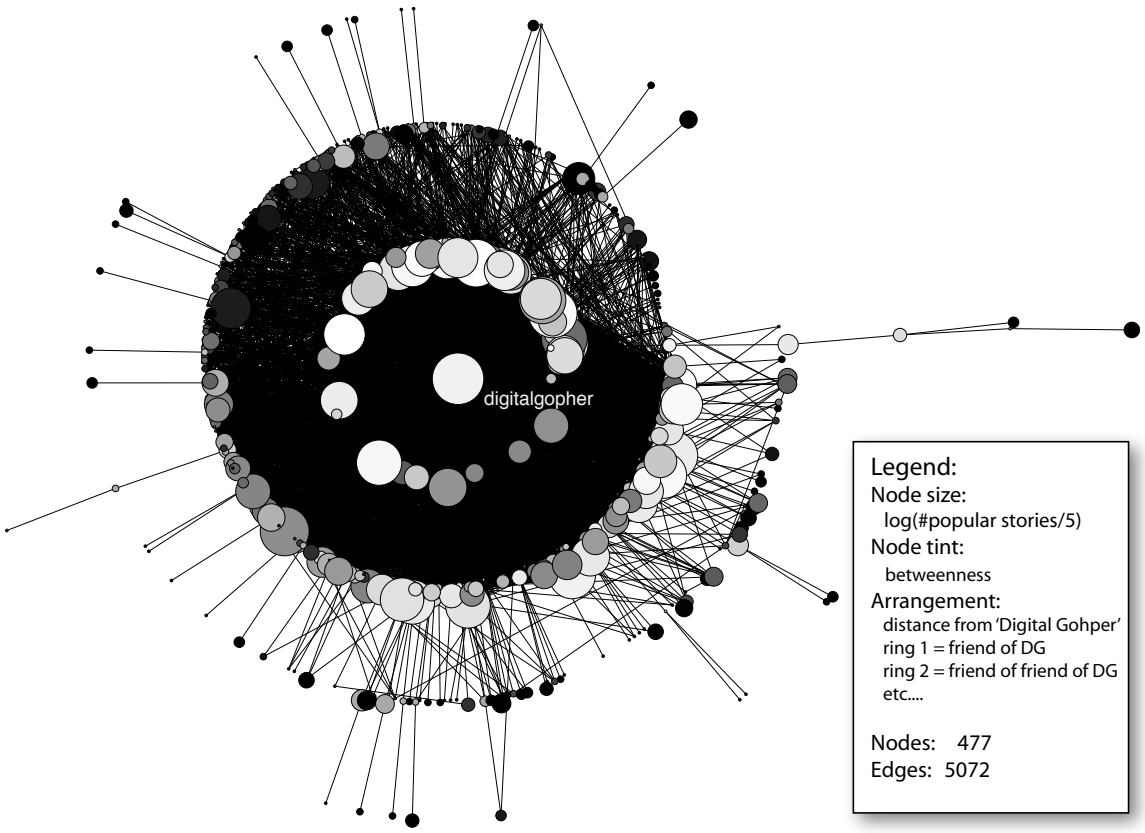

Fig. 3. A rendering of Digg.com's core 477 users. This network is the largest component among all Digg.com submitters who had 7 or more stories successfully make it to the front page. The radial layout is used to accentuate the relevance of the top poster, 'digitalgopher' who had 1007 stories make it to the top.

Table I shows the nested models predicting to the number of popular stories. Here we can see the benefits to a social understanding of online behavior. By merely counting and partitioning friends, we are able to explain forty percent of the variance in the number of popular stories, moreover, we can note that there is a nonlinear effect to friending. Having a fan among other top submitters carries more weight than having a non-top fan. Moreover, having numerous watched but unreciprocated ties actually has a negative effect.

The R^2 (the amount of variance explained by the independent variables) in the first model suggests that social network characteristics are intimately tied to the news stories that make it to the front page. The substantially lower R^2 in the second model suggests that while success is related to social structure, having friends does not guarantee than any story will make it to the top.

One must exercise much caution and subtlety in interpreting these models. Digg users accumulate both stories and friends. This model does not specify the causal arrow. For a longitudinal analysis, this network would need to be scraped at multiple points in time - a task outside the scope of this demonstration.

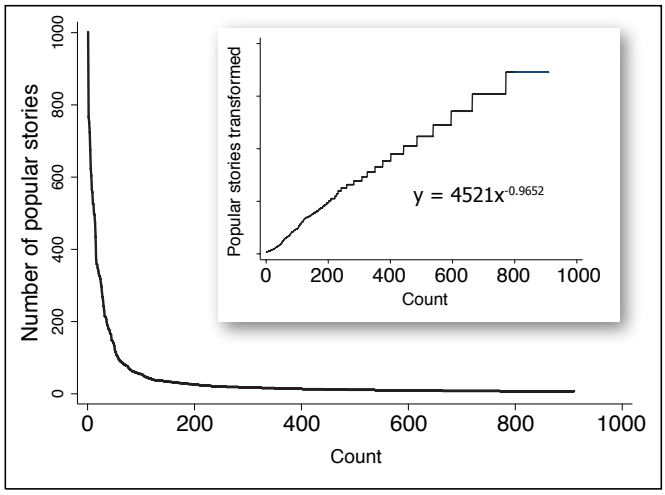

Fig. 4. The distribution of the number of stories made popular on Digg.com by user. The inset is the linearized transformation of this distribution.

C. How do online networks differ from offline ones?

The Internet used to be a "cyberspace" where "virtual communities" were linked by an "information superhighway". That is to say, it was considered as a separate sphere of activity apart from daily life. With increases in adoption and usability the Internet has become embedded in everyday life (Howard, 2004; Wellman & Haythornthwaite, 2002). It has become mundane as it has become ubiquitous. As numerous authors have shown, most of an individual's close online ties

TABLE I
OLS REGRESSION PREDICTING TO THE NUMBER OF STORIES MADE POPULAR AND THE RATIO OF STORIES MADE POPULAR BY NETWORK CHARACTERISTICS (NUMBER OF TIES IN, OUT AND MUTUAL).

	Number of Popular Stories				Ratio of stories made popular			
	Model 1		Model 2		Model 1		Model 2	
Fans (top)	8.37	***	7.66	***	0.05		0.32	***
Friends (top)	-3.17	**	-1.88		0.06		-0.21	*
Watched (top)	-0.65	+	-0.71	+	-0.04		-0.05	
Fans (others)	-0.42	***	-0.42	***	0.03	***	0.02	***
Friends (others)	-0.2		-0.66	**	0		0.08	***
Watched (others)	0.16		0.19		-0.03	**	-0.03	***
Submitted			0.09	***			-0.01	***
Dugg			0.01	***			>0.01	
Constant	-476.8	***	-479.27	***	16.72	***	18.07	***
Adjusted R^2	0.38	***	0.41	***	0.09	***	0.19	***

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, + $p < 0.10$

are really offline ties as well (Boase, Horrigan, Wellman, & Rainie, 2006; Baym, Zhang, & Lin, 2004; Wellman et al., 2006). This suggests that the clear dichotomy between online networks and offline ones is difficult to make. We are used to thinking of online data as a storehouse for robust objective relations such as 'sends at least 5 messages to' and offline networks as comprised of fuzzy subjective relations such as 'is close to'. However, there are a few considerations that make this simple dichotomy difficult:

Thresholding is still an arbitrary affair: While online networks indicate specific metrics, they do not let the researcher know which ones are the most relevant. *Precise behavioral metrics are also available offline:* Bernard, Killworth, and Sailer (1979) wrote a pivotal article on the difference between behavioural and cognitive networks long before the internet using logs from four different spheres of activity (ham radio operators, academics, a fraternity and an office). *With what media does one draw the line:* Is communication by telephone less related to email than instant messaging? In practice people use a host of media in concert to organize their lives and maintain their networks. Online media are a part of this ecology. Of course, all of the above points considered there are still some aspects of online networks that are difficult if not impossible to capture elsewhere.

Scope: The internet represents a massive store house of data. As (Newman et al., 2006) point out, this has led to the analysis of networks on a fundamentally different scale with datasets that often number in the millions of nodes, edges or cases. Also, at the personal network level, one can capture many acquaintances and weak ties that the individual might not have otherwise remembered in a self-reported study. *Passive data collection:* In most cases wiretapping is either illegal or infeasible, and capturing other communication relations beyond the level of a party or ethnography involves a great deal of work. By contrast, it is a straightforward task to see all of an individual's Live Journal friends, and only marginally more difficult to see the friends of each of these friends. *Novel structures and behaviors:* Online networks can reveal truly fascinating snapshots of human behaviour, some of which have no clear analog outside of the particular medium studied. From the idea of having (and negotiating) one's Top 8 friends to the presence of persistent altruists in newsgroups (Smith, 1999) and trolls in email lists (Herring, Job-Sluder, Scheckler, & Barab, 2002), online networks are a legitimate and compelling field of inquiry in their own right. To conclude this section, one can say that in general there is no hard distinction between online networks and offline ones. Some online networks and some offline networks share similar properties, such as whether they represent observed behavioural data or subjective states. What is different is the scope of data collection - which can now be massive and lead to the need for trimming and thresholding.

VII. SOFTWARE FOR SOCIAL NETWORK ANALYSIS

While it is not difficult to find examples of social networks via the Internet, it is still a nontrivial challenge to capture this data and work it into a usable form. Often data comes from a software package in one form and must be imported to a network analysis program in another form. As such, one should be prepared to massage the data accordingly. To clean the data, one can employ any number of scripting languages. Presently the most popular languages for this task are Python, Perl and Java.

At present, there are also a small number of pre-built programs available to academics. The Community Technologies Group at Microsoft is developing numerous tools such as SNARF, a email helper that builds a relational database of email and presents it to the user in novel ways and NetScan, a tool for querying the massive Usenet newsgroup archive[6]. The CASOS program at Carnegie Mellon offers numerous tools for network data retrieval and analysis[7]. Thelwall (2004) is not only an introduction to link analysis but also to SocSciBot which can perform numerous link spidering tasks. Likewise Schrenk (2007) has extensive online spidering examples and even a practice area for many complex spidering tasks. For the technically inclined, there are a number of software frameworks available as well to assist in visualization and analysis, including Viszter and prefuse(boyd & Heer, 2006),[8]

[6]http://research.microsoft.com/community/
[7]http://www.casos.cs.cmu.edu/computational_tools/tools.html
[8]http://prefuse.org/

JUNG (O'Madadhain, Fisher, White, & Boey, 2003),[9] and SNA for R (Butts, 2005).[10] In addition to these are the standard social network analysis packages, UCInet (Borgatti, Everett, & Freeman, 2006)[11] and Pajek (Nooy, Mrvar, & Batagelj, 2005).[12] Finally, numerous spiders exist for scraping online data and can be easily found through search engines.

One does not necessarily have to use any of these tools. Instead, it is possible to hand code relationships between individuals in a spreadsheet. However, the time it takes to hand code might be even greater than the time it takes to learn a language that parses an email header or the number of links on a webpage.

VIII. Conclusion

Social network analysis offers a powerful framework for detecting and interpreting social relationships online. They are accompanied by a host of analytic techniques ranging from simple centrality scores to sophisticated multilevel modeling. Yet gathering these networks is a time-intensive and challenging task. Online networks make this task somewhat easier through the use of passive networks (such as email stores and web pages), but the increase in efficiency leads to additional challenges about when to stop collecting, and what sorts of relations are substantively meaningful.

Overcoming these challenges takes patience, a good dose of technical skills with scripting languages or custom software and some trial and error. In return the results, as seen by many of the aforementioned studies, can inform our understanding of the interpersonal structures that affect online participation and online life in general. Yet, the techniques are relevant beyond the digital domain, hence the title 'via the Internet'. The discovered structures mirror and are a part of everyday life. It is not merely a gaze to distant shores, but a more crystallized view to the here and now.

IX. Acknowledgments

The author would like to thank the financial support of SSHRC, Bell University Labs and Intel's People and Practices Labs. The author has benefitted from the advice of the editors, Nigel Fielding, Ray Lee and Grank Blank as well as danah boyd, Danyel Fisher, Marc Smith, Ted Welser and Barry Wellman. Earlier versions of this paper were presented at the eSociety Handbook of Online Research Methods Colloquium in London, March 2007 and the 3rd Communities and Technologies conference, East Lansing, Michigan, June 2007. The author thanks the participants for their insightful feedback.

References

Ackland, R. (2005). *Mapping the u.s. political blogosphere: Are conservative bloggers more prominent?* Sydney.

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

Adamic, L., & Glance, N. (2005). The political blogosphere and the 2004 u.s. election: Divided they blog. *Working Paper*.

Adar, E. (2006). Guess: A language and interface for graph exploration. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 06)*.

Barabasi, A.-L. (2003). *Linked*. New York: The Penguin Group.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Bausch, S., & Han, L. (2006). *Social networking sites grow 47 percent, year over year, reaching 45 percent of web users, according to nielsen/netratings.*

Baym, N. K., Zhang, Y. B., & Lin. (2004). Social interactions across media. *New Media & Society*, *6*(3), 299–318.

Bearman, P., Moody, J., & Stovel, K. (2004, July). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, *110*(1), 44-91.

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.

Boase, J., Horrigan, J., Wellman, B., & Rainie, L. (2006). *Pew report: The strength of internet ties*. Washington, DC: Pew Internet and American Life Project.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2006). *Ucinet vi*. Harvard, MA: Analytictech.

boyd, d. (2006). Friends, friendsters and top 8: Writing community into being on social network sites. *First Monday*, *11*(12).

boyd, d., & Heer, J. (2006). *Profiles as conversation: Networked identity performance on friendster.* Kauai, HI: IEEE Computer Society.

Burt, R. (1984). Network items and the general social survey. *Social Networks*, *6*(4), 293–339.

Burt, R. (1992). *Structural holes: The structure of competition*. Cambridge, MA: Harvard University Press.

Butts, C. T. (2005). *Sna package: Tools for social network analysis*. Irvine, CA: University of California Irvine.

Callon, M., & Law, J. (1997). After the individual in society: Lessons on collectivity from science, technology and society. *Canadian Journal of Sociology-Cahiers Canadiens De Sociologie*, *22*(2), 165–182.

Carrington, P. J., Scott, J., & Wasserman, S. (Eds.). (2005). *Models and methods in social network analysis*. Cambridge, UK: Cambridge University Press.

Deleuze, G., & Guattari, F. (1987). *A thousand plateaus*. Minnesota, MN: University of Minnesota Press.

Emirbayer, M., & Mische, A. (1998). What is agency? *American Journal of Sociology*, *103*(4), 962–1023.

Fischer, C. (1982). *To dwell among friends*. Chicago: University of Chicago Press.

Fisher, D. (2004). *Social and temporal structures in everyday collaboration*. Unpublished doctoral dissertation, University of California, Irvine, Irvine, CA.

[9] http://jung.sourceforge.net/

[10] http://erzuli.ss.uci.edu/R.stuff/

[11] http://www.analytictech.com/ucinet/ucinet.htm

[12] http://vlado.fmf.uni-lj.si/pub/networks/pajek/ To note, the citation for this latter software refers to the excellent introductory network analysis text which guides the reader through Pajek while introducing many social network analysis concepts.

Fisher, D., Smith, M. A., & Welser, H. (2006). *You are who you talk to: Detecting roles in usenet newsgroups.* Kauai, HI: IEEE.

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

Freeman, L. C. (2004). *The development of social network analysis: A study in the sociology of science*. Vancouver, BC: Empirical Press.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology*, *78*, 1360–1380.

Gupta, S., Anderson, R., & May, R. M. (1989). Networks of sexual contacts: Implications for the pattern of spread of hiv. *AIDS*, *3*(12), 807–817.

Hargittai, E., Zehnder, S., & Gallo, J. (2006). *Mapping the political blogosphere: An analysis of large-scale online political discussions.*

Haythornthwaite, C. (2005). Social networks and internet connectivity effects. *Information, Communication & Society*, *8*(2), 125–147.

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). *Searching for safety online: Managing "trolling" in a feminist forum* (Tech. Rep. No. 02-03). Bloomington, IN: Indiana University. CSI Working Paper.

Hogan, B., Carrasco, J., & Wellman, B. (2007). Visualizing personal networks: Working with participant aided sociograms. *Field Methods*, *19*(2), 116-144.

Hogan, B., & Fisher, D. (2006). A scale for measuring email overload. *Microsoft Research Technical Report*, *TR-2006-65*, 1–3.

Howard, P. N. (2004). Embedded media: Who we know, what we know, and society online. In P. N. Howard & S. Jones (Eds.), *Society online: The internet in context* (pp. 1–27). Thousand Oaks, CA: Sage.

Koester, S., Glanz, J., & Baron, A. (2005, March). Drug sharing among heroin networks: Implications for hiv and hepatitis b and c prevention. *AIDS and Behavior*, *9*(1), 27-39.

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

Kossinets, G., & Watts, J., Duncan. (2006). Empirical analysis of an evolving social network. *Science*, *311*(5757), 88–90.

Lin, N., Fu, Y.-c., & Hsung, R.-M. (2001). The position generator: Measurement techniques or investigations of social capital. In N. Lin, K. Cook, & R. S. Burt (Eds.), *Social capital: Theory and research* (pp. 57–81). New York: Aldine De Gruyter.

Loch, C. H., Tyler, J. R., & Lukose, R. (2003). Conversational structure in email and face-to-face communication. *Working Paper*.

McCarty, C. (2002). Structure in personal networks. *Journal of Social Structure*, *3*.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2000). Comparing two methods for estimating network size. *Human Organization*, *60*(1), 28–39.

McGrath, C., Blythe, J., & Krackhardt, d. (1997). The effect of spatial arrangement on judgements and errors in interpreting graphs. *Social Networks*, *19*, 223–242.

McPherson, J. M., Smith-Lovin, L., & Brashears, M. (2006). Changes in core discussion networks over two decades. *American Sociological Review*, *71*(3), 353–375.

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

Mizruchi, M. S. (1982). *The corporate board network*. Thousand Oaks, CA: Sage.

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, *68*(1), 103–128.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Physical Review E*, *67, 026126*, 1–13.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Reviews*, *45*(2), 167–256.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*, 8577-8583.

Newman, M. E. J., Barabasi, A.-L., & Watts, D. (2006). *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.

Nooy, W. de, Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with pajek*. Cambridge, UK: Cambridge University Press.

O'Madadhain, J., Fisher, D., White, S., & Boey, Y. (2003). *The jung (java universal network/graph) framework*. Irvine, CA: UC Irvine.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of socail structure on the web. *Connections*, *25*(1), 49–61.

Rogers, E. (1995). *Diffusion of innovations, fourth edition*. New York: Free Press.

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

Schrenk, M. (2007). *Webbots, spiders, and screen scrapers*. San Francisco, CA: No Starch Press.

Seidman, S. B., & Foster, B. L. (1978). A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, *6*, 139–154.

Smith, M. A. (1999). Invisible crowds in cyberspace: Mapping the social structure of usenet. In M. A. Smith & P. Kollock (Eds.), *Communities in cyberspace* (pp. 195–219). London: Routledge.

Thelwall, M. (2004). *Link analysis: An information science approach*. Amsterdam: Elsevier.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, *32*(4).

Van Der Gaag, M. P. J., & Snijders, T. A. B. (2005). The resource generator: Social capital quantification with concrete items. *Social Networks*, *27*(1), 1–29.

Wallerstein, I. (1997). *The modern world system: Capitalist agriculture and the origins of the european world economy in the sixteenth century*. New York, NY: Academic Press.

Wasserman, S., & Pattison, P. E. (1996). Logit models and logistic regressions for social networks: I. an introduction to markov grahps and p*. *Psychometrika*, *61*, 401-425.

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

Watts, D. (2002). *Six degrees: The science of a connected age*. New York: W. W. Norton.

Wellman, B. (1979). The community question: The intimate networks of east yorkers. *American Journal of Sociology*, *84*(5), 1201–1233.

Wellman, B. (1988). The community question re-evaluated. In M. P. Smith (Ed.), *Power, community and the city* (pp. 81–107). New Brunswick, NJ: Transaction.

Wellman, B., & Haythornthwaite, C. (Eds.). (2002). *The internet in everyday life*. Oxford: Blackwell.

Wellman, B., Hogan, B., Berg, K., Boase, J., Carrasco, J. A., Cote, R., et al. (2006). Connected lives: The project. In P. Purcell (Ed.), *The networked neighborhood* (pp. 161–216). London: Springer.

Wellman, B., Salaff, J., Dimatrova, D., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, *22*, 213-238.

Whittiker, S., & Sidner, C. (1996). *Email overload: exploring personal information management of email.* ACM Press.