Conversations in the Blogosphere: An Analysis "From the Bottom Up"

Susan C. Herring, Inna Kouper, John C. Paolillo, Lois Ann Scheidt,
Michael Tyworth, Peter Welsch, Elijah Wright, and Ning Yu
School of Library and Information Science
Indiana University Bloomington
{herring,inkouper,paolillo,lscheidt,mtyworth,pwelsch,ellwrigh,nyu}@indiana.edu

Abstract

The "blogosphere" has been claimed to be a densely interconnected conversation, with bloggers linking to other bloggers, referring to them in their entries, and posting comments on each other's blogs. Most such characterizations have privileged a subset of popular blogs, known as the 'A-list.' This study empirically investigates the extent to which, and in what patterns, blogs are interconnected, taking as its point of departure randomly-selected blogs. Quantitative social network analysis, visualization of link patterns, and qualitative analysis of references and comments in pairs of reciprocally-linked blogs show that A-list blogs are overrepresented and central in the network, although other groupings of blogs are more densely interconnected. At the same time, a majority of blogs link sparsely or not at all to other blogs in the sample, suggesting that the blogosphere is partially interconnected and sporadically conversational.

1. Introduction

Weblogs (blogs)—web-based journals in which entries are displayed in reverse chronological sequence—are a recent addition to the repertoire of computer-mediated communication (CMC) technologies through which people can socialize online. Overt connections among weblogs include hyperlinks from one blog to another, mentions of other blogs and bloggers in entries posted to weblogs, and comments posted in response to other bloggers' entries.[1] Some writers and webloggers have suggested that these activities are conversational in nature. As Marlow (2004), for example, writes, "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*" [emphasis added]. Blood (2002) goes further: "I would go so far as to say that if you are not linking to your primary material when you refer to it—especially when in disagreement—[…] you are not keeping a weblog" (pp. 18-19). However, despite the existence of such claims, the "conversational" nature of blogging has yet to be empirically investigated. In particular, the claim that linking to other blogs constitutes a form of conversational interaction is not self-evident, and merits critical scrutiny. We do not generally think of web sites that contain links to other web sites as engaging in "conversation" (but cf. Coste, 2000). Why, then, should we characterize weblogs as conversing with one another when they do the same?

A second issue arises from Marlow's claim that conversation among blogs is "completely connected"—and similarly, from Blood's (2002) requirement that a blog link to other blogs in order to be characterized as a blog—which assume that some form of interlinking is frequent, if not absolute. Most discourse about weblogs focuses on a particular sub-type of blog, the so-called filter blog, which links to ("filters") web content. However, a recent study found that only 12.6% of currently active weblogs are filters, and 48.8% contain no links to other weblogs at all (Herring, Scheidt, Bonus & Wright, 2004). Most discussions of the 'blogosphere'—the universe of all weblogs—to date focus on an elite minority of blogs (Herring, Kouper, Scheidt & Wright, 2004), the most popular of which are sometimes referred to as "A-list".[2] "A-list" blogs—those that are most widely read, cited in the mass media, and receive the most inbound links from other blogs—are predominantly filter-type blogs, often with a political focus. The A-list appears at the core of most characterizations of the blogosphere (Blood, 2002; Delwiche, 2004; Marlow, 2004; Park, 2004). Given that a majority of blogs are not filters, however, and may contain few or no links, the question arises as to how representative such characterizations are of the blogosphere as a whole. Would the blogosphere appear to be a "completely connected conversation" from the perspective of a random, ordinary blog?

[1] Bloggers may also communicate with one another behind the scenes via email, instant messenger, and other forms of CMC, but this activity is beyond the scope of the present investigation, unless it is referred to in the blogs themselves.

[2] In this paper, we use 'A-list' as an operational shorthand to refer to the most popular blogs as determined by number of inbound links, a practice that also underlies the identification of "top-100" lists of blogs posted to the web. Some bloggers dislike this term, perhaps because of its implied elitism.

In this paper, we adopt precisely this perspective in an attempt to characterize empirically the nature and degree of interconnectedness of the blogosphere "from the bottom up," that is, taking as our point of departure randomly-selected "ordinary" blogs. We employ methods of social network analysis to identify link-based connections across blogs and to interpret (quantitatively and visually) the resultant patterns. We then turn to the micro-analytical level to consider (qualitatively) whether reciprocal linking correlates with more familiar manifestations of computer-mediated conversation, such as posting comments in response to others' entries. The results show that, consistent with our previous findings, most blogs contain no links or comments. Of those that do, filter blogs contain more links, and link preferentially to A-list blogs, which are consequently overrepresented in the network. At the same time, unexpected clusters of non A-list blogs emerge that are highly reciprocally-linked, forming topic-based communities (cf. Kumar, Novak, Raghavan, & Tomkins, 1999). Direct manifestations of online conversation are found in both A-list and non A-list blogs, but vary in extent and nature according to blog topic and gender of writer. On the basis of this evidence, we conclude that the blogosphere is partially interconnected, and that blog conversations, while occasionally intense, are the exception rather than the rule.

2. Background

2.1. The social network approach

Social networks consist of people connected through various social relationships or exchanges. Social network analysis (SNA) has developed over the past decades as a set of methods for mapping and analyzing relations among people, organizations, and objects. Grounded in graph and system theories, this approach has proven to be a powerful tool for studying networks in physical and social worlds, including on the web (Berkowitz, 1982; Adamic, 1999; Albert & Barabási, 2002).

SNA focuses on relations and ties in studying actors' behavior and attitudes (Degenne & Forsé, 1999; Wasserman & Faust, 1994). Thus the positions of actors within a network and the strength of ties between them become critically important (Knoke & Kuklinski, 1982). *Social position* can be evaluated by finding the centrality of a node identified through a number of connections among network members. Such measures are used to characterize degrees of influence, prominence and importance of certain members (Faust, 1997). *Tie strength* mostly involves closeness of bond. There is general agreement that strong ties contribute to intensive resource exchange and close communities, whereas weak ties provide integration of relatively separated social groups into larger social networks (Granovetter, 1973; Wellman & Wortley, 1990). A third network concept relevant to the present research is *small worlds*. The small world phenomenon, initially described by Milgram (1967), is a condition of interconnectedness in the real world whereby two random U.S. citizens were found to be connected by an average of six acquaintances.

The Internet poses new questions about the nature of social networks and opens new perspectives for social network analysis (Garton, Haythornthwaite, & Wellman, 1999; Wellman, 2001). In particular, the hypertextual structure of the web makes linking explicit. As a consequence, as Jackson (1997) points out, social network analysis is ideally suited to the web environment. In addition to being directly observable, hyperlinks are easy to map and analyze with specialized software programs.

A number of studies have analyzed patterns of linking on the World Wide Web. Adamic (1999) showed that the web is a small world network. Gibson, Kleinberg, and Raghavan (1998) developed a technique to identify hyperlinked communities in Web environments, which includes the identification of hubs (strong central points with high numbers of outbound links) and authorities (highly-referenced pages).

Most recently, social network analysis methods have begun to be applied to weblogs. Merelo-Guervos, Prieto, Rateb, and Tricas (in press) map the Blogalia weblog hosting site, which has around 200 members, using a neural-net like visualization method which reveals community features. Kumar, Novak, Raghavan, and Tomkins (2003) observe and model temporally-concentrated bursts of connectivity within blog communities over time, concluding that 'blogspace' has been expanding rapidly since the end of 2001, "not just in metrics of scale, but also in metrics of community structure and connectedness" (p.1). Adar, Zhang, Adamic, and Lukose (2004) identify blogs that initiate "information epidemics" and visualize the paths specific infections take through blogspace (see also Gruhl, Guha, Liben-Nowell, & Tomkins, 2004). Marlow (2004) uses social network analysis to identify "authoritative" blog authors, and compares them with measures of opinion leadership and authority in the popular press. As in Delwiche's (2004) study, the most authoritative authors turn out to be members of the A-list.

2.2. Weblog analysis tools

Recent years have also seen the rise of automated services intended to track blogs and provide indicators of their popularity and influence. These can be grouped roughly into three categories: 1) services that track and allow subscription to blogs and RSS[3] feeds; 2) services that rank blogs according to popularity, usually deter-

[3] Really Simple Syndication (also Rich Site Summary or RDF Site Summary) is a format for syndicating news and the content of news-like sites (http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html).

mined according to number of inbound links; and 3) services that track the spread of ideas across the blogosphere. Although not primarily intended as research tools in most cases, these services collect data and operationalize weblogs in ways that are relevant to the problems faced by researchers seeking to characterize the blogosphere empirically, as well as provide baseline data for purposes of comparison.

Currently the most comprehensive blog tracking services are the blo.gs (http://blo.gs) and National Institute for Technology and Liberal Education (NITLE) blog census (http://www.blogcensus.net/) sites. Blo.gs gathers data via an hourly check of updated blogs from antville.org, blogger.com, pitas.com, and weblogs.com, as well as from users who automatically or manually ping the service. The NITLE blog census continuously spiders its set of known blogs looking for links to new pages, archiving the results every two weeks. The system's "crawl queue" is seeded with the contents of nine blog studies, 16 blog directory services, six "recently updated" tracking sites like Weblogs.com and the Blogger updates page, along with individual submissions. Both sites define blogs broadly as regularly updated personal websites with posts that appear in reverse chronological order, a definition that extends to "community weblog" sites such as Slashdot and Metafilter. As of September 15, 2004, blo.gs was tracking approximately 2.7 million weblogs, and NITLE was tracking approximately 2.1 million weblogs.

The best-known sites that provide blog popularity rankings are Technorati (www.technorati.com) and The Truth Laid Bear (TTLB) (www.truthlaidbear.com). One of Technorati's principal services is to allow users to search for "conversations," or linked discussions, on specific topics. Technorati also ranks the top 100 blogs according to their number of inbound links. However, because Technorati monitors sites that make use of RSS feed, it includes non-blogs (e.g., news sites) as well as blogs. Similarly, TTLB's Blogosphere Ecosystem is an application that scans weblogs once daily and generates a list ranked by the number of inbound links they receive from other weblogs on the list, displaying the top 100 for visitors. The Ecosystem's list of blogs was originally collected by copying the blogrolls of two popular A-list blog authors, Instapundit and Vodkapundit. TTLB also provides a breakdown of lists according to average daily traffic, a measure that relies on data collected by Sitemeter logs. Although both of these blog popularity measures are biased—Technorati by including any site with an RSS feed, and TTLB by taking as its point of departure the blogrolls of A-list bloggers—their lists are often referenced by bloggers, and taken as evidence that a blog is important (or not).

The third set of tools aims to track trends in the blogosphere and includes Blogdex (www.blogdex.net), Blog Pulse (www.blogpulse.com), and Daypop (www.daypop.com). In as much as these tools track the spread of ideas rather than (conversational) connections among people, they are not central to the present analysis, and are not discussed further here.

The above services are limited by both the number of blogs that the system is aware of and a tendency to emphasize blogs that employ specific technologies or products like RSS and Sitemeter. Some are further biased by having taken as their starting point blogs already known to be linked to other blogs. The result is that these services are not comprehensive portrayals of the blogosphere as a whole, but rather snapshots of specific, sometimes self-selected, slices of the blog universe.

3. Methodology

3.1. Research questions

This study investigates the degree and nature of interconnectedness among blogs, from the perspective of ordinary, non A-list blogs, regardless of the technologies they employ. The larger goal is to shed light on the "conversational" practices that take place in the blogosphere. The general research questions guiding this study are:

1) How interlinked is the blogosphere from the perspective of a random blog?
   1a) Which blogs are central?
   1b) Which blogs are more interconnected? Are there cliques?
   1c) Is the blogosphere a "small world"?
2) Do other types of "conversation" take place between linked blogs, and if so, to what extent?

3.2. Sampling procedure

In order to begin to address these questions, we devised a sampling technique intended to identify the social networks of randomly-selected blogs. We began by selecting as our initial points of access four blogs, using the random blog selection feature of the blo.gs service described in section 2.2. Because this was the same selection mechanism we used in our previous blog research, the selected blogs can generally be characterized by the descriptions provided in Herring, Scheidt, et al. (2004), with two exceptions. First, we selected only blogs that contained links to other blogs in their sidebars (typically, but not always, in a 'blogroll,' or list of links to other blogs), since only blogs with links could be analyzed for patterns of interlinking.[4] Second, we included blogs in

[4] This method excluded sites that interconnect by means other than through lists of links in their sidebars, such as many LiveJournal and Xanga blogs. This was done to increase the interpretability of the results, since in addition to making use of different interconnection mechanisms, popular blog-hosting sites such as LiveJournal and Xanga attract a different user

languages other than English, provided that we could understand them sufficiently to determine if they linked to other blogs, although some in writing systems that we could not read (such as Arabic and Japanese) were discarded. In order to obtain four blogs that met these criteria, 16 blogs had to be discarded.[5] None of the 20 blogs generated was a member of the A-list. The four blogs selected were: pencilinyourhand.blogspot.com (blog *a*), www.danm.us (blog *b*), www.mysocalledblog.com (blog *c*), and orangetang.org/erica/blogger.html (blog *d*). Of these, *a* and *b* are filter blogs with mostly political content, *c* is a combination of filter and personal journal, and *d* is a personal journal (cf. Herring et al., 2004).

Taking these four blogs as our jumping-off points, we manually collected the URLs of the links to other blogs in the sidebars out three "hops" or degrees of separation (four degrees, in the case of blog *d*, in that three degrees led to the identification of relatively few links). That is, we collected the four original URLs, the URLs of all the blogs they linked to, and the URLs of all the blogs linked to by the latter. URLs were followed to ensure they linked to blogs, and those that linked to non-blog sites were discarded. All qualifying unique URLs were included, even if they led to more than one blog maintained by the same author. This process resulted in the collection of a sample containing 14,890 unique source-destination pairs and 5,517 unique URLs, all of which were verified by a human coder as blogs.[6]

We chose a snowball technique of sampling from the four source blogs based on the reasoning that the average blogger does not connect to blogs randomly, but rather through established patterns of links. Kumar, et al. (2003) found that a randomized blogspace displayed little community structure. Our interest is in identifying conversations among blogs, which we hypothesized would be most likely to take place in topical communities, as Kumar et al. observed. At the same time, a snowball sample is limited in that descriptions of it apply strictly only to the sample identified, rather than to all of blogspace. Nonetheless, a close analysis of a sub-part of a large network can be suggestive of some properties of the whole.

3.3. Generating an 'A-List'

In order to interpret the results of our analyses in relation to previous claims about the interconnectedness of the blogosphere, we identified as A-list or non A-list those blogs that appeared in our sample. Following 'top 100' lists of blogs such as those provided by Technorati and TTLB, we operationalized A-List blogs as the blogs with the highest number of other blogs that link to them. Top 100 lists, however, calculate the number of inbound links by disparate means (see 2.2.), and thus their contents differ. To get as accurate a list of A-list blogs as possible, we generated a composite ranking derived from the NITLE Blog Census, the Technorati Top 100, and the TTLB Blogosphere Ecosystem. Blogs that appeared in any two of the three ranking systems and were ranked in the top 100 of their system were included in the composite A-list measure.

The original ranking from each system was noted for each blog. For blogs appearing in two of the ranking systems but not the third, a third, arbitrary ranking of 114 was entered, and the three rankings were averaged. This was done to weight in favor of those blogs that appeared in all three ranking systems. This method resulted in a composite list, independent of the list of blogs generated as described above, of 45 ranked unique URLs.

3.4. Analytical methods

3.4.1. Quantitative analysis. In order to elicit empirical detail about the associations between and among blogs in our sample, we applied quantitative and statistical network methods drawn from social network analysis (SNA) and graph theory. Within graph theory and SNA more specifically, analysis methods relate nodes (in this case, individual blogs) via either their interrelationships or via shared properties. Specifically, the quantitative questions we asked involved calculation of the overlap between sets of weblogs, the distances between nodes, the reciprocity of linkage between weblogs, and the relationship between the number of inbound links (in-degree) and outbound links (out-degree) of each coded weblog.

Overlap, Distance: In the data collection process, it became apparent that more than one of our starting points sometimes led to the same weblogs. To determine the degree of convergence, we computed the overlap between the sets of nodes that are reachable from each starting point. Similarly, computing the distance between any two points in the network was possible because all start points eventually led to overlapping data points. Determining the overlap between start points and the distance (path length) allows us to make inferences about the properties of the network as a whole, including which blogs are central in it, rather than just those of the neighborhoods surrounding each source blog.

Reciprocity: To determine tie strength (cf. Granovetter, 1973) or strength of bond between weblog authors, we measured the frequency with which blogs link reciprocally to each other versus the number of links that flow unidirectionally from one blog to another. Reciprocal

demographic (younger, more females), and appear to display different social network dynamics.

[5] The source blogs were selected in late February 2004. An additional 200 blogs that we randomly generated and inspected during the second week of May 2004 included 28.5% with links in their sidebars to other blogs.

[6] We employed manual blog identification because all automated methods that we experimented with made too many incorrect identifications and/or failed to identify significant numbers of blogs. As a result, our sample is small, but reliable.

links were posited to indicate stronger ties than unidirectional links.

In-degree and out-degree: With our limited data set, it is not possible to measure the frequency with which a particular node is linked to within the blogosphere as a whole—such a measure would require that we have a copy of the entire blogosphere for analysis. However, we measured the in-degree (number of inbound links) within our sample as a rough substitute for global popularity. We also measured out-degree (number of outbound links to blogs) to determine whether a particular weblog is an active or a passive participant in the weblog ecology, and analyzed the relationship between the two measures. Weblogs which are mostly sources or mostly destinations were posited to occupy different roles in the network (Gibson, et al., 1998; Kleinberg, 1999).

Statistical tools were employed to test the quantitative measures, including averages and standard deviations, chi squares, and log-linear regression (see section 4.1.).

3.4.2. Visualization. In order to allow the patterns of relationships among the blogs in our sample to emerge more clearly, visualizations were generated from the combined data set, taking as input the 14,890 weblog pairs stored in the format A→B in plain text. A Perl program was written to process the data in order to extract the following information and assign the visualization attributes:

Phenomenon	Visualized as
In-degree	color of node
Out-degree	color of node boundary
One-way and reciprocal links	color of arc
Main clusters	setting a threshold at in-degree
Trajectories accessible from each source weblog	layers/classes
Location of A-list blogs	layers/classes

To create the visualizations, the Pajek visualization tool was used (http://vlado.fmf.uni-lj.si/pub/networks/pajek/), together with the Kamada-Kawai energy algorithm for the lay out. The results were exported in SVG format in order to enhance the legibility of the visualization. Visualization was used both as an initial means of identifying patterns and refining the study's research questions, and to provide independent results that could be triangulated with those of the quantitative analysis.

3.4.3. Qualitative analysis of blog dyads. For the last phase of the analysis, reciprocally-linked pairs of blogs were chosen from clusters identified by the visual analysis, and analyzed qualitatively for evidence of "conversational" interaction. The pairs were considered to be interacting conversationally if one or both blog author referred to the other author or the author's blog in an entry, either by means of a textual reference or by linking to the other's blog, or posted a comment on an entry in the other's blog. We hypothesized that if reciprocal linking constitutes a meaningful conversation, reciprocally-linked blogs will also interact in a more dynamic fashion, by referring to one another textually and exchanging comments.

Four blog pairs were selected from each of three clusters according to the criteria of reciprocal linking, topic similarity, and proximity to one another in the visualization. A total of 12 pairs, or 24 individual blogs, were examined. Each member of each pair was coded for blog type (filter, journal, k-log, mixed, or other; Herring, et al., 2004), gender of the blog author, and mentions of the other in entries and in comments. Mentions in entries were analyzed by reading all of the entries on the main blog page and by conducting a Google site search through the entire blog and its archives for mentions of the other's personal name or blog name. Comments were analyzed by reading the comments on the most recent 20 entries in each blog, again looking for mentions of the other's personal name or blog name. Distributions of mentions of each type were noted, and interpreted in relation to blog type and blog author gender, including whether the authors constituted a same-sex or mixed-sex pairing.

4. Results

4.1. Quantitative

4.1.1. Overlap. To answer the question of the extent to which different starting nodes converge on the same endpoints, we considered all the paths having three hops for blog *a*, blog *b*, and blog *c*, and four hops for blog *d*. We then classified all the nodes reached into sets that were reached by *a*, *b*, *c* or *d* individually, or in any combination. The members of these sets were classified as A-list or not and counted; these results are displayed in Table 1.

A chi-square test for independence was conducted on the reduced table of eight cells, where nodes are classified as only reachable by one, two, or three starting nodes, or by all four starting nodes. The test is significant (chi-sq=194.2, 3 df, $p<.0001$), suggesting a strong association between a blog's membership in the A-list and how reachable it is from the four starting blogs. The largest deviance (177.2) came from A-list nodes reachable from all four starting points, suggesting that A-list blogs tend to be reachable from any starting point, whereas the same is not true for non A-list blogs. This result indicates that A-list blogs are more central in the network than other blogs.

4.1.2. Distance. To measure the diameter of the sampled blogosphere, we calculated averages and standard deviations for all possible non-circular paths between any of the two starting nodes. These averages and the numbers of paths are given in Table 2. Note that starting blog *d* is unreachable from points *a*, *b*, and *c*, and it generally takes more hops to get from *d* to the other blogs, suggesting that *d* is somewhat peripheral to the network. Conversely, *a* has the shortest average paths to reach itself or any

other of the starting nodes, suggesting that it is more central to the network. The relationships between the four starting nodes are displayed schematically in Figure 1.

Table 1. Overlap of browsing paths

	A-list	A-list	Non	Non
a	3		2299	
b	2	6	303	3109
c	1		506	
d	0		1	
ab	0		462	
ac	1		63	
ad	0	4	6	1228
bc	3		628	
bd	0		65	
cd	0		4	
abc	3		442	
abd	2	5	3	465
acd	0		1	
bcd	0		19	
abcd	22	22	284	284
Total	37	37	5086	5086

Table 2. Average path lengths (and standard deviations) for all extant paths between starting blogs

	a	b	c	d
a	4.889 (1.556)	5.631 (0.888)	6.012 (0.894)	unreachable
b	6.637 (1.269)	7.378 (1.701)	7.759 (1.637)	unreachable
c	6.732 (1.159)	7.473 (1.693)	7.854 (1.810)	unreachable
d	8.888 (1.326)	9.604 (1.512)	9.994 (1.517)	unreachable

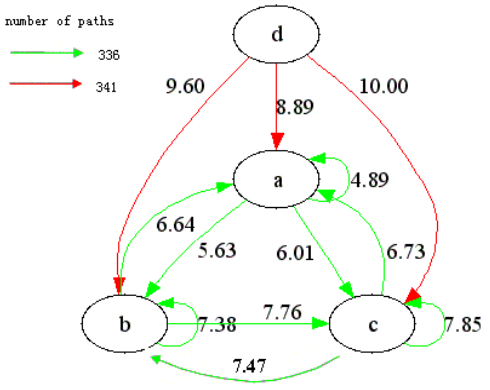

Figure 1. Path distances between the four starting nodes

4.1.3. Reciprocity. Reciprocity was measured by first generating the full complement of all undirected pairs of nodes in the sample, and then identifying which of these had a reciprocal link (labeled "R"), which had a directed or one-way link (labeled "D"), and which had no link at all (labeled Ø). In addition, the node pairs were classified as being between two A-list members (A-A), between an A-list member and a non A-list member (A-n) and between two non A-list members (n-n). The results are tabulated in Table 3. The table was modeled using a log-linear generalized linear model in GLMStat. For the nine cells in the table, a saturated model with nine parameters is required to explain the distribution (Table 4).

Table 3. Reciprocity of nodes in the sample

	R	D	Ø	Total
A-A	9	189	505	703
A-n	52	19535	8615	28202
n-n	1934	998863	14906184	15006981
Total	1995	118487	14915204	15035886

Table 4. Log-linear model for reciprocity

	est.	*se(est)*	*z ratio*	*Prob>|z|*
Constant	16.52	0.00026	63770	<0.0001
AA	-10.29	0.04450	-231.3	<0.0001
An	-7.456	0.01078	-691.8	<0.0001
R	-8.950	0.02274	-393.6	<0.0001
D	-5.016	0.00319	-1572	<0.0001
AA.R	4.923	0.33710	14.60	<0.0001
AA.D	4.033	0.08533	47.26	<0.0001
An.R	3.840	0.14090	27.25	<0.0001
An.D	5.834	0.01332	438.0	<0.0001

The model shows that there are far more Ø pairs than D pairs, and D pairs than R pairs, as well as more n-n pairs than A-n pairs and A-n pairs than A-A pairs, as expected. Nonetheless, there are significantly more A-A and A-n R and D pairs than expected, even given these general trends. That is, reciprocity is greater among pairs involving A nodes than among n nodes in general.

Since it matters for interpretation in which direction links among the A-n set go, we counted these separately according to whether the link starts from the A-list member or the non A-list member. The results show that non A-list blogs link preferentially to A-list blogs, but do not experience high rates of reciprocation. A-list blogs tend to be found in reciprocal relations with other A-list blogs.

4.1.4. In-degree and out-degree. To measure in- and out-linkage of the blogs sampled, we first excluded blogs that were at the ends of the paths in the sample, and that had not occurred elsewhere in the sample, as their out-degree is unknown. We then plotted the log of the in-degree against the log of the out-degree, as shown in Figure 2 (1 was added to both before taking the log to avoid logarithms of zero values). Red A's indicate A-list blogs,

and blue dots indicate non A-list blogs. The red and blue dotted lines are regression lines for A-list and non A-list blogs, respectively.

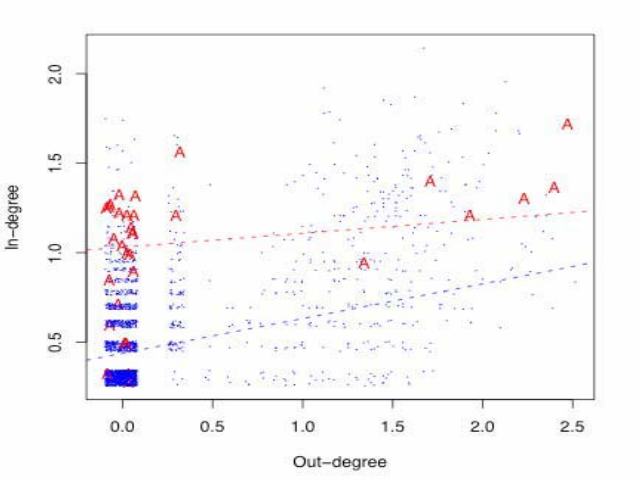

Figure 2. Relationship of in-degree to out-degree

Log-linear regressions of in-degree against out-degree show out-degree to be a significant predictor of in-degree but less strongly so for A-list blogs than non A-list. The slope of the regression line is steeper for the non A-list blogs (0.446 log in-degrees per log out-degree) explaining 35.8% of the variance, while that for the A-list is flatter (0.179 log in-degrees per log out-degree) explaining 27.3% of the variance. Hence, linking to other blogs on one's blog is likely to earn more links pointing back, but moreso among non A-list blogs that have a lower in-degree to begin with than among the A-list blogs.

Note that the majority of blogs in the sample have an out-degree of zero, for A-list as well as non A-list, but some nonetheless receive many inbound links. Such blogs can be considered 'authorities,' and those with a high out-degree can be considered 'hubs' (Gibson, et al., 1998). A-list blogs represent both types.

To summarize, quantitative analysis addressing social network concepts shows that A-list blogs are more central than non A-list blogs in the network generated by our data: most of the A-list blogs (37/45) can be accessed from all four source blogs, and other blogs link preferentially to them. They also have a higher out-degree than non A-list blogs. At the same time, the network that emerges from this analysis is quite closely interconnected overall, lending some support to the claim that the blogosphere is densely interconnected. All of the randomly-generated source blogs can reach one another. However, *d* is not itself reachable, and there is a difference in the closeness of connection of the four, with *a* being at a shorter distance, and *d* at a farther difference, than the others. These findings are both supported and qualified by the visualization results.

4.2. Visualization

The number of blogs in our sample is too large to display legibly in a social network diagram using Pajek. Figure 3 displays the results of visualizing the link-based connections in the sample with a cut-off of nodes at 10 in-degrees. This dramatically reduces the size of the network from 5,517 to 254 unique nodes, allowing clear patterns to emerge.

In the image in Figure 3, dark red nodes have the highest number of inbound links, while blue boundaries of nodes indicate the highest levels of outbound links. Green arcs (connecting lines between nodes) are reciprocal links; light grey arcs are one-way links. A-list nodes are indicated by triangles, non A-list nodes by ovals. A number of observations can be made on the basis of this visualization.

1) Three main clusters emerge, which we label Catholic weblogs (Figure 3, upper right) and Homeschooling weblogs (lower left), based on their topical content, and A-list blogs (lower right), which tend to focus on political commentary, although the A-list in this visualization also includes blogs by humorists Dave Barry and Tom Tomorrow.
2) The Catholic blogs and homeschooling blogs are densely reciprocally linked (via green lines).
3) The A-list blogs have many one-way in-bound links (grey lines) but few out-bound links. They tend to link to each other if there are reciprocal links.

These results paint a somewhat different picture from the quantitative analysis, which indicated that A-list blogs were more densely interlinked than non A-list blogs. The quantitative analysis did not recognize Catholic blogs or homeschooling blogs as separate subgroups; we identified these groups qualitatively only after they emerged from the visualization, by manually examining blogs in each cluster. The visualization results suggest that the most densely interlinked clusters of blogs are non A-list.

This interpretation is further supported by Figure 4, which represents successive increases in the cut-off point for in-bound links. As the cut-off point increases, blogs in the A-list cluster drop out. By the end of the winnowing process, only Catholic blogs remain (right-most image).

Figure 5 shows the number of links contributed by each source blog to the network. This visualization clearly shows the central status of *a* and the peripheral status of *d*. These findings suggest that for the method of random blog selection that we employed, selecting a well-connected start blog is crucial. It is notable that *a* and *b* are filter blogs, and they are also the best connected. Blog *c* is a mixed filter and personal journal, and *d* is a personal journal. From this limited evidence, it appears that filter blogs contain more links, and link to other blogs that contain more links, than do personal journal blogs.

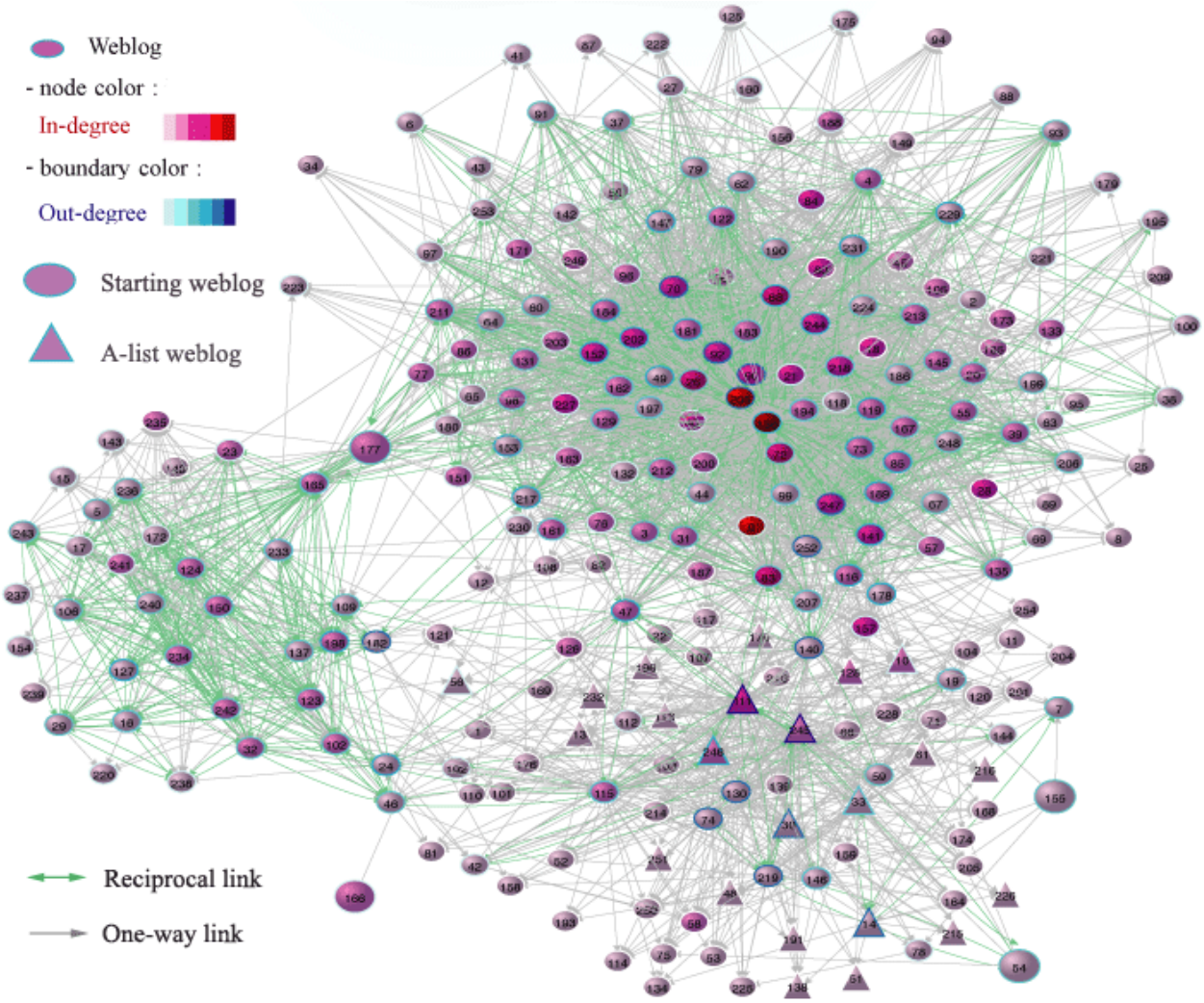

Figure 3. Blogosphere sample, cut-off at 10 in-degrees (total 254 unique nodes)

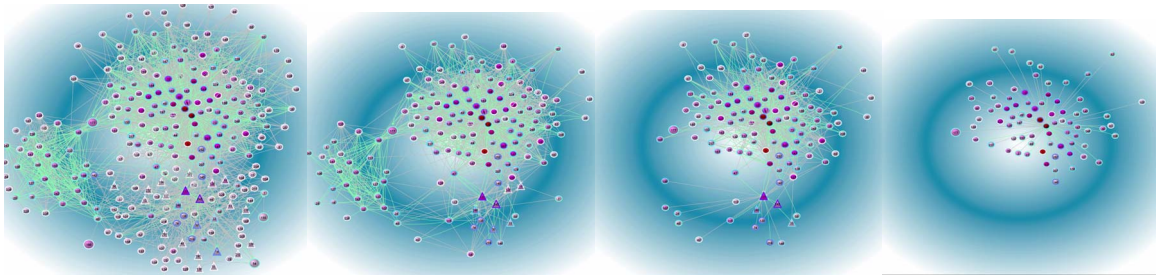

Figure 4. Effect of increasing in-degree cut-off (cut-off increases from left to right)

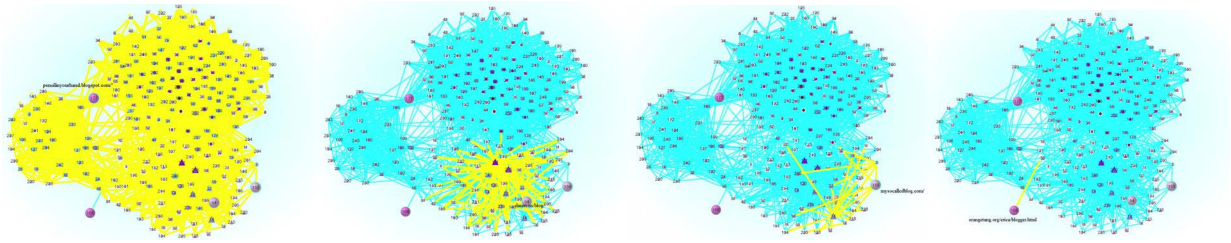

Figure 5. Trajectories (in yellow) accessible from each source weblog, from a (far left) to d (far right)

4.3. Qualitative analysis of blog dyads

The qualitative analysis of 24 blog pairs revealed that blogs that link to one another also tend to "converse" with one another more actively in their entries and comments, at least when the blogs are on closely-related topics. Eight out of the 12 reciprocally-linked dyads that we examined included mentions of the other blog author, and in all but one pair, mentioning was reciprocal. The largest number of mentions was comments posted to the other's blog (N=241), although only eight blogs made use of this option (another four blogs did not allow comments). Mentions of the other blogger's personal name, blog name, or a link to the other blog were less numerous (N=106), but were found in more of the blogs (N=13).

The latter type is common in two popular A-list blogs, Instapundit (Glenn Reynolds) and Vodkapundit (Stephen Green). For example, an Instapundit entry dated April 25, 2004 includes the mention:

> STEPHEN GREEN is correcting Tom Friedman.

In this example, the name STEPHEN GREEN is hyperlinked to an entry in the Vodkapundit blog. The form of address can also be more direct, as when Vodkapundit posts on May 06, 2004:

> Get Well Soon, Blogfaddah.

'Blogfaddah' is an affectionate reference to Glenn Reynolds, who has reported being ill in one of his entries. These mentions suggest a friendly relationship between the two A-list bloggers, although we are not aware of an off-line friendship between them.

Commenting tends to be more informal in tone, perhaps because it is less visible, and hence somewhat "off the record" (comments are not visible on most blogs unless the reader clicks on a link, in contrast to entries, which are displayed on the home page). The following comment was posted by Scott Chaffin of The Fat Guy (http://www.the fatguy.com) in response to a February 6, 2003 entry on dustbury.com (http://www.dustbury.com) discussing a neighbor who was spinning his vehicle's tires in his driveway on a thin layer of ice:

> I bet that old boy just wanted to spin his tires without fear of retribution from the coppers. I'm still enough of a teenager gearhead that I do that kind of dumb thing...maybe he is, too. Who doesn't like peeling out???

However, it is rare in this sample that the first blog author then responds to the second one's comment; that is, there is little evidence of extended exchanges.

There appears to be socially-based variation in the frequency and manner in which members of reciprocally-linked blogs converse. Blog topic conditions frequency of interaction: political blogs contain more conversation than Catholic blogs, which in turn contain more conversation than homeschooling blogs. However, it is also the case that the political bloggers are mostly male, and the homeschooling bloggers are mostly female (the Catholic bloggers include both males and females); thus gender, rather than topic, may condition this variation. Males post more mentions in entries and more comments than do females overall; this is especially evident in two mixed-sex dyads, in which a reciprocal conversation is taking place, but the male blogger makes more mentions of the female blogger than vice versa. For example, in his blog *Homeschool and Other Education Stuff* (http://www.cobranchi.com), Daryl Cobranchi posts entries referencing Isabel (Izzy) Azuola-Lyman or her blog *The Homeschooling Revolution* (http://icky.blogspot.com) more than 100 times, while she mentions him or his blog only 10 times. The tone of the messages on both sides is good-natured banter, as when Daryl blogs:

> Izzy and Chris [http://www.odonnellweb.com/mtarchives/001018.html] are having a mild disagreement over this [http://www.lewrockwell.com/latulippe/latulippe15.html] article. For the record, I'm with Chris on this one. Sorry, Izzy.

While our evidence suggests that mentioning another blog author tends to lead the other to reciprocate with a mention (indicating, crucially, that these authors are reading each other's blogs on a regular basis), they do not necessary reciprocate in the same manner. One member of the dyad sometimes mentions the other only in entries, while the other interacts primarily through comments posted to the other's blog. Such asymmetries may indicate a power dynamic (who goes to whose blog?); exploration of this possibility is left for future research.

5. Discussion

We have presented three types of evidence pertaining to the interconnectedness of a sample blogosphere generated from four randomly-selected blogs. The social network analysis results point to the centrality and influence of A-list blogs in the network, although they tend to be linked to other blogs by weak (one-way) ties. The visualization identified other, more strongly linked clusters, notably those comprised of blogs about Catholicism and homeschooling, and showed part of the sample blogosphere to be densely (albeit not completely) interconnected. A closer examination of selected pairs of reciprocally-linked blogs in each of the major clusters revealed that two-thirds were engaging in more dynamic textual interaction (e.g., posting comments), all but one pair reciprocally, although some dyads had no interaction beyond linking to each other's blogs in their sidebars.

To return to the research questions posed in section 3.1, our sample displays many characteristics of an interconnected network. There are central nodes (although we did not look at individual nodes closely), cliques, and a

tendency for preferential attachment (blogs link preferentially to A-list blogs), which defines a "small world" (Albert & Barabási, 2002). Moreover, we observed textual interaction between some blog pairs, involving reciprocal, verbal exchange. It is notable that in no case did a blog reference another blog only once—all the pairs that engaged in reciprocal referencing did so on multiple occasions, suggesting the existence of a relationship between them, not just a one-time exchange. These findings support the existence of interconnection and conversation in the blogosphere.

At the same time, the broader picture that emerges from our study indicates that these manifestations of interconnectedness are not uniformly characteristic of the blogosphere as a whole, but rather are restricted in occurrence. Throughout, our methods of analysis—by necessity, in as much as the social network approach can only apply when connections are present, but also because of our interest in seeking conversations—have focused on well-connected blogs, excluding less-connected blogs at successive stages of analysis. Yet only about one-quarter of all initially randomly-generated blogs were found to have any outbound links to other blogs. Including inbound links raises the percentage of random blogs connected in some way to other blogs to 58%,[7] but that still leaves 42% of the blogs tracked by blo.gs that appear to be social isolates, neither linking to nor being linked to by others. Further, those blogs that received fewer than 10 inbound links—95% of all the blogs in our already restricted sample—dropped out of the visualization, so that their patterns of connection (or lack thereof) are unknown.[8] Finally, the reciprocally-linked blog dyads analyzed for conversational interaction were selected to be as similar—and hence as likely to "converse"—as possible; this is a highly restricted sample. Even so, one-third of those otherwise closely connected blogs had no trace of textual conversation in the present sample, and it is unlikely that less interlinked blogs would refer and comment to each other more often (although this is an empirical question that could be investigated).

All of the available evidence thus suggests that less-connected (or unconnected) blogs represent the majority of blogs available on the Web today. That is, the blogosphere appears to be selectively interconnected, with dense clusters in parts, and blogs minimally connected in local neighborhoods, or free-floating individually, constituting the majority. Moreover, it seems likely that the much-touted textual conversation that all of the blogosphere is supposed to be engaged in involves a minority of blogs as well, and sporadic activity even among those blogs (for example, no dyads interacted publicly on their blogs every day, similar to the finding of Kumar, et al., 2003 that interactions among blogs are "bursty"). That participants, scholars and media commentators perceive conversation to be more widespread than it is may reflect the fact that relatively speaking, conversationality among blogs is considerably greater than that among traditional home pages or other HTML-based Web documents, linking aside (Herring, Scheidt, et al., 2004). Thus, while they may not be prototypically conversational or frequent in absolute terms, blog conversations appear to be a perceptually salient phenomenon.

[7] Based on 203 blogs randomly sampled from blo.gs and examined for both outbound and inbound links in mid-August, 2004.

[8] The quantitative analysis in Table 3 shows that 99% of all possible (as opposed to actual) pairings of blogs in the full sample were not linked to one another in any way.

6. Future directions

The analysis reported in this study was carried out on a sample of 5,517 weblogs that were manually identified by following hyperlinks from four random blogs. It was not our original intention to collect the blogs manually, as identification is tedious, but we encountered difficulties with the several blog identification algorithms we tried to use, all of which had unacceptably high rates of error. One of our future goals is to refine one of these algorithms to reduce the rate of error sufficiently so that much larger samples can be generated from random blogs, at further degrees of separation. The present analysis is proof of the concept, and succeeded in identifying patterns that would not have emerged as readily by other means, e.g., the Catholic and homeschooling blog cliques; it further revealed their network configurations and their relationships to A-list blogs. We hope to produce a more complete mapping of such topical communities when we are able to generate larger sample blogospheres from the bottom up, and/or to generate multiple local blogospheres from individual random source blogs. Simultaneously, we are directing efforts toward devising methods of sampling, analysis, and visualization that will render apparent patterns involving less-connected blogs.

7. Acknowledgments

The authors wish to thank Katy Börner for suggesting use of the Pajek visualization tool, and Sarah Mercure for assistance with the quantitative analysis.

8. References

Adamic, L. A. (1999). The small world Web. In *ECDL'99, Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries* (pp. 443-452). Berlin: Springer.

Adar, E., Zhang, L., Adamic, L. A., & Lukose, R. M. (2004). Implicit structure and the dynamics of Blogspace. Workshop on the Weblogging Ecosystem, *13th International World Wide Web Conference*. http://www.hpl.hp.com/research/idl/papers/blogs/blogspace-draft.pdf

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47-97.

Berkowitz, S. D. (1982). *An introduction to structural analysis: The network approach to social research*. Toronto: Butterworth.

Blood, R. (2002). *The weblog handbook: Practical advice on creating and maintaining your blog*. Cambridge MA: Perseus Publishing.

Coste, R. (2000). Fighting speech with speech: David Duke, the Anti-Defamation League, online bookstores, and hate filters. *Proceedings of the 33rd Hawaii International Conference on System Sciences* (HICSS-33). Los Alamitos: IEEE Press.

Degenne, A. & Forsé, M. (1999). *Introducing social networks*. London: Sage.

Delwiche, A. (2004). Agenda-setting, opinion leadership, and the world of web logs. Paper presented at the International Communication Association Conference, May 27-June 1, New Orleans, LA.

Faust, K. (1997). Centrality in affiliation networks. *Social Networks, 19,* 157-191.

Garton, L., Haythornthwaite, C., & Wellman, B. (1999). Studying on-line social networks. In S. G. Jones (Ed.), *Doing Internet research* (pp. 75-106). Thousand Oaks, CA: Sage Publications.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*.

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

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). Information diffusion through blogspace. *WWW2004*, May 17-22, New York.

Herring, S. C., Scheidt, L. A., Bonus, S., & Wright, E. (2004). Bridging the gap: A genre analysis of weblogs. *Proceedings of the 37th Hawaii International Conference on System Sciences* (HICSS-37). Los Alamitos: IEEE Press.

Herring, S. C., Kouper, I., Scheidt, L. A., & Wright, E. (2004). Women and children last: The discursive construction of weblogs. In L. Gurak et al. (Eds.), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*. http://blog.lib.umn.edu/blogosphere/women_and_children.html

Jackson, M. H. (1997). Assessing the structure of communication on the World Wide Web. *Journal of Computer-Mediated Communication, 3* (1). http://www.ascusc.org/jcmc/vol3/issue1/jackson.html

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM (JASM), 46*(5)*,* 604-632.

Knoke, D., & Kuklinski, J. H. (1982). *Network analysis*. Thousand Oaks CA: Sage Publications.

Kumar, R., Novak, P., Raghavan, S., & Tomkins, A. (1999). Trawling the web for cyber communities. *Computer networks*, *31*(11-16), 1481-1493.

Kumar, R., Novak, P., Raghavan, S., & Tomkins, A. (2003). On the bursty evolution of Blogspace. *Proceedings of the Twelfth International World Wide Web Conference*, Budapest, Hungary.

Marlow, C. (2004). Audience, structure and authority in the weblog community. Paper presented at the International Communication Association Conference, May 27-June 1, New Orleans, LA.

Merelo-Guervos, J-J., Prieto, B., Rateb, F., & Tricas, F. (in press). Mapping weblog communities. Submitted to *Computer Networks*.

Milgram, S. (1967). The small world problem. *Psychology Today*, *2,* 60-67.

Park, D. (2004). From many, a few: Intellectual authority and strategic positioning in the coverage of, and self-descriptions of, the "Big Four" weblogs. Paper presented at the International Communication Association Conference, May 27-June 1, New Orleans, LA.

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

Wellman, B. (2001). Computer networks as social networks. *Science Magazine, 293,* 2031-2034.

Wellman, B., & Wortley, S. (1990). Different strokes from different folks: Community ties and social support. *American Journal of Sociology*, *96*, 558-588.