Audience, structure and authority in the weblog community

Cameron Marlow
MIT Media Laboratory
cameron@media.mit.edu

Abstract

The weblog medium, while fundamentally an innovation in personal publishing has also come to engender a new form of social interaction on the web: a massively distributed but completely connected conversation covering every imaginable topic of interest. A byproduct of this ongoing communication is the set of hyperlinks made between weblogs in the exchange of dialog, a form of social acknowledgement on the part of authors. This paper seeks to understand the social implications of linking in the community, drawing from the hyperlink citations collected by the Blogdex project over the past 3 years. Social network analysis is employed to describe the resulting social structure, and two measures of authority are explored: popularity, as measured by webloggers' public affiliations and influence measured by citation of each others writing. These metrics are evaluated with respect to each other and with the authority conferred by references in the popular press.

Introduction

The medium of weblogging differs very little from other forms of online publishing which have constituted the web since its beginnings. During its infancy, only a handful of authors were writing daily to websites identified as weblogs, but undoubtedly there were many thousands of others who updated their personal homepages nearly as frequently and in a similar writing style. What distinguishes weblogging from previous web media is the extent to which it is *social*, and one can say that the medium came into existence when the set of web journal writers recognized themselves as a community.

In the early days, there were only a handful of individuals who practiced the form, but with the addition of simple, personal publishing tools the community began an exponential growth that persists today. What was once a small family has matured into a burgeoning nation of millions including immense sub-communities around tools such as LiveJournal and DiaryLand. While some of these webloggers identify with the progenitors of the medium, others feel that their practice is distinct from that form. Regardless of affiliation, the nation of weblogging exists as such because every individual who takes part is connected to all others through the social ties of readership (Marlow 2002).

Every informal social system has its own order, constituted by the attribution of friendship, trust, and admiration between members. These various forms of social association give rise to higher-level organization, wherein individuals take on informal roles, such as opinion leadership, gatekeeper or maven. Within the weblog community, these positions are sought after by many authors, as they convey a sense of authority that increases readership and ties with other webloggers.

This paper is an exploration of the concept of authority as it is manifested in the community of webloggers. The Blogdex aggregator has been collecting data on the referential information contained within weblogs for the past 3 years, namely the hypertext links contained within webloggers' writing. Using the links between weblogs as a proxy to social structure, we construct a representation of the social networks of the

weblog community and employ social network analysis to describe the aggregate effects of status. Two measures of authority are explored: popularity, as measured by webloggers' public affiliations and influence measured by citation of each others writing. These metrics are evaluated with respect to each other and with the authority conferred by references in the popular press.

Background

Social network analysis (SNA) is a discipline of social science that seeks to explain social phenomena through a structural interpretation of human interaction both as a theory and a methodology (Wellman 1997). SNA assumes a basic graph representation where individuals (actors) are characterized by nodes, and the relationships (ties) they form with each other are edges between these nodes. This graph may be undirected, assuming that all social relationships are reciprocal, or directed, where each interaction describes a one-way association between two people. The *degree* of any node is defined as the number of associates that node has; in the case of undirected graphs, the degree is separated into in-degree (links in) and out-degree (links out).

Social scientists have characterized power as an actor's ability to control resources and information within the network, typically by exerting some type of structural advantage over other actors. Katz and Lazarsfeld made the observation that influence is controlled by a two-step flow of communication wherein opinion trickles up to opinion leaders and then back down to the rest of the population (Katz and Lazarsfeld 1955). Social network researchers have validated this theory not only for opinion but for information as well, showing that innovations, rumors, and beliefs tend to move from those marginal in a network, to the central figures, and back to the rest of the population (Valente 1995; Rogers 2003; Weimann 1982).

Opinion leaders are typically observed by their centrality to a given network, or by their ability to exercise large portions of the population in question by controlling the flow of information (Granovetter 1973, 1983). Freeman has described centrality in three different measures: *degree centrality*, or the total number of ties an actor has, *betweenness centrality*, or the probability that an individual lies on a path between any two nodes in the network, and *closeness centrality*, the extent to which an actor is close to all other actors (Freeman 1978). Since betweenness and closeness centrality require a complete description of the network and considerable computational resources for large data sets, degree centrality is typically used as a simple and efficient means of calculating authority.

Network analysis is well suited for the study of weblogs as many of the social relationships between weblog authors are explicitly stated in the form of hypertext links. Webloggers have posited their own interpretation of popularity and influence based on the number of links a weblog has in various link aggregation systems. Many webloggers use the total number of links to their site to evaluate the effectiveness of their writing.

A recent debate that has raised quite a bit of attention among webloggers is related to the distribution of these links within the community. Clay Shirky wrote a piece documenting the fact that a small group of webloggers had an enormous number of links to their site while the great majority only had a few (Shirky 2003). This distribution, he claimed, followed a power law distribution, a widely observed phenomenon popularized recently by Albert-László Barabási. Barabási has posited that power law distributions in self-organizing networks often arise from a process of preferential attachment, where nodes with higher degree are more likely to receive new links than less connected ones (Barabási 2002). Shirky assumed this model to claim that within the weblog ecosystem, the "rich get richer," and that the longer one has been an author, the more central they will be.

In the conversation surrounding Shirky's piece, many assumptions were made as to the nature of weblog authority. People assumed that every weblogger wanted to be recognized as an opinion leader, central to the network, and heavily linked by other webloggers. More importantly, it was assumed that links to a give weblog were somehow a proxy to the authority of that weblog. The remainder of this paper will explore this question in depth, namely what a link to a weblog means, the different types of social links that can occur, and how to understand authority in this social environment.

Defining Weblog Social Ties

Weblogs are a massively decentralized conversation where millions of authors write for their own audience; the conversation arises as webloggers read each other and are influenced by each others' thoughts. It is through the constant process of reading, writing and referencing that authors come to know each other at an informal level. Links are the social currency of this interaction, allowing webloggers to be aware of who is reading and commenting on their writings. A number of distinct subtypes of links have emerged within the medium, each one conveying a slightly different kind of social information:

Blogrolls

Nearly every weblog contains a list of other weblogs that the author reads regularly, termed the *blogroll*. This form evolved early in the development of the medium both as a type of social acknowledgement and as a navigational tool for readers to find other authors with similar interests. While the term pays homage to the practice of logrolling (the exchange of political favors and influence), a link within a blogroll indicates a general social awareness on behalf of the author. In some hosted services, such as LiveJournal and Xanga, the blogroll is a core part of the interaction, allowing users to be notified when their friends make a post or even to create a group dialog represented by the sum of the group's individual weblogs.

Permalinks

Weblogs are comprised of many individual entries, each covering a different interest or line of thinking. During the development of the first weblogging systems, it became apparent that it would be necessary to refer to specific posts instead of an entire weblog (Dash 2003); this feature allowed authors to have a sort of distributed conversation, where one post can respond to another on an entirely different weblog. These entry reference points are called *permalinks* and they are a core element of nearly every weblog today.

Comments

The most basic form of weblog social interaction is the *comment*, a reader-contributed reply to a specific post within the site. Comment systems are usually implemented as a chronologically ordered set of responses, much like web bulletin board systems. Depending on the amount of traffic a particular weblog might entertain, comments serve a range of usefulness; on extremely popular sites, the amount of response a post receives can render the comments long and unreadable, while on smaller sites a lack of any response can give the author and readers the sense that the site is generally unread. In between these two extremes, the comment serves as a simple and effective way for webloggers to interact with their readership.

Trackbacks

A recent feature of weblog tools is the *trackback*, an automatic communication that occurs when one weblog references another. If both weblogs are enabled with trackback functionality, a reference from a post on weblog *A* to another post on weblog *B* will update the post on *B* to contain a back-reference to the post on *A*. This automated referencing system gives authors and readers an awareness of who is discussing their content outside the comments on their site.

Design and Methodology

The Blogdex project was launched in 2001 as an effort to track the diffusion of information through the weblog community. The system currently tracks over 30,000 weblogs, updating its index when weblogs are changed, and keeping a record of each link made on a weblog along with the time the citation occurred. These links are aggregated into an index of the most rapidly diffusing content at any given point in time. These data are made available publicly on the project's homepage (Marlow 2001).

Of the four types of explicit social ties made by webloggers, blogrolls, permalinks, comments and Trackbacks, the first two types are available via the front page of a site while the latter two occur within the deep content of the archives. Since Blogdex collects every link on the front page of a weblog daily, the data for constructing a social network from either blogrolls or permalinks exists for the entirety of its operation. To extract the social network from the database, we first normalize the URLs of known weblogs to deal with potential duplicates, removing any leading "www" string and any trailing file name:

http://www.myweblog.com/index.html → myweblog.com

The resulting string is termed the *weblog ego* as it represents a unique key to all links that come from a particular site. These strings are then queried as substrings of links in the database; when a match is found, if the normalized form of the resulting URL is the same as the weblog ego, we assume this is a blogroll link. When the URL points to content other than the front page, we presume the link is a permalink. For example, if the ego in question is "myweblog.com," we would identify the following blogroll and permalinks as such:

http://www.myweblog.com/ → blogroll

http://myweblog.com/archives/001385.html → permalink

The social network is then represented by recording the weblog the link occurred on, the weblog the link pointed to, and the type of link (blogroll or permalink). Since one weblog can link to many permalinks on another given weblog, we associate a weight score with these links, the value of which is simply the number of permalinks from one weblog to another.

Blogrolls and permalinks represent two different types of social reference. A link made on a blogroll is made explicitly as a statement of social affiliation. By placing a link to another weblog, one assumes that the author either endorses that weblog, wishes to promote it, or claims to read it on a regular basis. Blogroll links are also updated much less frequently than the weblog itself, allowing these references to go out of date. Furthermore, once a link is made, there is a disincentive for removing it, as one feels guilty about taking traffic away from someone who was once an associate.

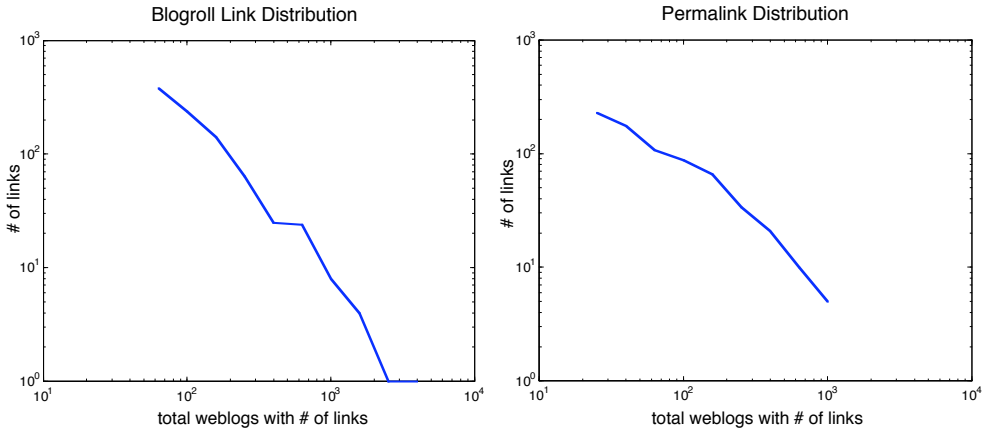
Permalinks, on the other hand, are made implicitly in the process of contextualizing a post. They reflect definite readership each time they are made, and through repeated links we can observe the strength of a tie increasing. Permalinks also represent influence to some degree, as the link signifies that some amount of thought has passed from one individual to another.

Because of the striking differences between these two types of links, we wish to explore the difference between them as measures of authority. To achieve this comparison, we will examine the top 1,000 weblogs by degree for both sets of network data, qualitatively analyzing the top weblogs for each ranking and looking quantitatively at the distribution of authority across these sites.

Another outside perspective on authority is the one conferred by citations in the popular press. Weblogs have achieved a high level of media attention since their inception, and many individual authors have been singled out for their content and opinions on the medium. To better understand how authority is perceived given our two models we will assume that journalists are looking to find central weblogs in the community. Searches for the terms "weblog," "web log" and "blog" will be executed on Lexis Nexis for all magazines and newspapers from 1995 to the present. URLs will be culled from these articles and matched against our database of known weblogs, and this mass media authority compared to our other findings.

Results

The network data collected by Blogdex contains 27,976 weblogs that have at least one inbound or outbound tie. The blogroll network data consists of 116,234 ties between these weblogs while the permalink data contains 285,970 ties. The higher density of permalink ties can be attributed to the fact that these relations accrue over time, and during the process of writing a weblog, while blogroll links require an explicit effort.



Figures 1 and 2 - Blogroll and Permalink distributions

In ranking these sites, the first observation is that most of the top sites are standard weblogs per se, they are weblog-like tools supported by communities of authors. Metafilter, Slashdot, Plastic, Fark and others depend upon tens of thousands of people for their content, while the rest of the list consists of sites operated by one or a small group of people. We can think of these systems as playing a role that a single human cannot, i.e. maintaining social ties with thousands of individuals. These sites play a crucial role in connecting large parts of the weblog network, resonating the important information that is diffusing through the community.

Figures 1 and 2 show the distribution of rank across the weblog social network for both blogroll links and permalinks. The data has been log-binned to remove noise from the tail and plotted on a log-log axis. The distribution looks like the characteristic power law seen by others for other weblog data sets. The first observation from these plots that the slope of the blogroll distribution is slightly steeper than the permalinks, suggesting that the falloff for authority in blogrolls is quicker, leaving a bigger separation between those at the tail of the curve. Other than this feature, the distributions are quite similar, and

one would might expect there to be a strong correlation between placement in both of these sets.

Rank	Blogroll Degree Rank		Permalink Degree Rank	
1	2581	metafilter.com	1322	boingboing.net
2	2434	slashdot.org	1270	diveintomark.org
3	2146	boingboing.net	1096	metafilter.com
4	1825	kottke.org	1073	slashdot.org
5	1604	instapundit.com	982	kottke.org
6	1527	scripting.com	976	weblog.siliconvalley.com/column/dangillmor
7	1307	evhead.com	956	instapundit.com
8	1220	andrewsullivan.com	828	andrewsullivan.com
9	1062	memepool.com	827	themorningnews.org
10	1007	doc.weblogs.com	826	rathergood.com
11	977	megnut.com	819	textism.com
12	961	littlegreenfootballs.com/weblog	683	denbeste.nu
13	899	diveintomark.org	626	doc.weblogs.com
14	880	littleyellowdifferent.com	625	asmallvictory.net
15	848	textism.com	582	rightwingnews.com
16	846	rebeccablood.net	577	microcontentnews.com
17	758	plasticbag.org	568	joi.ito.com
18	737	dashes.com/anil	560	buzzmachine.com
19	719	ftrain.com	553	waxy.org
20	714	plastic.com	522	a.wholelottanothing.org

Table 1 - Top authoritative sites by Blogroll and Permalink degree

Table 1 shows the top 20 sites by degree for both the blogroll and permalink network data sets. While some sites, such as Kottke, Boingboing, Andrewsullivan and Instapundit maintain high rank in both lists, as the list continues it becomes increasingly divergent. Many of the earlier "A-List" weblogs, such as Rebecca Blood, Scripting News, Megnut and LittleYellowDifferent do not place in the top 20 for permalinks, suggesting that while their names are recognized and placed on many blogrolls, they are not writing content as widely influential as those with high permalink rank. While it has been argued that "the rich get richer," as observed by the blogroll distribution, this is not true of those with high permalink rank. The power law distribution observed there contains many authors whose weblogs are half as old as those at the top of the blogrolls.

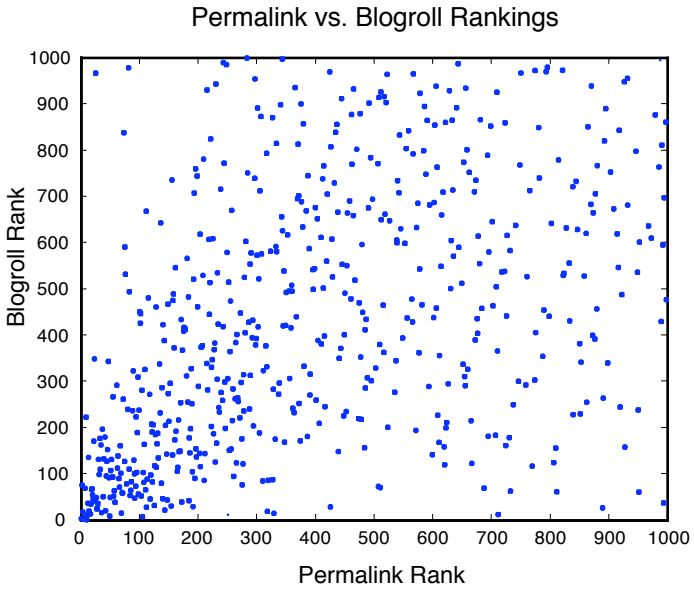


Figure 3 - Permalink and blogroll rank differential

Figure 3 demonstrates the differential between these data sets by plotting the rank in permalink versus the rank in blogroll. While the highest ranked data points tend to cluster around similar ranks, as soon as the rank passes 100 the correlation becomes much less apparent. This is further evidence that age is not the only factor in determining rank, otherwise these two data sets would be more tightly clustered around the line with slope 1. The fact that these two measures are not closely related implies that authority as measured by popularity cannot be interpreted as authority of influence.

Our queries made to Lexis Nexis returned 4,728 articles from both magazines and newspapers, of which 310 contained at least one known weblog URL. These documents yielded 545 total weblog URLs representing 212 unique sites. The twenty most cited weblogs are listed in Table 2 along with their rankings for both permalinks and blogroll citations.

Rank	News citations	Blogroll Rank	Permalink Rank	Site
1	24	9	8	andrewsullivan.com
2	21	5	7	instapundit.com
3	14	6	102	scripting.com
4	12	19	39	rebeccablood.net
5	11	1	3	metafilter.com
6	11	41	144	robotwisdom.com
7	10	7	46	evhead.com
8	7	11	708	memepool.com
9	6	14	66	megnut.com
10	6	147	166	bgbg.blogspot.com
11	5	22	23	plasticbag.org
12	5	42	18	buzzmachine.com
13	5	61	54	benhammersley.com
14	5	117	617	danbricklin.com/log
15	5	184	223	links.net
16	5	962	564	voxpolitics.com
17	4	28	151	camworld.com
18	4	38	2376	obscurestore.com
19	4	72	1052	loobylu.com
20	4	86	110	ntk.net

Table 2 - News citation rank

Among the top cited weblogs are a number of high ranking sites for both the blogroll and permalink rankings. One would expect this to be the case, as any journalist reading a weblogs at random should have a greater chance of running across a highly referenced site. This is especially true for blogrolls because the more weblogs one peruses at random, the more certain names will appear familiar as they are listed on the front pages of many sites. Any article about the phenomenon of weblogs would probably contain many of these since they are assumed to be highly popular, and thus representative.

But one cannot assume that journalists read weblogs just by chance and exclusively for stories about weblogs–they are often directed by specific stories and tips. It is in this case that those sites with higher permalink rank would be more likely to be seen. A story itself might be directly related to a post made on a weblog, with the press citation acting almost like a permalink in and of itself. Authority in the popular press thus seems to be a combination of popularity, in the case that a representative author is required, and influence, when weblogs themselves drive the media.

Discussion

The websites collected by Blogdex were originally culled from lists of weblogs available at the time of its creation, but has since become an opt-in service for any weblogs who wish to participate. One caveat of the system is that the data set includes a selection bias based on the individuals who choose to participate. Newer weblog aggregators such as Technorati operate on an opt-out policy that creates a much more comprehensive set. Blogdex is currently transitioning towards a similar model, but the results contained within this paper are constructed from data collected under the opt-in system.

In the process of migrating the system from opt-in to opt-out, the addition of new sites to the system has been halted, resulting in a data set that is missing new weblogs over the past 5 months. However, since we are focusing only on the top 1,000 ranked weblog in each category, missing these newer sites should not drastically affect our results. When a more complete data set is collected, unless we have missed entire sub-networks of authors, we expect our distributions to simply be shifted up by the factor of increase.

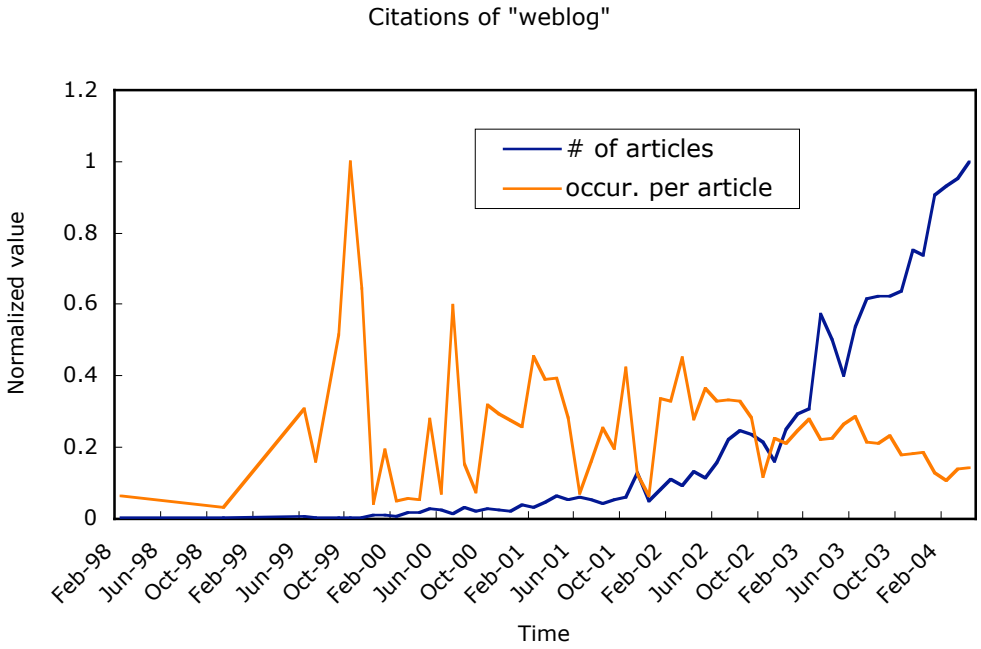


Figure 4 - Articles containing the term "weblog," "blog," or "web log," and average number of usages per article

Another probable explanation for the news citation rankings is that the role of weblogs in journalism is changing. Figure 4 shows the number of articles containing a weblog term versus the average number of times these terms are used per article. While the number of articles about weblogs shows exponential growth, the number of times the term is used per article has started to wane. This is a sign that the concept of the weblog has become part of our vernacular, and as such articles about weblogging alone are probably on the decline. More recent articles are likely to be influenced more by weblogs, and less about the medium itself.

Future Work

This work is by no means complete, and could be benefited from a number of future research directions. First, this work completely ignores the dynamic element of social networks; the suspicion that blogrolls reflect a “rich get richer” scenario more than

permalinks could be easily validated by examining the growth of degree for network members over time.

Second, not all permalinks are created equal as some weblog posts receive many orders of magnitude more traffic than others. In some cases, one post can define a weblog's permalink rank entirely, even though this attention is quickly lost. Looking at the distribution of permalink citations for each individual weblog may allow us to renormalize the data to avoid this phenomenon.

Finally, the data collected by Blogdex, while useful for the task of measuring authority, is not complete. Moving into an opt-out system and crawling weblogs at large will present a much more accurate picture of authority and influence.

Conclusion

The initial excitement over the weblog power law made many webloggers uncomfortable. How can a person get excited about a medium where attention is garnered by the number of weeks one has participated? Looking only at popularity by blogroll rank, it does appear that the "rich get richer," but another assessment of authority, permalinks, might be an equally good proxy to authority and a better measure of influence.

Barabási has noted that the growth of scale free networks is not only determined by the age of nodes, but also by the *node strength*, an undefined property related to a node's ability to acquire links. Permalink rank might be an accurate way of measuring node strength, and a better proxy to authority and influence at a given point in time.

Bibliography

Barabási, Albert-László. 2002. *Linked: The new science of networks*. Cambridge, MA: Perseus Publishing.

Dash, Anil. 2004. *Interview with Paul Bausch* 2003 [cited March 24 2004]. Available from http://www.sixapart.com/log/2003/09/interview_with_.shtml.

Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1 (3):215-239.

Granovetter, Mark. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78 (6):1360-1380.

———. 1983. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory* 1:201-233.

Katz, Elihu, and Paul F. Lazarsfeld. 1955. *Personal influence*. Glencoe, IL: Free Press.

Marlow, Cameron. 2004. *Blogdex* 2001 [cited May 2004]. Available from http://blogdex.net/.

———. 2002. Getting the Scoop: Social Networks for News Dissemination. Paper read at Sunbelt Social Network Conference XXII, at New Orleans, LA.

Rogers, Everett M. 2003. *Diffusion of innovations*. 5th ed. New York: Free Press.

Shirky, Clay. 2004. *Power laws, weblogs and inequality* 2003 [cited May 15 2004]. Available from http://www.shirky.com/writings/powerlaw_weblog.html.

Valente, Thomas W. 1995. *Network models of the diffusion of innovations*, *Quantitative methods in communication*. Cresskill, N.J.: Hampton Press.

Weimann, Gabriel. 1982. On the importance of marginality: One more step in the two-step flow of communication. *American Sociological Review* 47 (6):764-773.

Wellman, Barry. 1997. Structural analysis: From method and metaphor to theory and substance. In *Social structures: A network approach*, edited by B. Wellman and S. D. Berkowitz. Greenwich, CT: JAI Press.