Using Information Networks to Study Social Behavior: An Appraisal

Bernie Hogan

Abstract

Social network analysis investigates relationships between people and information created by people. The field is presently in flux due to the increasing amount of available data and the concomitant interest in networks across many disciplines. This article reviews some of the recent advances in the field, such as p modeling and community detection algorithms alongside some of the societal transitions that facilitate these advances. The latter third of the article raises some issues for data engineers to consider given the state of the field. These issues focus mainly on querying and managing large and complex datasets, such as those commonly found through online scraping.*

1 Introduction

1.1 Social Networks and Digital Traces

This is a particularly exciting time to be a social network researcher; advances both within social network analysis and within society at large are making our work increasingly relevant. To be clear, by social networks we refer to associations between humans, or information created by humans. Presently, widespread use of information and communication technologies (ICTs) such as email, social software and cell phones have made possible the creation, distribution and aggregation of these relationships on an entirely different scale. But as we broaden our reach and seek ever more sophisticated answers within this paradigm, it is clear that we cannot do it alone. Quite frankly, there is much work to do and a substantial chunk of this work can get very technical, very fast. This article presents a brief overview of social network analysis as a field, where it is heading given current advances in the field, and where it may head as social scientists forge stronger links with computer scientists. As Gray and Szalay [1] note about science in general, there is now a data avalanche in the social sciences, and despite much of our previous expectations, we have indeed become data rich.

The use of digital media means that information can be copied and aggregated for analysis with very little extra work on the part of the respondent. Prior to the proliferation of digital media, gathering data about relations between ties was a long and expensive affair. Most data was self-reported, meaning an interviewer had to ask each respondent in turn. Not only did this place practical limits on the size of populations, but it introduced methodological issues of recall bias and concordance [2][3][1]. Some researchers have persuasively argued that individuals are good at recalling trends over time [4], but this is still not as robust as passive traces.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[1] Concordance refers to the extent to which two individuals will report a relationship of similar strength or frequency.

Digital traces are particularly significant artifacts for researchers. With email, for example, we are not left with empty postmarked envelopes or a caller ID from a telephone call - we are left with the names, dates, content and recipients. And because this interaction is digital, there is virtually no marginal cost in making a perfect replica of the messages for analysis. This level of fidelity and efficiency means that studies of social activity can (and have) scaled from previous ’huge’ data sets of a thousand people, to a data set of millions of messages. The consequence of this is to shift the burden from the practicality of collecting the data to the practicality of managing it.

These digital traces are probably more significant for social network studies than traditional social science methods. This is because much network analysis does not deal with sampling very well. The absence of a few key ties could completely alter the profile of a network [5][6]. While the probability that one would miss a key tie is low, their absence is significant. This confined network analysis to small whole groups, such as an office team or corporate board (with cases numbering from 10 up to a few hundred) or samples of ’personal networks’ where a person reports on the *perceived* ties between friends and family. Now, however, complete social networks such as the world wide web, or massive corporate communication networks are available. Even snowball sampling, which in face-to-face methods is costly and slow, can now be done simply by following hyperlinks from MySpace pages or a weblog (blog).

We can now scrape company inboxes and map the entire communication network for a large scale firm [7], plot links between ideologically charged blog discussions [8], or even map the entire email network for a university domain [9]. Asking all of the participants for this information directly is still possible, but non-obtrusive measures are so comprehensive and often efficient that many researchers start with trace data rather than consider it mere supplement. This situation has brought with it the usual issues about data quality (i.e. how do we capture this data, what data is worth keeping and what should be discarded), but social scientists deal with people, not star clusters or genetic maps. This leads to a host of additional questions that we are only now learning how to ask with trace data, let alone answer. Issues of data aggregation, privacy, contamination (e.g. the ”Hawthorne Effect”[2]), partial/incomplete data sources and ethical approval are relevant when dealing with human subjects. Moreover, many of these issues are located at the database level, where strict policies (such as one-way salted hashing of names) come into play and represent compromises between researchers and subjects. Finally, there are complex ontological issues that must be addressed - is a Facebook friend really an important friend [11]? What sort of additional data must be captured to distinguish ’friendsters’ from friends?

The remaining bulk of this article will present an overview of various advances in social network analysis, and exogenous advances that affect social network analysis. Throughout, issues of data management will either be in focus, or at least in the peripheral vision. Before concluding, I will also introduce some recent practical issues at the nexus of data engineering and social science methodologies.

2 A brief history of social network analysis

2.1 Early years

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”[12]. In 1973, Mark Granovetter 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 [13]. As he showed, job searchers are most successful not by looking to their strong ties (who have similar information as the searcher) but to their weak ties who link to other pockets of information. This decade also saw the first major personal network studies [14][15], an early, but definitive, statement on

[2] This effect refers to any situation where a general stimulus alters the subject’s behavior. It is particularly problematic because it is not a specific stimulus that creates a change, but mere attention[10]

network metrics [16], 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[19], the core discussion networks of Americans [20], the logic of diffusion,whether it's the diffusion of the latest music format or the spread of a new disease[21],and even the social structure of nation states [22].

2.2 The last decade

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 started contributing to a larger paradigm of 'network science'. Massive data sets could be gathered and analyzed in reasonable time frames, leading 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. This era produced advances which I will categorize in three sections: Endogenous advances - namely those advances coming from the field of social network analysis proper, parallel advances - those coming from related fields and scientific endeavors and exogenous advances - those coming from society outside academia but having a significant effect on the field.

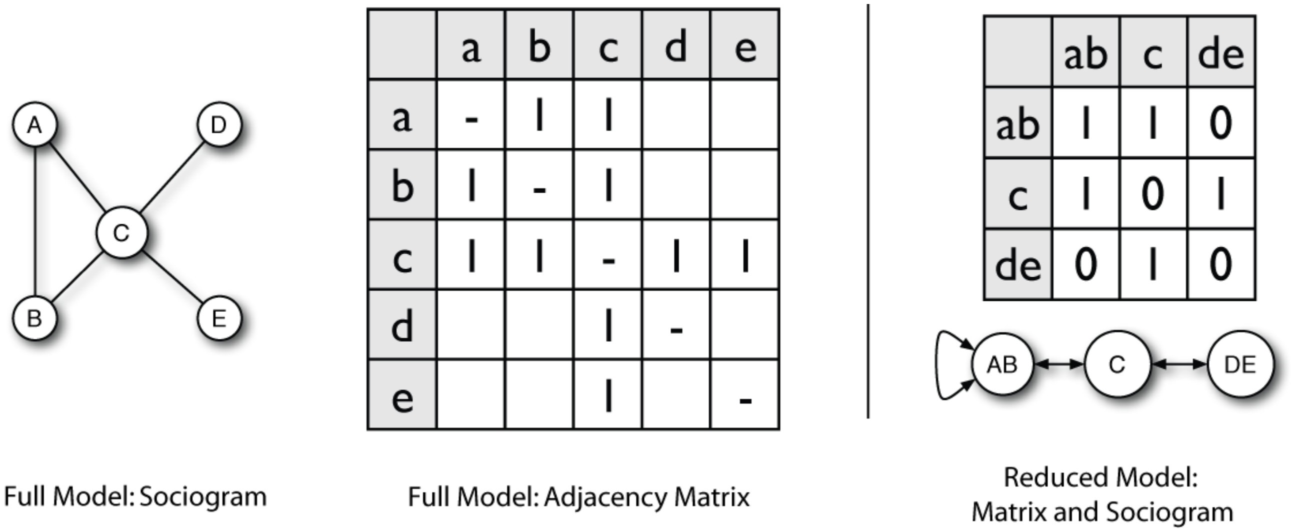
3 Endogenous advances

As was mentioned above, up until recent years, network analysis has traditionally been concerned with relatively small data sets or sampled populations. For example, numerous tutorials have been given on "Krackhart's High tech managers", a study of a whole network with a mere 21 respondents (see [23][24]). These specialized populations have led to analytic tools for such small networks. In general, the techniques either focus on relations, positions or motifs.

3.1 Relations

Relations are perhaps the most intuitive concept in social network analysis. Simply stated, these are methods for describing the importance of direct links between nodes in the network. These generally fall into two categories: examinations of the most prominent individuals and examinations of subgroups or communities of individuals. For measures of prominence, most research continues to use classic centrality measures (degree, closeness and betweenness centrality), even though two of them are computationally expensive. Specifically, betweenness centrality and closeness centrality both look for the shortest path between two nodes [16]. Since the creation of these measures a few other notable measures have cropped up for evaluating the prominence of an individual. These include eigenvector centrality which weights a nodes centrality score by the score of one's neighbors (thus one might be considered central not because they have many ties, but ties to popular people) [17]. While not a social network method, Google's PageRank is similar to Eigenvector centrality except it doesn't require complete information about the network [18]. Recent advances in this area have paid more attention to the robustness of these measures under varying conditions, than to the elaboration of new measures [5][22][6].

The relational analysis of subgroups is a field of active research. Here one looks at the ties between nodes to come up with subgroups or 'communities'. Early sociological work looked at the analysis of cliques (or complete subgraphs) whereas computer science solutions examined network flows. Questions about automatic detection of community structure are becoming increasingly relevant, as sites seek to categorize the structure of their users. In recent years, methods have moved beyond various max-flow min-cut solutions towards methods based on expected density and edge betweenness [25][26].



	a	b	c	d	e
a	-	1	1		
b	1	-	1		
c	1	1	-	1	1
d			1	-	
e			1		-

	ab	c	de
ab	1	1	0
c	1	0	1
de	0	1	0

Full Model: Sociogram

Full Model: Adjacency Matrix

Reduced Model:
Matrix and Sociogram

Figure 1: Example network - represented as a sociogram and a matrix. The right hand side shows the reduced matrix as a result of blocking the network.

3.2 Positions

Positions are a more lateral concept, stemming primarily from the work of Harrison White and his contemporaries in the early 1970s. If relations are concerned about *who* one is connected to, then positions are concerned with *how* individuals are connected. For example, the world system has a core-periphery structure, with most international finance moving through a few key cities. Those cities on the periphery might not be connected to each other, but they might be connected to the main cities in the same manner - hence they have similar positions in the system of global trade. Two nodes are considered structurally equivalent if they are connected to the same specific nodes in the same way. Two nodes are considered to be regularly equivalent if they are connected to any nodes in the same way. Partitioning a network into equivalent sets is referred to blockmodeling [27].

Figure 1 shows how nodes A and B are connected in equivalent ways, as are nodes D and E. Once the graph is partitioned using a blockmodeling algorithm, one can reduce the graph to its clusters, as is seen in the right-hand side of figure 1. To note, when you reduce a graph to its clusters you can have self-loops as nodes within each cluster can link to other nodes in that cluster.

As a technique, blockmodeling has a number of distinct strengths. Most particularly, this technique can find clusters of relationships which might otherwise be hidden. Community detection algorithms generally base their insights on higher connectivity within a subgraph than between subgraphs. Blocks, by contrast focus on the pattern of connectivity rather than the prevalence of connections. One non-sociological example of this is the work of Broder et al. [28] in characterizing the web based on positions (this includes a strongly connected core, an in-group, an out-group, tendrils, tubes and islands).

The last decade has seen two major improvements in blockmodeling. The first is generalized blockmodeling, enabling partitions of smaller, more precise block types, counts of ’errors’ for fitting models and predefined block structures based on attribute data[29]. The second is stochastic blockmodeling which compares a partition to partitions from similar networks to attain a probabilistic goodness-of-fit statistic [30]. In both cases, there are still a number of open questions. One is how to interpret partitions of massive networks, when it is unclear what will constitute an optimally fitting partition. The second is what to do with various edge types. When we consider an edge valued as 1 or 0, partitioning is straightforward, but blocking on data that is signed (positive, neutral or negative ties), sequential or weighted is still very experimental.

3.3 Motifs / Configurations

Motifs are small easily defined network structures which can be combined to create larger networks [31]. In social network analysis these are generally called configurations and represented in p* / Exponential Random Graph models [32]. There are only 3 dyadic configurations between two nodes (a symmetric tie, an asymmetric tie and no tie), but numerous configurations between the sixteen possible triadic motifs[33][30]. There are numerous theoretical reasons to believe that these configurations can be interpreted meaningfully, and that their distribution can inform the processes of tie formation and decay that characterize network dynamics [34].

Exponential Random Graph models refer to a family of probability distributions used to assess the likelihood of a particular network configurations appearing by chance. Testing these models is often done using either Maximum Pseudolikelihood Estimation or Monte Carlo Markov Chain Maximum Likelihood Estimation. While the former is far more efficient, it is often too conservative with standard errors and certain distributions and therefore should only be considered a proximate tool (Wasserman and Robins 2005). The latter is so computationally expensive that some models may not converge after days of iterations. Nevertheless, a robust model can illustrate quite clearly the relative importance of certain micro structures (such as 2-stars) on the emergent macro structures. Some of the most significant open problems in this area are related to the optimization of these methods and the use of robust estimation techniques. Because of the complexity of these dependency structures (e.g. requiring so many triads or four-cycles), and the fact that many of these problems appear to be NP-complete, advances in both computational algorithms and storage are welcome additions to the field.

4 Parallel Advances

4.1 Physicists, Biologists and Computer Scientists, Oh My!

Presently, there are far more individuals working in network data than social scientists and mathematicians. Biologists, physicists, are computer scientists are among the many disciplines that are finding network research particularly relevant to many of their research questions. Take the web, for example. It was created by humans and its linking structure is the result of many individual decisions. Yet, physicists have been characterizing the structure of the web as an emerging from many small decisions. In this vein, Watts and Strogatz showed that Milgram's small worlds (which they formally characterized as networks with high clustering and short paths between actors) could be found in movie actor networks and neural structures alike [35]. Through an analysis of virtually the entire World Wide Web, Barabasi and Albert [36] illustrated a major class of networks known as ”scale-free networks”, which have been subsequently found in traffic patterns, DNA and online participation [37]. All of these scale-free networks are based on the incredibly straightforward logic of preferential attachment: as a network grows, nodes with more links are likely to attract even more links, thus creating massive asymmetries between the few nodes with many links and the many nodes with few.

Biologists are finding that the human genome is an incredibly efficient means of encoding the data necessary for life. Genes do not work like on-off switches for direct effects, but work in combination, such that if two or more genes are present there is a particular phenotypical expression, but without all of them, there is none. This will have great consequences in the understanding of genetic diseases, as certain diseases depend on particular genes - but - other parts of the genome also depend on these focal genes.

As is probably evident to this audience, the use of network models in computer science has led to a number of very efficient solutions to problems. The 500-pound gorilla in the room is no doubt Google, who have used PageRank (and certainly a modified version thereof) to make search results more credible. Google succeeded as many people found their solution more useful than other patterns based on keywords or categories.

4.2 Visualization

Network researchers can find their data represented by numerous talented information visualization specialists. This task can become very technical very quickly, as people seek to represent an underlying structure merely by calculating parts of it in different ways. The Fructerman-Rheingold force directed algorithm gives the classic ’network-ish’ aesthetic, leading to insights about subgroups and clusters. By superimposing email networks over a company hierarchy (and measuring the overlap accordingly), Adamic and Adar show how the communication network conforms to the company hierarchy [7]. Representing mainly the dyads and the temporal structure (rather than the network structure) can also be insightful. Viegas and Smith’s newsgroup crowds visualization enables one to interpret long-term patterns of reciprocity in chat forums [38]. But perhaps the most striking as of late is Boyack’s representation of the ’Web of Science’. By analyzing co-citation patterns in ISI’s Citation Index, he has shown how disciplines as diverse as Organic Chemistry, Sociology and Computer Science are all part of an interconnected, but intelligible web of knowledge [39].

5 Exogenous advances

5.1 Advent of the internet

Perhaps the most obvious, and significant, recent change is the advent of the internet. By allowing us to communicate with digital bits, communication gets encoded merely through its use. That is to say, data sets start creating themselves not because the sociologist asked for them, but because they are part of the maintenance of an internet-oriented communication.

Even something as passive as shopping is grist for the sociological / computational mill. Krebs has been annually producing a series of network diagrams based on the purchasing habits of U.S. liberals and conservatives using only the Amazon API and his inFlow network software.[3]

5.2 Computational power

Both the visualization of networks and the calculation of structural metrics can be a time intensive process. Some important metrics , like betweenness, have only been reduced to *$O(n^2)$* time, while others are even *$O(n^3)$*. Alternative metrics (such as PageRank) help, but they are not a direct substitute given the theoretically meaningful nature of many of the metrics. With advances in computational power, we are beginning to play with our data instead of very rigidly and deductively grinding out a specific set of metrics.

One attempt to leverage this computational power is the ongoing NetWorkBench Cyberinfrastructures project at the University of Indiana[4]. This project is halfway between an algorithm repository and a web services framework for piping large data sets to a central supercomputer for processing.

5.3 Cultural changes

The world at large is becoming more responsive to social network analysis. There are numerous reasons for this. They include the presentation of clever and visually appealing maps [39], the advent of social software (which explicitly requires an individual to demarcate and maintain their network), and the inclusion of network ideas in common vernacular (”six degrees of separation”). As is the case with most sciences, there is still quite a disjuncture between scientific knowledge and lay understanding, but in this field people often ’get it’, and network literacy is, I would surmise, increasing.

[3] http://www.orgnet.com/divided.html

[4] http://nwb.slis.indiana.edu/

One interesting halfway point between rigorous scientific network knowledge and lay understanding is the new phenomenon of data mash-ups. Network data can be piped and displayed using Yahoo Pipes, IBM's Many Eyes and a host of small java applications. Insights from these simple interfaces may not be the most profound, but they stimulate discussion, and perhaps more importantly raise general network literacy. It is also the case that the interfaces for these tools represent significant improvements over scripting and they may pave the way to more interactive live data analysis in the future.

6 A cohesive programme of network science

6.1 Interdisciplinary collaboration

It is not too much to suggest that there is a emerging cohesive programme of network science, which has many foundations in sociology, but is by no means limited to it. There is presently no professional organization, but NetSci, the International Conference in Network Science is emerging as an interdisciplinary complement to the established social science-oriented International Sunbelt Social Networks Conference. Within this paradigm, the social scientists will most likely use their particular skill sets to frame questions, develop valid research designs and interpret figures reflexively. However, it is unlikely that many will graduate with the technical prowess necessary to implement a rich programme of network science. To complete this programme, we need to employ the aid of others with knowledge of unsupervised learning techniques, relational databases and scripting languages. Dealing with this data is becoming easier through the use of APIs. Most online social forums now have at least a basic mechanism for querying data. This includes sites like Yahoo, Google, Digg and Facebook. The big challenge for sociologists now is to bridge the gap between these lists of data and the usual rectangular data sets necessary for both network analysis and standard regression modeling.

6.2 Data issues within this programme

Accessing and analyzing quality data is an essential but often overlooked condition of possibility for the sorts of analysis described above. Presently, there are few sources for best practices regarding online and social network data. As such, there are still numerous open problems in the area of data quality and retrieval. Below are a list of particular issues that I suggest will become increasingly relevant.

Thresholding: How strong does a tie have to be for the relationship to be meaningful? Thresholding is the process of limiting ties between nodes to those that fulfill a specific threshold of activity. In an email network, one might consider a threshold of 6 messages between individuals in two months as sufficient to assume there is a 'strong tie'. While all authors agree that there is a need to somehow filter out irrelevant mail and spam from the analysis of network data, the specific scope conditions vary from project to project. By using a reciprocal threshold (i.e. a minimum of one message between two correspondents) one can ensure that there is at least some communication - but beyond that is a methodological "no man's land". The same can be said for links on web pages, replies in bulletin boards, calls on a cell phone, etc... Of course, one can keep all ties in the data set no matter how trivial, but then many models of diffusion, influence and community structure might not give answers that are particularly meaningful.

Algorithms for weighted graphs: Thresholding could partly be ameliorated with better algorithms for weighted graphs. There is some work on centrality algorithms for weighted graphs, but the field is still quite unsettled. Interpretations of these algorithms remain at the statistical, and not the substantive, level. One large challenge is the presence of exponential distributions for most measures - there's always a handful who either communicate more frequently, post more often, search more books, etc.

Robust formats for storing large rich data sets: Network analysis has as many data formats as there are programs (if not more). One of the emerging standards is GraphML. However, like all XML files, it contains a significant amount of supplementary text. For massive data sets this additional text scales linearly with an increase in nodes or edges leaving files many times larger than they need to be. Alternative formats such as Pajek are very lightweight but do not do as good a job of ensuring that certain data are associated with particular nodes or edges. Designing a halfway point between the austerity of Pajek and the clarity of GraphML with the ability to conveniently append data, particularly time sensitive data, will be a great improvement.

Better methods for slicing data (particularly temporal slices): Cleaning data is generally an unpleasant experience. For the expert programmer, the SQL queries can be tedious, and for the scripting-challenged, it is down right arduous. Presently, it is done by filtering the interactions which are then processed into networks, not by slicing the networks themselves (that is to say, they are sliced in SQL first, then exported to a network analysis package and then analyzed). A robust object model that maintains a sense of links over time should be able to dynamically slice the network without requiring the researcher to rebuild the network after every slice. Such techniques will enable more efficient sensitivity analyses for thresholds as well as facilitate more exploratory analysis of temporally-oriented network data (such as changes on Wikipedia).

Social network-oriented support in APIs: Support for networks is implicit in numerous APIs. However, this can be leveraged even more successfully (and reduce the number of queries to a site) by anticipating network data. For example, presently if one wishes to capture ”Joe's” network on facebook the steps are unnecessary clumsy. First, the program reads all of Joe's ties as a list. To find out who on Joe's list have befriended each other, the program has to then go to (almost) every other list, and compare these lists. By providing an option to query for mutual ties, one can reduce this querying from all of the friend lists of all of Joe's friends to a single list of user-user pairs. This puts an additional processing burden on the server, but it is a simple query for the server, rather than a series of steps for the user (and it reduces data and bandwidth).

People and relations as first class objects: Some frameworks will allow people to be considered first class objects. This allows individuals to be sorted, indexed and have numerous attributes, all of which are easily accessible through the program. However, a significantly harder technical challenge is to design a framework or language that will implement relations between individuals as first-class objects. Obviously, the dependencies between relations and people will make this challenging. But the end result will facilitate easier and perhaps faster querying of relations as well as enable more straightforward code and perhaps even simpler algorithms. It would certainly make network support in APIs dramatically easier to implement.

7 Conclusion

7.1 Concluding thoughts

The field of network analysis has been changing at a blistering rate. There is an influx of talented researchers from a host of disciplines. Network analysis is being done by MacArthur fellows and at the Santa Fe institute. It is featured in the Museum of Modern Art and on numerous blogs. It is an essential part of epidemiological modeling and our notions of social cohesion. Underlying all of this progress is an interest in a deceptively simple type of data that records and tracks links between entities. It has come a long way in the last half a century. With new data sources like the World Wide Web and new tools to examine this data more efficiently, it is likely that we will be busy for the next fifty years at least.

References

[1] J. Gray and A. Szalay, “Where the rubber meets the sky: Bridging the gap betweed databases and science,” *Bulletin of the Technical Committee on Data Engineering*, vol. 27, no. 4, pp. 3–11, December 2004.

[2] H. R. Bernard, P. D. Killworth, and L. Sailer, “Informant accuracy in social network data iv: A comparison of clique-level structure in behavioral and cognitive network data,” *Social Networks*, vol. 2, no. 3, pp. 191–218, 1979.

[3] j. adams and J. Moody, “To tell the truth: Measuring concordance in multiply reported network data,” *Social Networks*, vol. 29, no. 1, pp. 44–58, January 2007.

[4] L. C. Freeman, A. K. Romney, and S. C. Freeman, “Cognitive structure and informant accuracy,” *American Anthropologist*, vol. 89, no. 2, pp. 310–325, 1987.

[5] E. Costenbader and T. W. Valente, “The stability of centrality measures when networks are sampled,” *Social Networks*, vol. 25, no. 4, pp. 283–307, 2003.

[6] G. Kossinets, “Effects of missing data in social networks,” *Social Networks*, vol. 28, no. 3, pp. 247–268, July 2006.

[7] L. Adamic and E. Adar, “How to search a social network,” *Social Networks*, vol. 27, no. 3, pp. 187–203, 2005.

[8] L. Adamic and N. Glance, “The political blogosphere and the 2004 u.s. election: Divided they blog,” *Working Paper*, 2005.

[9] G. Kossinets and J. Watts, Duncan, “Empirical analysis of an evolving social network,” *Science*, vol. 311, no. 5757, pp. 88–90, 2006.

[10] E. Mayo, *The Human Problems of an Industrial Civilization*. New York, NY: MacMillan, 1933.

[11] d. boyd, “Friends, friendsters and top 8: Writing community into being on social network sites,” *First Monday*, vol. 11, no. 12, 2006.

[12] S. Milgram, “The small-world problem,” *Psychology Today*, vol. 1, no. 1, pp. 60–67, 1969.

[13] M. Granovetter, “The strength of weak ties,” *American Journal of Sociology*, vol. 78, pp. 1360–1380, 1973.

[14] C. Fischer, *To Dwell Among Friends*. Chicago: University of Chicago Press, 1982.

[15] B. Wellman, “The community question: The intimate networks of east yorkers,” *American Journal of Sociology*, vol. 84, no. 5, pp. 1201–1233, 1979.

[16] L. C. Freeman, “Centrality in social networks conceptual clarification,” *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[17] P. Bonacich, “Power and centrality: A family of measures,” *American Journal of Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[18] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” *Working Paper*, 1999.

[19] M. S. Mizruchi, *The Corporate Board Network*. Thousand Oaks, CA: Sage, 1982.

[20] J. M. McPherson, L. Smith-Lovin, and M. Brashears, “Changes in core discussion networks over two decades,” *American Sociological Review*, vol. 71, no. 3, pp. 353–375, 2006.

[21] E. Rogers, *Diffusion of Innovations, Fourth Edition*. New York: Free Press, 1995.

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

[23] D. Krackhardt, “Cognitive social structures,” *Social Networks*, vol. 9, no. 2, pp. 109–134, 1987.

[24] K. Faust and S. Wasserman, “Blockmodels: Interpretation and evaluation,” *Social Networks*, vol. 14, no. 1-2, pp. 5–61, 1992.

[25] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[26] M. E. J. Newman, “Modularity and community structure in networks,” *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577–8583, 2006.

[27] H. C. White, S. A. Boorman, and R. L. Breiger, “Social structure from multiple networks. i. blockmodels of roles and positions,” *American Journal of Sociology*, vol. 81, no. 4, pp. 730–780, 1976.

[28] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, “Graph structure in the web,” *Computer Networks*, vol. 33, no. 2, pp. 309–320, 2000.

[29] P. Doreian, V. Batageli, and A. Ferligoj, *Generalized Blockmodeling*, M. Granovetter, Ed. Cambridge, UK: Cambridge University Press, 2005.

[30] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge, UK: Cambridge University Press, 1994.

[31] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: Simple building blocks of complex networks,” *Science*, vol. 298, pp. 824–827, 2002.

[32] S. Wasserman and P. E. Pattison, “Logit models and logistic regressions for social networks: I. an introduction to markov grahps and p*,” *Psychometrika*, vol. 61, pp. 401–425, 1996.

[33] P. Holland and S. Leinhardt, “An exponential family of probability distributions for directed graphs,” *Journal of the American Statistical Association*, vol. 76, pp. 33–65, 1981.

[34] S. Wasserman and G. Robins, “An introduction to random graphs, dependence graphs and p*,” in *Models and Methods in Social Network Analysis*, P. J. Carrington, J. Scott, and S. Wasserman, Eds. Cambridge, UK: Cambridge University Press, 2005.

[35] D. Watts, *Six Degrees: The Science of a Connected Age*. New York: W. W. Norton, 2002.

[36] A.-L. Barabasi and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, pp. 509–512, 1999.

[37] A.-L. Barabasi, *Linked*. New York: The Penguin Group, 2003.

[38] *Newsgroup crowds and Author lines: Visualizing the Activity of Individuals in Conversational Cyberspaces*, 2004.

[39] K. Boyack and D. Klavans, “Scientific method: Relationships among scientific paradigms,” *Seed Magazine*, vol. 9, pp. 36–37, March 2007.