

Cyberspace, the Web Graph and Political Deliberation on the Internet

Kenneth N. FARRALL
Annenberg School for Communication, University of Pennsylvania
Philadelphia, PA 19104, USA

and

Michael X. DELLI CARPINI
Annenberg School for Communication, University of Pennsylvania
Philadelphia, PA 19104, USA

ABSTRACT

Ongoing debate about the effects of the internet and computer mediated communication on political deliberation and democracy has been limited by traditional research methods and a general failure in social science to recognize that cyberspace is a fundamentally new social space with its own laws. Although the mapping of symbolic networks emerging within cyberspace remains a difficult problem both theoretically and methodologically, this paper argues that shadows of these networks can often be discerned from the lower order web graph. The paper demonstrates that established graph patterns in social network theory such as centrality, marginality and betweenness can be readily observed in visualizations based on crawls of web sites oriented toward targeted issues. Further, there is a correspondence between these patterns and the virtual social network behind them, although in ways that are not always consistent with traditional interpretations. Finally, the paper begins to outline the broad strokes of a hybrid, interactive research method involving both network and content analysis for the purposes of understanding the growing roll of cyberspace in political deliberation.

Keywords: web graph, cyberspace, hyperlink network analysis, web graph sociology, deliberation, democracy, social network analysis.

1. NETWORKS AND THE STUDY OF POLITICAL DELIBERATION

There is continuing debate in social science circles about the internet's contribution to or hindrance of the kinds of public deliberation deemed critical to democratic life. Some [1] [2] [3] have argued that the internet is inherently democratizing, dramatically increasing access to information and allowing for interactive exchanges among citizens without regard to geographic or temporal boundaries. Others worry that the internet's ability to increase consumer and producer selectivity and allow for information filtering may be inherently fragmenting and polarizing [4] [5]. Empirical investigations of the internet's effects have done little to settle this debate. Until recently, most studies have depended on survey research of net users or content analyses of a small number of targeted online

domains. Results have been mixed. Survey research can give us a snapshot of the population of cyberspace participants, but tells us little about the vectors of influence. Because online social groupings are so numerous and are growing at such a rapid rate, the representativeness of targeted domains remains a significant weakness with these types of studies. Dahlberg summarizes the nature of the research problem:

...how can such an evaluation be undertaken given the vast amount of diverse activity taking place through many thousands of Internet sites? Many studies of online discourse, including cyber-democracy research, limit their focus to very specific aspects or sites of cyber-communications that can be closely scrutinized with particular social science research tools. Unfortunately, the findings of narrowly focused cyber-research are also quite limited. Given the many and complex dimensions of online discourse, generalizations about the relationship between the Internet and society at large cannot be made from research into single sites and/or by applying single methods. [6]

Social Network Theory

Implicit in Dahlberg's critique is the issue of whether one conceptualizes the internet as simply a communication tool or medium akin to newspapers, telephones or television, which individuals use, or, more ambitiously, as a virtual (i.e., cyber) space in which individuals, and increasingly groups, institutions and sociopolitical processes, are located. Social network theory (SNT) offers a potentially rich framework for this latter conceptualization, as well as for theorizing and testing the relationship of cyberspace to democratic practice. Cyberspace is clearly some form of network (or more accurately, a network of networks) and SNT already has formal, structural definitions of certain key relational attributes. These network attributes - for example, isolation, cohesion, centrality and marginality - which can be useful in identifying patterns of fragmentation or integration within social systems, are of obvious relevance to political deliberation.

One problem, however, is that until recently SNT has traditionally identified individual people as the unit of analysis, or node. To graph the network in which an individual found himself the researcher would ask them to name people who have certain important connections to them and the type of relation

that was involved. Obviously, to ask an individual to identify all the cyberspatially mediated social networks of which she is a part would be an exercise in futility. To sidestep this problem, the Computer Mediated Communication (CMC) field has attempted to use SNT to measure how traditional social networks (for example, within a corporation) are affected by the introduction of new communication technologies such as such as email or newsgroups [7].

2. CYBERSPACE: A NEW SOCIAL SPACE

While useful, such approaches avoid the larger question of the nature and structure of the growing number of globally-distributed virtual social networks, and their relationship to or independence from more traditional networks. Are these virtual networks as fleeting, ethereal and trivial as they often appear, or are they in fact fundamentally transforming the dynamic of public deliberation and democratic discourse? While some CMC and SNT researchers are beginning to address at least aspects of this more fundamental question [8] [9] [10] [11] it remains a controversial area of study, in large part because of continuing skepticism over whether cyberspace represents a fundamentally new social space with its own dynamics, laws, networks and implications for community:

Ham radio operators have a global network of friends and acquaintances who came together solely through their use of that instrument. Do they exist in "hamspace"? And why is the manner in which people make first contact so significant? Do pen pals exist in "penpalspace"?

One reason that cyberspace is described as a place is to avoid downgrading it to the status of a mere medium, and perhaps especially to avoid comparisons with television. Those who would distinguish the Internet from television point out that Web denizens are not mere passive recipients of electronic signals. That may be (partly) true. But telephones and the postal system are also communications media that allow two-way communication. We don't regard them as places. [12]

This argument is, in our opinion, easily dismissed. The reason we do not regard the postal system and telephones as places is that they are mediums of transmission and transfer, not storage. For a space to exist in any sense of the word it must be able to contain objects, objects which tend to persist over time and can interact on the basis of certain system rules (the laws of nature, for example). Space must be able to store the objects, their attributes and the laws which govern their interaction, and must also be able to mediate their interactions. Telephones and short wave radio sets can mediate interactions, but do not store anything. The spatiality of cyberspace is manifested in its unique ability to combine storage, transmission, and algorithmic processing into a single medium. This has provided an environment for the persistence and rule-based interaction of digital objects over time. Further, these digital objects (data) exist at multiple levels of ontological expression, from numbers, to texts, to images, to procedures (algorithms) affording the evolution of complex spatial ontologies and symbolic networks.

Wynn and Katz have argued that, because the Internet in no real sense disembodies the person as postmodernists claim, the Internet is not fundamentally different from any other communications medium. Social issues about the Internet are best addressed as "questions of emerging structures of

interaction and reorganization of social boundaries that can occur in any medium of communication [13]."

The realization that cyberspace is a space does not require that human bodies can occupy this space in any sense (though it does not preclude the possibility either.) Our bodies remain where they are, but cultural products and social meta information such as "reputation" and "trust" now increasingly reside in cyberspace [14]. And these objects interact under different rules, as digital bits instead of physical atoms [15].

3. SHADOWS IN THE WEB GRAPH

Although the rhizomic nature of virtual symbolic networks in cyberspace is virtually impossible to visualize or map in any coherent way, shadows of this order may be discernible if we look at lower dimensional symbolic networks within this space such as the *web graph*. Using the mathematical language of graph theory, the web graph, in its entirety, contains a node for every document on the World Wide Web as well as directed edges for every hyperlink between them.

A range of literature in computer science and mathematics has discussed the macro structure and character of the web graph [16] [17] [18]. Understanding of the deep structure of the web and how it encodes authority and trust has powered the success of Google and other search engines. A common topic in this literature is the idea of a web community defined by formal graph definitions [19] [20] [21] [22]. Although much of this literature recognizes the potential sociological benefit of these concepts, their focus is on the achievement of better search algorithms. From this perspective, a community is often a topic cluster with actors who may have very little awareness of each other.

We believe that the web graph, because of the extreme range of socially contingent data that it encodes and its rapid rate of growth and change, should be recognized as an important domain for sociological research. The local structure of the web graph, defined by the link space surrounding a specific node or group of nodes, can be discerned and often visualized through targeted crawls of these links. Many questions about the meaning of these graphs, however, require further study, most importantly the nature of virtual data-based networks and their interaction with the human social networks that produce them.

Does a hyperlink link between a pro-choice web and a pro-life web community necessarily indicate that the two are mutually aware and interact with each other? The number of links, whether or not they are reciprocal, and the nature of the links will clearly provide a more nuanced picture, but there will always be a gap between the hyperlink connections and their human counterparts. And the attributes of the links between nodes vary considerably as ones eye scans the two-dimensional surface of the map. Social science researchers must develop theory and technique to bridge this gap with confidence.

Although this variance and uncertainty in the web graph would appear on the surface to discount the value of visualizations, link attributes often occur in clusters and recurring patterns. The web graph can often collapse complex social information into patterns which provide direct insights into the nature (situations and practices) of the social systems they mediate as well as provide an efficient orientation for further investigation.

between human minds by their shadows in the space of the hypertext graph.

Effective deliberation is usually preceded by dialogue between more like minded actors. Web links between actors can reflect both dialogic and deliberative relationships. Marginality has a difference essence depending on the relationship. As we will see, marginality in the web graph might also reflect a logistical support orientation. And there is inevitably more variation in the web graph. An obvious way of getting past this limitation is to go beyond the map and consider the space it reflects. What is the nature of the specific web sites that appear in the graph? What function do individual links play? To be used effectively, analysis of the web graph must be integrated with additional contextual data.

Issues are deliberated within and between political parties, as well as religious groups and nation states. At times the

participants share world views while at other times they are in opposition. Issues can range in specificity and comprehensiveness. Often, issues can be conceived within hierarchies of their comprehensiveness, such as moving from the question of toxic pollution in the oceans, to the health of our biosphere, to the nature of life in our universe. Issues can exist in clusters, like the Patriot Act, Privacy, and Constitutional Rights; they can exist in hierarchies (such as oceans to the biosphere to life); or the more complex associations that inform world views.

Our initial exploration of the web graph using the Issue Crawler tool suggests that many aspects of deliberation discussed above are encoded in the map visualizations.



Electronic Voting Issue Network (figure 2)

4. CASE STUDIES

The remainder of this paper presents graph visualizations from three targeted issue crawls using the Issue Crawler. To follow the general arguments made about these issue maps, you will not need to be familiar with the issue at hand. To evaluate the specific interpretations we provide or to form your own interpretations, however, some familiarity with the issue is

critical. The first case study, the *progressive left issue network* (figure 1), should be familiar to virtually all academics. The *electronic voting issue network* (figure 2), the second case study, is an issue of growing debate in the United States concerning the use of privately-owned voting machines to count non-physical ballots in local, state and national elections. The *Radio Frequency Identification (RFID) technology issue network* (figure 3), the third and final case study, concerns the insertion

of small, digital identification tags in the world' s physical and living objects (from consumer products to human beings) for a wide range of applications.

The Issue Network and its Seed URLs

The Issue Crawler generates an association matrix from a set of seed URLs provided by the researcher. The seed is an attempt by the researcher to represent some of the key actors in the issue network of interest. To assume that the actual results of the crawl represent the issue network as it has been defined is perhaps one of the most common errors of web graph

interpretation. A common result , for example, is to target a narrow issue and get back a web graph of a more general issue that includes it. When the researcher has some familiarity with the issue he is mapping, recognizing such differences is usually a simple matter of consulting the nodes' domain names. At other times, the crawler does not return any form of network. This might be because there is no virtual network within the web graph that corresponds to the issue under question, or it might be that the researcher is not familiar with the more important, central nodes.



RFID Issue Network (figure 3)

It is our general experience and belief that the value of web graph visualization, assuming a researcher's basic familiarity with an issue area, is not highly dependent on specific choices within the initial crawl seed. It remains an important question to be explored with further research, however.

Our goal is to point towards interesting areas for more research, not to offer specific conclusions about the issues that we targeted in the crawls or to offer final determinations about the meaning of particular web graph patterns. The patterns themselves and their interpretation is a dynamic process between the researcher, the raw space of the web graph, and the virtual

symbolic network it encodes. There are numerous subjective decisions involved in an issue crawl such as the initial choice of seed URLs and the breadth (number of iterations) and depth of the crawl. Further, once the association matrix is generated, a specific visualization from the graph is determined by a range of parameters, such as inter-node link thresholds (the higher the number the lower the graph density.) Nevertheless, we believe it is apparent from the resulting graphs that these visualizations encode a very large amount of information, information that may be perceived instantaneously, or which can be coaxed out via consultation of the content and context of the graph nodes. There is a correspondence between these patterns and the virtual

social networks behind them, although in ways that are not always consistent with traditional interpretations.

The first graph (figure 1) resulted from a broad crawl of the progressive left issue space using the following seed domains: *alternet.org*, *buzzflash.com*, *commondreams.org*, *counterpunch.org*, *democraticunderground.com*, *guerrillanews.com*, *indymedia.org*, *moveon.org* and *tompaine.com*.

The first set of valuable data is the total set of network actors identified in the visualization. Using a seed of only 9 URLs, the issue crawler has identified a network approaching 100 members. For any one with existing familiarity with this issue network, an inspection of the domain names labeling each node will reveal that many (perhaps most) key actors in the progressive left web space have been identified. In this visualization, the size of the nodes varies according to in-link centrality, or the total number of links within the network which point to the node. High in-link centrality relative to other nodes in a web graph tends to be a strong indicator of a node's social authority within the web graph [17]. Thus, large nodes in the graph, such as *thenation.com*, *fair.org*, *moveon.org* and *alternet.org* carry significant authority within this progressive left issue network.

Many of the smallest nodes in this graph are clustered on the left-hand side. Though their lack of size and obvious network marginality would suggest an interpretation that these nodes lack authority and importance within the issue network, in this case marginality reflects something different. In the process of discourse within this issue network, actors will reference daily news papers. These papers do not link back to the network but may be referenced again and again over time. In this case marginality reflects a certain logistical or functional orientation of the primary network under study within the broader social space. It is interesting to note here that *foxnews.com* is one of the most highly referenced of these daily news sources. An inspection of the context of these links would no doubt reveal a better interpretation of this form of in-link centrality than simple authority.

The second graph (figure 2) resulted from a more targeted crawl of sites in the electronic voting issue network and used seed URLs from the following domains: *lorrie.cranor.org*, *blackboxvoting.org*, *epic.org*, *notablessoftware.com*, and *verifiedvoting.org*. There is an obvious difference in the network structures of the first and second graph. The electronic voting issue network is not nearly as dense, as well developed, as the progressive left issue network. There are at least two obvious reasons for this difference. First, the progressive left "issue cluster" is far broader and far more inclusive than the specific issue of electronic voting. Second, the electronic voting issue is newer, giving it less time to develop than older issues. Those familiar with the issue of electronic voting will recognize that key research, industry and government actors are identified in this graph. In this visualization, the size of the node varies according to total degree centrality (in and out-links). *EPIC.org*, the web site of the *Electronic Privacy Information Center*, exhibits a high degree of both in and out-link centrality. The *EPIC* site has direct, outgoing links to most of the key actors in this network, suggesting a high degree of issue awareness and influence within this space. Those familiar with eigenvector centrality [24] will note that this graph clearly illustrates this measure through the overall positioning of *epic.org*.

Finally, the third graph was generated through a targeted crawl of the *RFID* issue space. Seed URLs from the following domains were used to produce the crawl: *wikipedia.org*, *aimglobal.org*, *epic.org*, *junkbusters.com*, *nocards.org*, *rfidprivacy.org*, and *spychips.com*. The size of the nodes in this graph represent in-link centrality. Those familiar with this issue may note that the graph does not include some key actors in this issue space such as *epcglobalinc.com*, the multi-industry global *RFID* standards body, *autoidlabs.org*, the MIT research center, and *stoprfid.org*, an emerging *RFID*-awareness activist site. *Stoprfid.org* is less than a year old and has not yet established itself in the web graph, despite being present in the initial seed. Instead of *EPC Global* and *Autoid Labs*, the graph includes *autoidcenter.org*, a now defunct MIT-sponsored web site that preceded them. This raises the possibility of "graph lag" where more recent developments in the virtual network may not yet be apparent in the graph, as well as the possibility of seeing the emergence of more established, mature networks over time in successive web graph visualizations.

A significant result from this crawl was the emergence of the higher order issue network of electronic privacy and human rights from the *RFID* seed domains and the orientation of this network toward the narrower *RFID* issue space. It appears from the graph that the electronic privacy and human rights issue network, which consists of such sites as *epic.org*, *eff.org*, and *privacyrights.org*, is more mature, more cohesive and more dense, than the *RFID* awareness network, which is clustered around two primary actors: *nocards.org* and *boycottbenneton.org*. Further, it is apparent from the map that the privacy network appears to be collectively oriented toward the *RFID* cluster in a reciprocal (bi-directional) relationship. Given that *RFID* is one issue among a cluster of issues that the "privacy community" considers, it is understandable that the privacy network may be more mature and advanced, but that this relationship is visible in the web graph is a further indication of important correspondence to real social relations.

5. CONCLUSION

We hope that the few examples above serve to demonstrate, at least, the utility of web graph mapping for understanding social phenomena and further, the promise of web graph sociology for increasing our understanding of the nature of political deliberation and the internet. Further interface between the web graph generation process and reflection on social phenomena will no doubt help to build on our understanding of the web graph taxonomy as will continued experimentation with the techniques of graph generation and display.

One obvious advantage of web graph research is the dramatically improved access and retrievable possibilities of the targeted data. It is out there in cyberspace and can be retrieved at regular intervals with considerable efficiency and speed. Data collection improves dramatically compared to person to person survey techniques common in *SNT* research.

High node centrality in the web graph is well established as providing significant social indicators such as authority, trust, knowledge and reputation. Nodes that appear as "marginal" on a given web graph can reflect a wide variety of system relations, from socially marginal issue positions to supporting roles that may be limited in time and space. An understanding of these

various flavors of marginality can often be drawn from deeper inspection of the graph, especially the content of the nodes. For example, daily newspapers serve a reference function for the progressive left issue network without being network members in a traditional sense. Nevertheless, this orientation of the network within other network spaces is of obvious analytical relevance. And relationships between and across issues may be discernible within certain web graphs, such as the higher order issue network of electronic privacy and human rights and its orientation to the highly relevant but far more focused issue of *RFID* technology.

Although the presence or lack of correlation between web graphs (where web documents, web sites, or clustered web communities might act as nodes) and social networks (where people act as nodes) remains an important research question, it is important to recognize that these virtual networks represent something in and of themselves. The emergence of cyberspace as a new social space with its own laws requires new forms of network analysis that go beyond traditional assumptions.

Most research today generates graphs using web documents or web domains as nodes. It should also be possible to graph relationships between clusters of web sites or documents that reflect certain forms of "web communities." Better, more illustrative pictures of virtual deliberation may emerge at this level of analysis. The definition of web community, however, has had varying levels of strictness in computer science literature. A highly restrictive formal definition of web community will yield a far smaller number of communities than looser definitions. Further, similar structural communities can have very different characteristics and functions that are critical to understanding their role in community formation and fragmentation.

The development of hybrid graphing methodologies using both link crawls and node-by-node content analysis is likely to further the understanding of virtual social communities and their deliberative dynamics in cyberspace. Links are not just structures; they are specific forms of relation. Content-analysis can help us to identify some of the more common forms of association and perhaps distinguish between discursive and deliberative communities.

Finally, further research should explore the evolution of particular web graph visualizations over time. In the second case study, we noted that the graph had not yet picked up the emergence of certain key *RFID* actors such as *EPC Global*. It is highly likely that these actors will appear in subsequent crawls, along with other nodes as they network grows and matures. Issue networks which grow rapidly over time exhibit different dynamics than those which remain static or even disappear.

Web graph analysis may help measure the success of various interventions that have been suggested to improve the internet's function as a virtual public sphere. For example, one could conceivably measure the effectiveness of public supported online deliberative spaces that have been proposed by Sunstein, et al. What does the local web graph look like? Can we identify interconnected communities of conflicting issues and world views? How does the picture compare to more self-interested web communities emerging spontaneously on the Web? Does a heavy commercial orientation such as the AOL internet experience, make a difference in fragmentation tendencies? Can

we see evidence of walled gardens [25] in the corresponding web graph?

We must learn to identify what kinds of parameters matter most (and what parameters are most readily mappable) in understanding and ultimately fostering healthy global societies with interactions across world view and issue communities. Such an understanding is critical if humans are to use cyberspace to revitalize democracy and not let it usher in new levels of polarization and extremism.

6. REFERENCES

- [1] M. Hauben. & R. Hauben, **Netizens: On the History and Impact of Usenet and the Internet**, Wiley-IEEE Computer Society Press, 1997.
- [2] H. K. Klein, "Tocqueville in Cyberspace: Using the Internet for Citizen Associations," **The Information Society**, 15, 1999, pp. 213-220.
- [3] D. Kellner, "Habermas, the Public Sphere, and Democracy: A Critical Intervention," in L.E. Hahn (Ed.), **Perspectives on Habermas**, Open Court Press, 2000.
- [4] C. R. Sunstein, **Republic.com**, Princeton, NJ: Princeton University Press, 2001.
- [5] N. Zatz, "Sidewalks in Cyberspace," **Harvard Journal of Law & Technology**, vol. 12, 1998, pp. 151-240.
- [6] L. Dahlberg, "Computer-Mediated Communication and The Public Sphere: A Critical Analysis," **Journal of Computer-Mediated Communication**, 7(1) 2001.
- [7] C. Haythornthwaite, "A social network theory of tie strength and media use: A framework for evaluating multi-level impacts of new media. Technical Report," UIUC LIS -- 2002/1+DKRC, **Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign**, Champaign, IL, 1999.
- [8] L.A. Adamic, "The small world Web," **Proceedings of 3rd European Conference of Research and Advanced Technology for Digital Libraries**, ECDL, 1999.
- [9] R. Rogers, & N. Marres, "Landscaping climate change: A mapping technique for understanding science and technology debates on the world wide web," **Public Understanding of Science**, 9, 2000, pp. 141-163.
- [10] C. Hine, **Virtual ethnography**. London, UK: Sage, 2000.
- [11] A. Halavais, & M. Garrido, "Mapping networks of support for the Zapatista movement," In M. McCaughy, M., & M.D. Ayers (Eds.), **Cyberactivism: Online activism in theory and practice**, London: Routledge, 2003.
- [12] J. Koppell, "Why cyberspace isn't anyplace," **Atlantic**, August 2000, pp. 16-18.
- [13] E. Wynn. & J. Katz, "Hyperbole over Cyberspace: Self-presentation & Social Boundaries in Internet Home pages and Discourse," **The Information Society**, 13(4):, 1997, pp. 297-328.
- [14] P. Resnick, "Beyond Bowling Together: Sociotechnical Capital," In J. Carroll (Ed.), **HCI in the New Millennium**, Addison-Wesley, 2002, pp. 247-272.
- [15] N. Negroponte, **Being Digital**, New York: Alfred A. Knopf, Inc., 1995.
- [16] L. Terveen & W. Hill, "Evaluating Emergent Collaboration on the Web," **Proceedings of the ACM CSCW'98 Conference on Computer Supported Cooperative Work, Social Filtering, Social Influences**, 1998, pp. 355-362.
- [17] J. Kleinberg, "Authoritative sources in a hyperlinked environment," In **Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms**, 1998.

- [18] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar,, A. Tomkins, et al., "The Web as a Graph." **Proceedings 19th ACM SIGACT-SIGMOD-AIGART, Symposium on Principles of Database Systems**, 1-10, 2000.
- [19] G.W. Flake, S. Lawrence, C. Giles. & F. Coetzee, F., "Self-Organization and Identification of Web Communities," **IEEE Computer**, 35(3), 2002, pp. 66–71.
- [20] D. Gibson, J. Kleinberg, & P. Raghavan, "Inferring Web Communities from Link Topology," **Proceedings of 9th ACM Conference on Hypertext and Hypermedia**, 1998, pp. 225-234.
- [21] R. Kumar, P. Raghavan, S. Rajagopalan, & A. Tomkins, A., "Trawling the Web for Emerging Cyber-Communities," **Computer Networks**, 31(14), 1999, pp. 1481-1493.
- [22] P. Reddy, & M. Kitsuregawa, "An approach to build a cyber-community hierarchy," Paper, **Institute of Industrial Science, University of Tokyo**, 2002.
- [23] H. W. Park, "Hyperlink Network Analysis: A New Method for the Study of Social Structure," **Connections**, 25(1), 2003, pp. 49-61.
- [24] P. Bonacich, "Technique for analyzing overlapping memberships," in **Sociological methodology**, H. L. Costner (ed.), San Francisco: Jossey-Bass, 1972, pp. 176-185.
- [25] D. Winseck, "Netscapes of Power," In David Lyon (Ed.), **Surveillance as Social Sorting**, London and New York: Routledge, 2003, pp. 176-198.