

**Exploring the Network Structure of Virtual Communities:
A Web Structure Analysis of a Korean-American Community**

Sun-Ki Chai
Department of Sociology, University of Hawai`i
2424 Maile Way 247, Saunders Hall, Honolulu HI 96822 USA
Email: sunki@hawaii.edu

Mooweon Rhee
Shidler College of Business, University of Hawai`i
2404 Maile Way, Honolulu, HI 96822 USA
Email: mooweon@hawaii.edu

Author's note: Sun-Ki Chai (Ph.D., Stanford University) is Associate Professor in the Department of Sociology, at the University of Hawai`i. Mooweon Rhee (Ph.D. Stanford University) is Shidler College Distinguished Associate Professor of Management at the University of Hawai`i. Correspondence concerning this article should be addressed to Sun-Ki Chai at the address above. Telephone information includes: Office +1 808 956-7234, Fax +1 808 956-3707, Cell +1 808 741-4843.

**Exploring the Network Structure of Virtual Communities:
A Web Structure Analysis of a Korean-American Community**

Abstract

We introduce new general-purpose agent software, which crawls the web using user-specified “on-the-fly” social theory-based criteria to identify and analyze web sites in a particular virtual community. It then renders information about the sites into data in standard formats suitable for additional analysis by statistical or network software. In this study, a simple “two-layer” centrality measure is used to identify 250 highly interconnected websites forming a virtual community on Korean-American affairs. We then examine how theoretical propositions about centrality and prestige in the literature on social networks apply to cyberspace. Our analysis suggests that web sites with greater centrality within the virtual community are more prominent and popular, attracting a greater number of visitors.

Keywords: Virtual Community; Agent Software; Web Crawler; Centrality; PageRank;
Korean-American

**Exploring the Network Structure of Virtual Communities:
A Web Structure Analysis of a Korean-American Community**

Despite the recent explosion of interest in the internet among sociologists and the prominent role that the analysis of the internet plays in advancing diverse sub-disciplines in sociology, particularly social network analysis, there are fewer empirical, especially quantitative, sociological analyses of the internet than one might hope or expect (DiMaggio et al., 2001; Wellman and Haythornthwaite, 2002). Perhaps the greatest barrier to the analysis of the internet may be the limited access to suitable data from the internet. For instance, a search on the topics of “web” or “internet” and “social networks” in the EBSCO academic database allows us to locate only a handful of quantitative studies of virtual communities in sociology journals, each using their own ad hoc methodology of collecting data. Furthermore, most quantitative analysis of the internet tend to focus on email correspondence, chatrooms, message boards, or newsgroups as their research settings, using the content analysis methods developed for traditional text documents (Krippendorff, 2004). In contrast, quantitative analysis of the World Wide Web (WWW) is nearly nonexistent apparently because of considerable methodological challenges researchers face. This would seem paradoxical, since by its nature, content on the web is widely accessible and already in electronic form. Moreover, the “social” aspects of the web are obvious in the sense that web sites contain not only text content but also link to other web sites with related content (Stewart, 2003).

The most persuasive explanation for the shortfall in quantitative analysis of the WWW is the lack of suitable technologies, which allow one to crawl the web using

sociologically relevant criteria for sampling and to code the downloaded content into data suitable for statistical analysis (Weare and Lin, 2000). This paper presents a new technology for addressing those challenges, based upon agent software developed by one of the authors. The software is intended to download web sites in a virtual web community, while compiling quantitative information about the sites.

The basic methodology implemented by the software begins with an initial site or an initial set of sites specified by the user. The initial site(s) might be located through means of a web directory or link page, or based upon the user's own selection of representative sites on a particular issue of interest. As each site is downloaded, detailed information is compiled on its content and links to outside sites/ Data about its links from other sites, traffic patterns, and registration data os also compiled based on information from third-party web information databases.

The methodology then incrementally adds additional sites to the community, using a priority ranking algorithm built that takes into account strength of links in and links out to the existing set, as well as other criteria specified by the user. A link-based measure helps ensure that the incrementally growing set is highly cohesive as a network, and thus can appropriately be viewed as a virtual community (Moody and White, 2003; Wellman, Boase, and Chen, 2002). An advantage of the software is that it identifies sites that are closely related to sites in the original set, even if the user was not originally aware of their existence. Once a suitable set of sites is downloaded within a community, a variety of network measures can be obtained. For example, measures of site centrality within the community are calculated by application of link information, and are supplemented with data about the basic characteristics of each site, such as domain, age,

and popularity. It is important to note that there is no attempt in this paper to ensure that additional sites added are “about” the same subjects as the original site. While the software does include routines for collecting content data and for using content similarity as an additional criterion inclusion in a set, these are not implemented in the current investigation. This is because our intent in this study is not to investigate sites about a particular topic, but rather to examine the structural aspects of the sites that are tightly interconnected.¹

Nonetheless, the initial set of sites chosen for an investigation clearly will influence the final set, and it makes sense to choose an initial set that can be identified with a certain basis for common interest. The initial set chosen for this paper contains sites that all share a focus on the common substantive issue of Korean-American culture. This issue is chosen partly due to the importance of social networks presented by the literature in ethnicity and identity (Mehra, Kilduff, and Brass, 1998; Saxenian, 2000), but also because the paucity of quantitative data available for the study of the internet as a medium of social interaction is mirrored by a comparable lack of data for the study of ethnicity as a cultural basis for interaction (Chai, 2001). In particular, while some U.S. research groups provide case studies of ethnic identification and interactions using surveys and interviews (Portes, 1995; Saxenian, 1999), there are few systematic databases of broad breadth that provide information about ethnic interactions within an ethnic group or between ethnic groups.

¹ We have also developed agent software that aims to investigate the relationships among web sites in terms of their content. Access to the software may be available from the first author upon request.

Theoretical Discussion on Centrality and Popularity

The formal sociological literature that translates most straightforward into propositions about the WWW is the theoretical literature on social networks (cf. Scott, 2000; Wasserman and Faust, 1994). This is because there are numerous characteristics of web sites that are analogous to the characteristics of individuals and groups used in network. Most importantly, the connections between individuals and organizations, which serve social network theory as its core research theme, can be viewed as analogous to the hypertext links that exist between web sites (Dunne, Williams, and Martinez, 2002; Kleinberg et al., 1999; Park, 2003; Park, Kim, and Barnett, 2004). The main difference between hypertext links and conventional social network connections is that hyperlinks are always “directed,” from one page or site to another, producing asymmetric matrix of ties among pages/sites.

In this study, we focus on the centrality, one of the most prominent concepts in social network theory (Freeman, 1979), of web sites within a virtual community and its consequence on site popularity. The measurement and analysis of centrality, introduced by Bavelas (Bavelas, 1948; 1950), has certainly generated a huge amount of theoretical and empirical writing. Much work in the past few decades has been devoted to conceptual clarification of centrality, as well as to investigations of its linkages to other variables of the theoretical or empirical interests (e.g., Bonacich, 1987, 2007; Everett and Borgatti, 1999; Everett, Sinclair, and Dankelmann, 2004).

Work by Freeman has led to the distinction being made between point centrality, the characteristics of a single node or web site, and graph centrality, the characteristic of an entire community (Freeman, 1977; 1979). Freeman proposes three different types of

node centrality that are derived from the ‘degree,’ ‘betweenness,’ and ‘closeness’ of ties among nodes. Bonacich has proposed a measure of centrality which recursively takes into account the centrality of each of a node's partners (Bonacich, 1972; 1987). In addition to incorporating partner’s centrality, Bonacich’s centrality measure is also known to account for the intensity and direction of linkage, which are not adequately addressed in Freeman’s measures of centrality. This enhances the validity of Bonacich centrality for the network study of virtual communities where tie intensity and directions are usually identifiable (Bonacich 2007). Thus, one version of this measure is adopted in our study as described below.

It may be of interest to social network researchers to know that measures closely resembling Bonacich centrality are now widely being implemented by computer scientists to map the internet (cf. Chin and Chignell, 2006). Certainly the most prominent such effort has resulted in the generation of “PageRank” algorithm, which forms the basis of the Google search engine (Brin, 1997; Brin and Page, 1998; Page, 1998). The PageRank (PR) of a particular page $p \in S$ is defined as $PR(p) = (1 - d) + d (\sum_{p,q \in E} PR(q) / L(q))$, where S is the set of pages in a database, E is the set of pairs in $S \times S$ such that the first page in the pair links to the second, $L(q)$ is the number of links out from page q , and $0 < d \leq 1$ is a “damping factor” to take into account closeness (Brin and Page, 1998: 107-108). It is straightforward to see that PageRank is similar to Bonacich's measure of centrality, with intensity being defined as being the inverse of the number of degrees originating from each page linking to the page in question.

A somewhat different measure is employed to enhance search for web documents by IBM's Clever Project. The project’s HITS algorithm examines both forward and

backward link structure to differentiate “authorities” and “hubs” (Kleinberg, 1998; Chakrabarti, Dom et al., 1999). Each page p is associated with an authority weight $x(p)$ and a hub weight $y(p)$ such that $\sum_{q \in S} x(q)^2 = 1$ and $\sum_{q \in S} y(q)^2 = 1$. Furthermore, $x(p) = \sum_{q, p \in E} y(q)$ and $y(p) = \sum_{p, q \in E} x(q)$, with S and E defined similarly as for PageRank. Note that a page's authority weight thus depends on the number and hub weight of the pages that link to it, while a page's hub weight depends on the number and authority weight of pages it links to. Connection intensity, however, is not adjusted for the number of links in or out that a page's counterpart has (Kleinberg, 1998: 8-9). Besides being used as the basis for the design of a search engine, the HITS algorithm has been used to supplement content analysis as a basis for resource classification (Chakrabarti et al., 1998). It has also been shown how such analysis can be done in real-time, without the benefit of a large existing database of web sites (Chakrabarti, Van den Berg et al., 1999). Finally, it has been used recently to identify emerging virtual communities on the web (Gibson et al., 1998; Kumar et al., 1999).

In the social network literature, Bonacich centrality has typically been used as an explanatory variable, often to predict outcomes related to power, prestige, and other status characteristics (Bonacich, 1987; Faust and Wasserman, 1992; Mizruchi and Potts, 1998; Podolny, 2005). In particular, there is a persistent scholarly interest in how such characteristics help central actors enjoy a great popularity from the audience. In their study of technological changes in the semiconductor industry, Podolny and Stuart (1995) show that subsequent innovations in the industry are more likely build upon the ideas generated by central actors. Washington and Zajac's (2005) study of which college is more likely to be invited to the NCAA postseason basketball tournament suggest that

people will want to watch the games of teams with higher centrality scores in their prior game network. A reasonable link underlying the relationship between centrality and popularity would be that an actor's prestige or status stemming from its centrality enhances perceived quality and prominence, which in turn helps the actor attract positive attentions from the audience (cf. Podolny, 2005; Rhee and Haunschild, 2006; Rindova et al., 2005). While there is no direct counterpart to the human or organizational quality of status among web sites, the goal of those who design most sites is clear-cut. They will be successful if and only if many people view them. Therefore, we propose:

Hypothesis: Web sites that are central within a virtual community are more popular, attracting a greater number of visitors.

While Bonacich centrality itself is often employed as a direct measure of popularity (Bonacich, 1987, 2007; Scott, 2000), those two concepts are distinct in this study (similar to the above empirical studies) in that our measure of site centrality draws upon hyperlinks among web sites whereas site popularity is measured in terms of the number of recorded visits to a web site. (cf. Marlow, 2004).

Methodology

The agent software used in this paper was implemented in the scripting language Rebol.² It is a specialized web crawler containing routines that use social science constructs to generate algorithms for exploring and analyzing web links and content. It also contains routines for parsing html text, storing and indexing content and link

² Newer versions are being implemented in Java, with a much wider palette of admission and post-processing models available, but the original Rebol version was more than sufficient for this particular research.

information, and generating datasets in a form usable by statistical and social network software programs.

In contrast to the normal computer science approach, the unit of analysis for our data collection is an entire site rather than a single web page. For the purposes of this project, a site is defined as a host root url (e.g. with a url such as *http://www.mysite.com/*) and all linked pages that share the same host. While search engine algorithms and existing technical analysis of link structure on the web typically tend to take a web page as its primary unit of analysis, a collection of pages sharing the same host do not really correspond to the typical definition of virtual community, since they are likely to have been produced by the same individual or group. Thus, web site appears to be a valid unit of actor within a virtual community.

The centrality algorithm used in the study adopts elements of both the PageRank and Clever algorithms. Its formula for calculating centrality is based upon a simplified version of the PageRank algorithm. To lessen computational burden, in-link centrality was normalized by weighting each in-link for a site in proportion to the number of own in-links the source of the in-link has, but not did not weight the source's in-links. However, a "backward" centrality somewhat analogous to Clever's hub weights was also calculated by using links out rather than links in as the basis for comparison.

The notion of the set of sites in the database is complicated somewhat by the fact that there are two sets of sites, a database S containing all the sites that are either in the virtual community or candidates for inclusion, and the virtual community itself $R \subset S$. Furthermore, each database expands with each iteration of the program, with R absorbing an additional site or sites from S , and S expanding to include all sites that are

linked to or from any site in R . More formally, let R^{in} indicate the set of sites that link to at least one site in R , and likewise R^{out} indicate the set of sites that are linked to from at least one site in R . If we use t to indicate time period, $S_{t+1} = S_t \cup R_t^{out} \cup R_t^{in}$.

For this research, the virtual community R was expanded one site per period, by adding to it the site in S/R with the highest *omnidirectional local centrality*, i.e. the site $p \in S, p \notin R$ that had the highest product of forward and backward centrality based on links to and from sites in R . Hence omnidirectional local centrality for site p is $OC(p) = (\sum_{q \in R, p, q \in E} LC_{in}(q, S) / L_{in}(q)) * (\sum_{q, p \in E} LC_{out}(q) / L_{out}(q))$, where L_{in} and L_{out} measure the degrees links in and out to all websites, and LC_{in} and LC_{out} degrees links in and out to websites in R . After each iteration, the highest ranking site in S/R was added to R , and centralities of each site in S were recalculated to take into account the expanded set.

Links out data for each site were obtained by crawling down the site and analyzing its parsed html pages for hyperlinks to outside sites. Links in data for a site cannot be obtained by examining the site itself, but were obtained indirectly via the Altavista search engine, which featured search flags that allowed users to filter search results by url strings, i.e. including only sites that link to a particular specific webpage or any page which starts with the same url, and also placed no restrictions on the use of its search engine by agent software. Site clustering was specified in the search string to reduce multiple links from the same site, and remaining duplicates were eliminated.

Data on popularity were the perhaps most difficult to obtain, since server data across sites is not systematically collected or stored in any centralized fashion. Hit counters are notoriously unreliable, and are not available for most sites. Instead, data was obtained in an indirect fashion via Alexa Internet, a web database based on a free browser

“assistant” that was installed by default on later versions of Netscape and many copies of Internet Explorer, and is a popular add-on for other browsers. Browsers with Alexa installed and enabled automatically send data on a user’s browsing history back to a central server, where the information is used to compile information on related sites that can be queried from the browser. One side effect of this is that the data allows Alexa to compile information on how many times its users as a whole have visited a particular site. While this information is not built into a browser’s query features, it can be accessed publicly through appropriate url request to the Alexa web database server.

Certain implementation decisions were made to facilitate data collection and generate clear criteria for inclusion and exclusion:

Only links corresponding to a root URL were considered as sites. For instance, links to urls such as *http://www.geocities.com/mysite/* were excluded from the set *a priori*. The reason for this is that outside links into a site may point to multiple levels of a single directory tree, and it is very difficult to determine what level of the directory tree, if not the root level, corresponds to the highest level of a single site. While this does exclude a number of personal homepages on sites such as Geocities, it does mean that most sites, especially the larger ones, which tend to have their own hosts and servers, are retained. Furthermore, the tendency more recently has been for personal homepages on free hosting services such as tripod and xoom/nbci to have their own root URL, e.g., *mysite.tripod.com*.

Urls containing or lacking a third-level domain name or fourth-level domain name in the case where the first-level name specifies a country were automatically given a prefix of “www”, i.e. *mysite.com* → *www.mysite.com* and *mysite.co.kr* →

www.mysite.co.kr. This is because truncated and untruncated urls of this type virtually always point to the same IP address, and hence to the same site. Failing to add a prefix would create duplication, since the two types of urls may be used interchangeably by people linking to a site. This implementation decision reflects a methodology adopted by some search engines, as well as by browsers (such as in their auto-completion feature).

When crawling a site, the agent software downloaded only pages containing html content, since graphics, applications, and multimedia would rarely contain the link information crucial to the program, and were difficult to parse given the myriad of formats that non-html content may be stored in. Dynamically generated pages such as produced by CGI scripts (e.g., those containing “?” in URLs) were not followed because such scripts tend to accept multiple URLs for the same content, hence causing unwanted duplications.

Additionally, only URLs using the http protocol were followed, while those using the ftp or gopher protocol were ignored. The main justification in the ftp case is that ftp sites typically are used to archive large-size application or media files rather than html or other text content. Gopher sites are now quite rare, and their inclusion therefore seemed very unlikely to affect the results significantly.

Furthermore, sites were only crawled to a depth of two units. This was to avoid the agent software becoming bogged down downloading all the html content of huge sites such as *http://www.yahoo.com/* or *http://www.microsoft.com/*. A single unit was defined an html “href” link along the shortest network path from a page to its site’s root page. Html “src” links, such as one found in a frameset, were considered to be zero units, since

the linked pages would be rendered by a browser in the same document as the linking page.

The initial set of sites was obtained by downloading all links found in the Korean-American page of the Open Directory (<http://dmoz.org/>), a popular web portal. Only 15 sites whose urls met the criteria described above were included, and site urls were normalized as indicated above. Through the crawling process, sites were then incrementally added to the set until a target sample of 250 sites was obtained.

Analysis and Results

While the primary purpose of this paper has been to demonstrate a new way of generating quantitative data for sociological analysis of a virtual community, it also provided information about a fairly large set of interrelated sites that could be qualitatively analyzed. A qualitative examination of the final set revealed some anticipated and some unanticipated results. First of all, as expected, most of the sites that were incrementally downloaded related to Korean-American, Korean, or Asian-American culture and identity, as indicated by their page titles or actual contents. Such a result confirmed the prediction of homophily principle that sites that are closely tied to sites on Korean-American issues will themselves tend to be about similar or related topics. (Lazarsfeld and Merton, 1954; McPherson, Smith-Lovin, and Cook, 2001).

However, one thing that was perhaps surprising was the extent to which Non-Korean Asian-American sites were restricted to Japanese-American topics, including such sites as <http://www.nikkeiwest.com/>, <http://www.japantownsanjose.org/>, <http://www.taiko.org/>. In all, about 20 downloaded sites could be classified as Japanese-

American in content, and several more that were pan-Asian in character, while there was not a single downloaded site that could be classified as specific to another single Asian-American group. This might imply relatively strong network closeness between Korean-American and Japanese-American groups in cyberspace, probably reflecting the great intensity of political, historical, cultural proximity and exchanges between the two ethnic groups.

Besides sites that focused on Korea or Asian America, other sites in the downloaded set were generally large news sites or portals such as <http://www.cnn.com/>, <http://www.excite.com/>, and <http://www.webcrawler.com/>. This is not surprising given the large number of overall links in and out that such sites have. However, it is important to note that such sites on average do not have a particularly high centrality scores compared to other sites in the downloaded virtual community, based either on links in or out to other sites in the community.

Quantitative analysis was performed by identifying the following independent and control variables: The three main explanatory variables were normalized links-in centrality, i.e. $(\sum_{p,q \in E, q \in S'} 1 / C_{in}(q))$ for site $p \in E$, and normalized links-out centrality, i.e. $(\sum_{q,p \in E, q \in S'} 1 / C_{out}(q))$, as well as their product. Given the possibility that degree centrality is an alternative antecedent of site popularity (Freeman, 1979; Jansson, 2000; Zemljic and Hlebec, 2005), we also include overall degrees of links in and links out, as well as their product, as both substitute terms and additive terms. Models with the additive terms thus provide a conservative test of our hypothesis.

We first control for the age of the site in days (as obtained from Alexa data) as age sometimes signals an actor's quality and position (Hannan, 1998). We also include

three dummy variables to control for the nature of the site's URL: whether it had country-specific domain name, e.g., *mysite.com*, *mysite.edu*, or *mysite.org*; whether it was a corporate URL, e.g., *mysite.com* or *mysite.co.uk*; and whether it was an educational URL, e.g., *mysite.edu* or *mysite.ac.jp*.

Table 1 reports OLS (ordinary least square) regression results for models used in the study. Model 1 includes only three terms related to degree centrality (“Links In,” “Links Out,” and “Links In * Links Out”) and shows that their effects on site popularity are all insignificant. These insignificant effects are persistent across all subsequent models. Model 2 includes our three main independent variables (“Links-in Centrality,” “Links-out Centrality,” and “Links-in C * Links-out C”). As shown in Model 2, of the three variables, only normalized links-in centrality displays a significance effect ($p < .05$) on site popularity. This significance is maintained in Model 3 (which combines Model 1 and Model 2), Model 4 (which adds site age), and Model 5 (which adds three dummies regarding the nature of the site and excludes site age). However, normalized links-in centrality loses its significance in Model 6 which includes all independent and control variables. Among the control variables, only “dot-com” sites significantly and positively affect the extent of site popularity (Models 5 and 6).

Table 1 About Here

Conclusion and Discussion

A primary contribution of this study is its presentation of a novel software agent, which automatically crawls web sites to locate and download data about a virtual community that shares particular interests and topics. The agent then provides

information about the network structure of downloaded sites into quantitative data available for statistical analysis. We believe that scholars in a variety of sociological disciplines, particularly social networks and community sociology, would greatly benefit from the software agent by generating valuable databases because the data would be otherwise difficult to collect from actual communities, would have considerable sociological implications, and would develop distinctive understanding and operationalization of virtual communities (DiMaggio et al., 2001; Wellman, 2001; Wellman and Haythornthwaite, 2002).

While it is difficult to generalize our findings to all types of virtual communities, this study provides consistent results with prior studies of actual communities, demonstrating the positive effects of centrality on its popularity (Podolny, 2005; Podolny and Stuart, 1995; Rhee and Haunschild, 2006; Rindova et al., 2005; Washing and Zajac, 2005). Higher centrality makes an actor more prominent and visible, and thus helps the actor draw greater awareness and attention from an audience. More specifically, our study suggests that at least within the Korean-American and related virtual communities, two conditions hold.

First, the ascendancy of normalized links-in centrality over normalized links-out centrality suggests that a site's popularity depends more on how many sites link to it than on how many sites it is linked to. That is, in the context of ethnic virtual communities, it appears to be that prominence and attractiveness relate more to the extent of an actor being referred to or endorsed by others than to the extent of an actor referring to or endorsing others (cf. Merton, 1968). However, given that each centrality measure has different implications for site outcomes depending on the nature of flows between nodes

(Borgatti, 2005), a different type of virtual communities may observe that site popularity is affected by other centrality measures. For example, in a virtual community that aims to inform researchers of useful data sources, links-out centrality or betweenness centrality may be a crucial determinant of site popularity.

Second, the lack of significance for either total links in or total links out (i.e. degree centrality) may suggest that a site's popularity depends in particular on having connections to sites which share a common overall network and topic, and that connections to other sorts of sites may be less useful. Our brief qualitative investigation into the sites supports this speculation, confirming the need to consider network content as well as network structure in the prediction of network outcomes (Lin, 2001; Podolny and Baron, 1997).

Clearly, the analysis being presented here is quite preliminary, and further analysis on this topic should address several issues. First, while the sample size in this study was deliberately limited to 250 to facilitate qualitative examination of downloaded sites, there is little (save disk space) that precludes studies encompassing larger samples using the same software. A study of virtual communities with a long arm may test for how reliable our software agent is for large samples.

Second, it would be useful to examine how results would be changed by the employment of inclusion criteria based on a more fully recursive centrality algorithm (such as a fuller implementation of localized eigenvector centrality) or, alternatively, by a criterion based only on degree centrality to other sites within a community. The relevance of a particular inclusion criterion might be also affected by the nature of virtual

communities (Borgatti, 2005). Such an algorithm is present in more recent versions of the software.

Third, as mentioned above, the software already contains routines to compile text data from site contents, as well as to analyze it using the wide range of established content analysis semantic “dictionaries”, and such data could be used either as part of the crawling algorithm (as criteria for inclusion of new sites in a set) or for the generation of variables for statistical analysis. While we did not incorporate this function in the study due to the desire to use simple criteria rather than potential contentious complex inclusion criteria, future research should be able to achieve more abundant utilization of our software agent.

In the longer run, furthermore, the software can be used to investigate a wider range of virtual social phenomena, both in terms of empirical topic and in terms of theoretical literature. We hope that a discipline centered around the quantitative analysis of the web will have ample room for expansion and can ultimately start to take its place within the realm of sociological methodology.

References

- Bavelas, A., 1948. A mathematical model for group structure. *Applied Anthropology* 7, 16-30.
- Bavelas, A., 1950. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America* 57, 271-282.
- Bonacich, P., 1972. Technique for analyzing overlapping memberships. *Sociological Methodology* 4, 176-185.
- Bonacich, P., 1987. Power and centrality: a family of measures. *American Journal of Sociology* 92, 1170-1182.
- Bonacich, P., 1991. Simultaneous group and individual centralities. *Social Networks* 13, 155-168.
- Bonacich, P., 2007. Some unique properties of eigenvector centrality. *Social Networks* 29, 555-564.
- Borgatti, S.P., 2005. Centrality and network flow. *Social Networks* 27, 55-71.
- Brin, S., 1997. *Extracting Patterns and Relations from the World Wide Web*. Stanford University, Computer Science Department.
- Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems* 30, 107-117.
- Chai, S.-K., 2001. *Choosing an Identity: A General Model of Preference and Belief Formation*. University of Michigan Press, Ann Arbor, MI.
- Chakrabarti, S., Berg, M., Dom, B.E., 1999. Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. Paper read at 8th Annual WWW Conference.
- Chakrabarti, S., Dom, B.E., Kumar, S.R., Raghavan, P., Rajagopalan, S., Tomkins, A., Gibson, D., Kleinberg, J., 1999. Mining the Web's Link Structure. Paper read at IEEE Computer.
- Chakrabarti, S., Dom, B.E., Raghavan, P., Rajagopalan S., Gibson, D., Kleinberg, J., 1998. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. Paper read at 7th Annual WWW Conference.
- Chin, A., Chignell, M., 2006. A Social Hypertext Model for Finding Community in Blogs. Paper read at HT: Hypertext and Hypermedia, at Odense, Denmark.

- DiMaggio, P., Hargittai, E., Neuman, W.R., Robinson, J.P., 2001. Social implications of the internet. *Annual Review of Sociology* 27, 307-336.
- Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Food web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences* 99, 12917-12922.
- Everett, M.G., Borgatti, S.P., 1999. The centrality of groups and classes. *Journal of Mathematical Sociology* 23, 181-201.
- Everett, M.G., Sinclair, P., Dankelmann, P., 2004. Some centrality results new and old. *Journal of Mathematical Sociology* 28, 215-227.
- Faust, K., Wasserman, S., 1992. Centrality and prestige: a review and synthesis. *Journal of Quantitative Anthropology* 4, 23-78.
- Freeman, L.C., 1977. A set of measures of centrality based on betweenness. *Sociometry* 40, 35-41.
- Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. *Social Networks* 1, 215-39.
- Gibson, D., Kleinberg, J., Raghavan, P., 1998. Inferring web communities from link topology. Paper read at 9th ACM Conference on Hypertext and Hypermedia.
- Hannan, Michael T., 1998. Rethinking age dependence in organizational mortality: logical formalizations. *American Journal of Sociology* 104, 85-123.
- Jansson, I., 2000. Popularity structure in friendship networks. *Social Networks* 21, 311-410.
- Kleinberg, J., 1998. Authoritative sources in a hyperlinked environment. *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 668-677.
- Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A., 1999. The web as a graph: measurements, models, and methods. *Proceedings of the International Conference on Combinatorics and Computing*, 1-17.
- Krippendorff, K., 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Sage, Thousand Oaks, CA.
- Kumar, R., Raghavan, P., Rajagopalan S., Tomkins, A., 1999. Trawling the web for emerging cyber-communities. Paper presented at 8th Annual WWW Conference.
- Lazarsfeld, P.F., Merton, R.K., 1954. Friendship as a social Process: A substantive and methodological analysis. In: Berger, M. et al. (Eds.), *Freedom and Control in Modern Society*. Litton, New York, pp. 18-66.

- Lin, N., 2001. *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press, Cambridge.
- 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.
- McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: homophily in social networks. *Annual Review of Sociology* 27, 415–444.
- Mehra, A., Kilduff, M., Brass, D. J., 1998. At the margins: a distinctiveness approach to the social identity and social networks of underrepresented groups. *Academy of Management Journal* 41, 441–452.
- Merton, R.K., 1968. *Social Theory and Social Structure*. The Free Press, New York.
- Mizruchi, M. S., Blyden B. P., 1998. Centrality and power revisited: actor success in group decision-making. *Social Networks* 20, 353-387.
- Moody, J., White, D.R., 2003. Social cohesion and embeddedness: a hierarchical conception of social groups. *American Sociological Review* 8, 1–25.
- Page, L., 1998. *Pagerank: Bringing Order to the Web*.
- Park, H. W., 2003. Hyperlink network analysis: a new method for the study of social structure on the web. *Connections* 25, 49-61.
- Park, H. W., Kim, C. S., Barnett, G., 2004. Socio-communicational structure among political actors on the Web. *New Media and Society* 6, 403-423.
- Podolny, J. M., Baron, J. N., 1997. Resources and relationships: social networks and mobility in the workplace. *American Sociological Review* 62, 673–693.
- Podolny, J. M., Stuart, T. E., 1995. A role-based ecology of technological change. *American Journal of Sociology* 100, 1224-1260.
- Podolny, J., 2005. *Status Signals*. Princeton University Press, NJ.
- Portes, A., 1995. *The Economic Sociology and the Sociology of Immigration: Essays on Networks, Ethnicity, and Entrepreneurship*, Russell Sage Foundation, NY.
- Rhee M, Haunschild P R., 2006. The liability of good reputation: a study of product recalls in the U.S. automobile industry. *Organization Science* 17, 101-169

- Rindova V. P., Williamson I. O., Petkova, A. P., Server, J. M., 2005. Being good or being known: an empirical examination of the dimensions, antecedents, and consequences of organizational reputation. *Academy of Management Journal* 48, 1033–1049.
- Saxenian, A.L., 1999. *Silicon Valley's New Immigrant Entrepreneurs*. Public Policy Institute of California, San Francisco, CA.
- Saxenian, A.L., 2000. Networks of immigrant entrepreneurs. In: Lee, C.-M., Miller, W.F., Hancock, M.G., Rowen, H.S. (Eds.), *The Silicon Valley Edge*, Stanford University Press, Stanford, CA, pp. 248-268.
- Scott, J., 2000. *Social Network Analysis: A Handbook*. Sage Publications, London.
- Stewart, K. J. Trust transfer on the World Wide Web. *Organization Science* 14, 5-17.
- Washington, M., Zajac, E. J., 2005. Status evolution and competition: theory and evidence. *Academy of Management Journal* 48, 282-296.
- Wasserman, S., Faust, K., 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press, NY.
- Weare, C., Lin, W., 2000. Content analysis of the world wide web: opportunities and challenges. *Social Science Computer Review* 18, 272-92.
- Wellman, B. and Haythornthwaite, C., 2002. *The Internet in Everyday Life*. Blackwell, Oxford.
- Wellman, B., 2001. Physical place and cyber place: the rise of personalized networking. *International Journal of Urban and Regional Research* 25, 227-252.
- Wellman, B., Boase, J., Chen, W., 2002. The networked nature of community: online and offline. *IT & Society* 1, 151-165.
- Wellman, B., Haythornthwaite, C., 2002. *The Internet in Everyday Life*. Blackwell, Oxford.
- Zemljič, B. and Hlebec, V., 2005. Reliability of measures of centrality and prominence. *Social Networks* 27, 73–88.

Table 1. The Effects of web site centrality on site popularity

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	-33,203.00 (199,939.09)	42,288.00 (136,774.42)	-180,317.00 (229,474.38)	-332,999.00 (320,539.69)	-598,991.00* (260,732.70)	-672,882.00* (339,764.47)
Degrees In	3,836.01 (2,892.49)		3,454.84 (2,951.92)	2,719.59 (3,639.45)	5,910.67 (3,052.10)	6,525.74 (3,795.61)
Degrees Out	230.45 (768.80)		2,95.46 (768.12)	318.42 (817.10)	305.56 (774.41)	314.99 (828.51)
(Degrees In)× (Degrees Out)	11.04 (13.21)		10.53 (13.60)	9.87 (14.54)	8.28 (13.77)	7.87 (14.79)
Two-Step In-Centrality		70.64* (31.50)	67.83* (31.39)	74.22* (36.47)	61.07* (30.82)	66.50 (35.94)
Two-Step Out-Centrality		47.00 (37.76)	10.79 (43.20)	15.41 (47.50)	7.32 (42.40)	7.63 (46.74)
(TS In C)× (TS Out C)		-0.005 (0.004)	-0.003 (0.004)	-0.004 (0.004)	-0.003 (0.004)	-0.004 (0.004)
Site Age				133.08 (142.31)		4.49 (144.25)
Non-Country					210,974.00 (231,232.31)	203,489.00 (261,777.31)
Dot-Com					590,272.00* (222,808.73)	649,593.00* (258,046.53)
Dot-Edu					51,488.00 (583,544.00)	69,720.00 (621,740.81)
<i>N</i>	249	249	249	222	249	222
<i>R</i> ²	0.0295	0.0234	0.0486	0.0537	0.1043	0.1075

Note: * $p < 0.05$. Numbers in parentheses are standard errors.