From Wayback Machine to Yesternet: New Opportunities for Social Science

William Arms[1], Dan Huttenlocher[1], Jon Kleinberg[1], Michael Macy[2], and David Strang[2]

[1]Department of Computer Science, Cornell University, USA

[2] Department of Sociology, Cornell University, USA

Corresponding author: Michael Macy, mwm14@cornell.edu

Abstract. Social scientists have stores of data on individuals and groups but relatively little on social interactions, the basis of all social life. That is likely to change due to the spread of computer-mediated interactions that leave a digital record. The flood of available on-line information – from corporate web pages to news groups, wikis, and blogs – has the potential to open up new frontiers in social science research on the diffusion of innovations and beliefs, the self-organization of on-line communities, and the collective behavior of individuals. The Cornell Yesternet project will create a research laboratory for social science research based on the Internet Archive's 40-billion page Web collection. These snapshots of the Web have been captured and archived every two months for nearly ten years. The Yesternet project will copy and reconfigure large portions of this massive collection as a relational database that can be used for research on social and information networks. The Cornell team, composed of social, computer, and information scientists, will develop, test, and refine the necessary tools as part of a series of testbed research applications that track the diffusion of innovation on the Web.

The Challenge

Over the past century, social scientists have collected stores of data about individuals, using surveys and records kept by governments and employers. Individual-level data is then aggregated as population statistics for groups of varying size, from local communities to nation states. In comparison, we have frustratingly little data about the interactions between people who influence one another in response to the influences they receive. Yet social relations and interactions are the foundation of social life. What we most need to understand is also what we know the least about. The reason is obvious. It is much easier to observe friends than to observe a friendship. Social interactions are fleeting – one needs to be watching at precisely the right moment. Moreover, the data are also tedious to hand-code and record, given the nuances of interaction and the exponential increase in the number of

relations as the size of the group increases. As a consequence, studies of interactions in networks tend to be static (Moody, McFarland, and Bender-deMoll 2005), limited to the structures of interaction without regard to content (Emirbayer and Goodwin 1994), and based on very small numbers of nodes. For example, Newcomb's (1961) classic study of an acquaintance network involved only 17 residents of a boarding house yet required several years of work by a team of graduate students who lived with and recorded observations of interactions.

The agonizing difficulty observing social life at the relational level is about to change, thanks to computer mediated social interactions that leave a digital trace. In particular, the explosive growth of on-line networks over the past decade has created vast quantities of semi-structured data such as Web pages, email messages, blog postings, and chat room logs. These data hold enormous potential for empirical social science research that has thus far been largely untapped.

The problem now is not the scarcity of data but its abundance – how to download, store, structure, and search Web-scale data. The challenges for researchers are multi-fold:

- Content and structure are implicit rather than explicit, unlike traditional tabular databases with specified fields and relations. New tools are needed to parse data into meaningful parts and structures.
- Data results from independent actions of many agents (both individuals and organizations) rather than by being gathered centrally for particular purposes. Thus the paradigm shifts from one of data collection and analysis to one of data mining and information discovery.
- The sheer scale of billions of items poses both computational and conceptual challenges. Data intensive computing, tightly coupling petabytes of storage with teraflops of processing power, is at the edge of what is technically feasible over the next two years. Manually coding such vast quantities of data is beyond human capabilities and requires new tools such as semi-supervised machine learning.
- Even for data gathered from publicly available sources such as the Web, privacy is a substantial concern. Combining, analyzing and mining data can easily reveal information that was not apparent in the original sources. Thus privacy preserving data mining and discovery techniques must be employed.

To address these challenges, we assembled a team of social scientists whose research stands to benefit from the use of on-line network data, alongside computer scientists with expertise in very large semi-structured datasets, privacy, machine learning, natural language processing, data mining, information retrieval, relational databases, and digital libraries.

The project team is drawn together by the enormous promise of a unique and largely untapped dataset: the Internet Archive's 40-billion page collection of Web pages. The Internet Archive consists of snapshots of the Web collected and archived about every two months for nearly ten years. Realizing this potential will require successfully meeting a daunting technical challenge. The Internet Archive is larger in scale and more heterogeneous in content than any other semi-structured social science dataset we know of. In its current form, this dataset comprises a vast archive that can be accessed only via individual Web page URLs. We are copying and transforming the dataset into a relational database we are calling the "Yesternet" that will be an accessible resource for social science research. The project exploits a unique facility currently being built at Cornell with NSF support, the "Petabyte Storage for Data Driven Science."

Computer scientists have learned through experience that it is usually best to build software tools in close collaboration with users. Our success depends critically on using the tools, as they are being developed, to address specific research problems. Hence, our project is two-fold – to build an intelligent front-end that will make the Internet Archive widely

accessible to social scientists, and to develop, test, and refine these tools through specific research applications on a problem of broad theoretical and practical interest – the diffusion of innovations.

II. The Yesternet and the Diffusion of Innovation

Studies of the diffusion of innovation have attracted widespread interest across the social sciences. Sociology, anthropology, geography, economics, organizational behavior, population ecology and communication studies all have long-standing traditions of research on the means by which ideas and practices spread across the social landscape. Applications include the diffusion of medical technologies like new prescription drugs (Coleman, Katz and Menzel 1966), diseases like AIDS (Watts 2003), organizational policies like affirmative action (Edelman 1990), and collective actions like the sit-down strike (Morris 1981). Studies such as these have developed a wealth of empirical insights about the characteristics of innovators, early adopters, and laggards; the networks that channel contagion; and the opportunities and obstacles faced by "change agents" (Rogers 2003).

Despite the accumulation of knowledge about particular cases and contexts, cumulative theoretical progress has been halting. One reason for this lack of theoretical progress is that research strategies and technologies have been remarkably static. Ryan and Gross's (1943) landmark hybrid corn study "established the customary research methodology to be used by most diffusion investigators: retrospective survey interviews in which adopters of an innovation are asked when they adopted, where or from whom they obtained information about the innovation, and the consequences of adoption" (Rogers 2003, p. 33). These painstaking reconstructions suffer from important limitations:

- Restricted scale, on the order of hundreds of adopters, with hard-to-collect network data often a limiting factor.
- Restricted geographical scope, often country specific, rarely global.
- Little analysis of the earliest stages of diffusion, before the innovation has spread sufficiently to attract the attention of pollsters, reporters, or scholars.
- Limited knowledge of the full life cycle of diffusion. While upswings in adoption are much studied, little is known about downswings and innovation abandonment (which is often an unpublicized non-event).
- Limited attention to unsuccessful innovations that fail to diffuse widely.
- Limited analysis of the branching, evolving, and mutating content of the innovation.
- Limited comparative inquiry across cultures, innovations, or units of analysis.

We can address these limitations by tracking the diffusion of innovation using data from Web pages. Our strategy moves away from separate collection of hand-coded data for each innovation, in favor of a vast on-line repository that can be accessed for many large-scale studies using intelligent search and manipulation tools. In marked contrast to small, local case studies, the Yesternet will allow a global and cross-cultural perspective. This breadth need not come at the expense of depth. For example, Web reports of the spread of protest events can be used in much the same way researchers currently use newspaper accounts, with the added benefit that Web reports often provide links to extensive primary source material from the protest organizations themselves. Because the Web has no editors, it provides access to innovations that never took off and which therefore never reached the attention of those who screen which data get collected and archived. We can observe not only the spread of innovations but also the decline, as well as the mutations and branching – tasks that are not possible when data collection is hand coded using predefined terms to search for a specific entity.

We suspect the greatest potential of the Yesternet will be to suggest new questions and modes of inquiry. For example, intelligent search tools could permit researchers to discover innovations that spread the fastest or the farthest over a given time period and those that failed. More generally, the inductive approach allows researchers to focus on the profile of an innovation, and then find content that fits the form. Instead of generalizing from innovation-specific cases, we will be able to create macrosocial datasets in which innovations, not adopters, are the units of analysis. By comparing diffusions between different network structures and cultures, we will be able to gain insight into contextual effects.

To date, the use of Web data to study diffusion of innovations is itself an innovation that has not yet diffused. A few studies have appeared (Preece and Maloney-Krichmar 2003, Kollock and Smith 1996, Price et al. 2003), and there are proposals for using Web data to study diffusion processes (Rangaswamy and Gupta 1999). The Web has also been used to study the spread of social movements and political influence. For example, Adamic and Glance (2005) mapped the "political blogosphere" of "red" and "blue" advocates in the 2004 election (see Figure 1). They found that conservative blogs tend to link to one another more frequently than do their liberal counterparts.

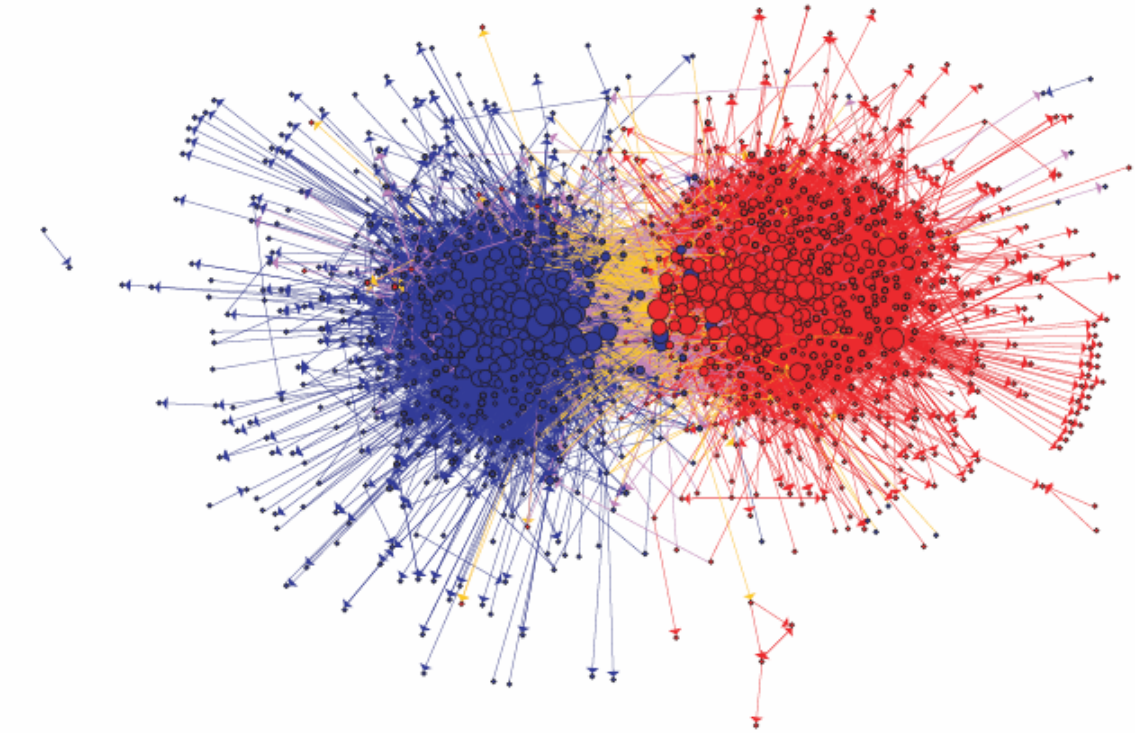

Figure 1: Community structure of political blogs, from Adamic and Glance (2005, p. 4). Orange links go from liberal (blue) to conservative (red), and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Following Adamic and Glance, Robert Ackland's team at Australian National University is also using Web data to study differences in network structure between left-wing and right-wing social movements in Australia, including pro-life/pro-choice, left and right extremist parties, and environmental groups (Ackland et al. 2006). They developed the Virtual Observatory for the Study of Online Networks (http://voson.anu.edu.au), Web-based software that incorporates Web mining and data visualisation. Using VOSON, they also found that right-wing groups tend to be more densely tied to one another compared to those on the left.

They are currently working hard to see if this pattern extends to still other groups in other places, and to look for clues as to why this might be the case.

The primary reasons that more researchers are not yet using the Web are the technical difficulties in gathering and analyzing the data. Researchers must not only be able to use a Web crawler to collect data, but must also master techniques for natural language processing and machine learning in order to build effective filters that can search through gazillions of pages. Researchers then confront massive amounts of poorly structured data that must be organized into a usable form. Moreover, a crawl of the Web yields only a current snapshot. Temporal phenomena such as diffusion are impossible to study without gathering many crawls over a substantial time period.

Pioneers in this area know what we need: "a centralized solution for archiving and access" (Halavais 2003). The data for this centralized solution already exists: the Internet Archive. The question is whether this data source will remain limited to its current archival purposes, or whether social and computer scientists will work together to aggressively exploit this remarkable opportunity to observe very large-scale social interactions that leave a digital record, without the prohibitive need to start each study from scratch.

III. Research Applications: Opportunities and Limitations

A wide variety of innovations are studied within a diffusion framework, including the spread of new technologies, organizational forms, fads and fashions, norms, beliefs, and urban legends. Other potential applications include tests of mimetic models of the spread of virus-like social, political, and religious innovations. Below, we sketch in greater detail several applications that illustrate the research opportunities that lie ahead and how they might be exploited. It is important to recognize, however, that these applications barely scratch the surface of the research opportunities that will open up when the Yesternet comes on-line.

The Evolution of Innovation

By combining natural language processing with structural analysis of changes of Web data over time, it becomes possible to track not only diffusion but also the branching that occurs when an innovation mutates into new forms that then spread independently. This is very hard to do when data collection is hand coded using predefined terms to search for a specific entity. Language processing tools developed by our project will make it possible to develop search algorithms that can not only locate a particular innovation (such as an urban legend) but all its "first cousins," "second cousins" and so on. Researchers will be able to compute not only the rate of diffusion, but also the mutation rates. This will extend research interest to areas including organizational ecology, memetics, and the study of social movements.

From Sigmoids to Bursts

Classical diffusion models posit a characteristic S-shaped growth curve that reflects the need for a critical mass on the one side and the limiting effects of a bounded population of potential adopters on the other. Feedback loops are one of the mechanisms for critical mass. Book sales increase gradually until the book makes the bestseller list and then leap ahead. Protests grow in dribbles until they get large enough to attract media attention. Recent research using on-line data suggests that the classical model may obscure local bursts that can only be observed using fine-grained longitudinal methods. Web data gives us an unprecedented ability to observe these bursts and their repercussions. In studying profiles for the adoption of innovations, “burst analysis” suggests the beginnings of a framework for understanding the interplay between discrete and continuous phenomena in the diffusion of ideas and content. Consider, for example, the typical dynamics by which a Web site or blog

rises to popularity. Is it a gradual climb, or a series of relative plateaus punctuated by abrupt rises? Or is it a mixture of the two? Do bloggers tend to become well-known because of the exposure gained from a few widely-read postings, or because of steady effort over a period of months or years? Can we characterize ideas, innovations, and content by those whose rise to prominence follows a more continuous trajectory, and those whose rise is more discrete?

Niche Formation

Not all contagions propagate to the global population, even in small worlds. Innovations that compete for adopters tend to carve out niches on the social landscape. Niche formation is difficult to observe using traditional content-specific retrospective studies, due to the narrow focus on one innovation at a time. This is true as well for the related field of organizational ecology, despite the theoretical importance of niche formation. Data collection directed towards specific innovations has focused attention on competition among organizations within a given niche, to the neglect of competition between innovations to attract organizations. This competition is most apparent among technologies, but it also includes competition among beliefs, opinions, and fashions. Open access to the Internet Archive will open up new opportunities for research on niche formation by organizational and population ecologists, as well as cultural sociologists. Research on the spread of religion shows that theological competition promotes diffusion rather than undermining faith. Does this extend more broadly? Do uncontested ideas and practices spread differently than those that are hotly debated (such as stem cell research, abortion rights, same-sex marriage, or euthanasia)? Does contested diffusion promote local clustering and temporal bursts?

Subcultures and Community Formation

The Internet is a great place for subcultures to develop and thrive. The archive will allow anthropologists and ethnographers to track their birth, development, and failure. Moreover, researchers will now be able to compare a large number of different subcultures, which would have been practically impossible using traditional methods of data collection. For example, how do users with common interests gather around a central space and acquire an identity as a distinct community? Tracing the growth of Slashdot, Arstechnica, or any well-defined group of bloggers can help identify patterns of interaction. How do these communities self-organize? What happens once the group defines an explicit identity? Does identity limit growth by defining unambiguous boundaries, or does it promote growth by affirming the group's legitimacy? Is growth limited by the exhaustion of available recruits (as posited in classical diffusion models) or by the carrying capacity of self-organizing communities that lack the material and institutional resources for enforcing norms? Groups like Slashdot have developed a system of user-based moderation that can be used to test theories (including game-theoretic predictions) about the origins of social order.

Kleinberg, Huttenlocher, and two of their graduate students, Backstrom and Lan, are using Web crawls to study the "processes by which communities in a social network come together, attract new members, and develop over time" (2006). They crawled 875 LiveJournal communities to find individuals who were one degree removed, that is, who were not members themselves but were friends with at least one member. They then calculated the probability these individuals would join the community, as a function of the number of friends who were members (Figure 2) and the clustering of ties among these friends (Figure 3).

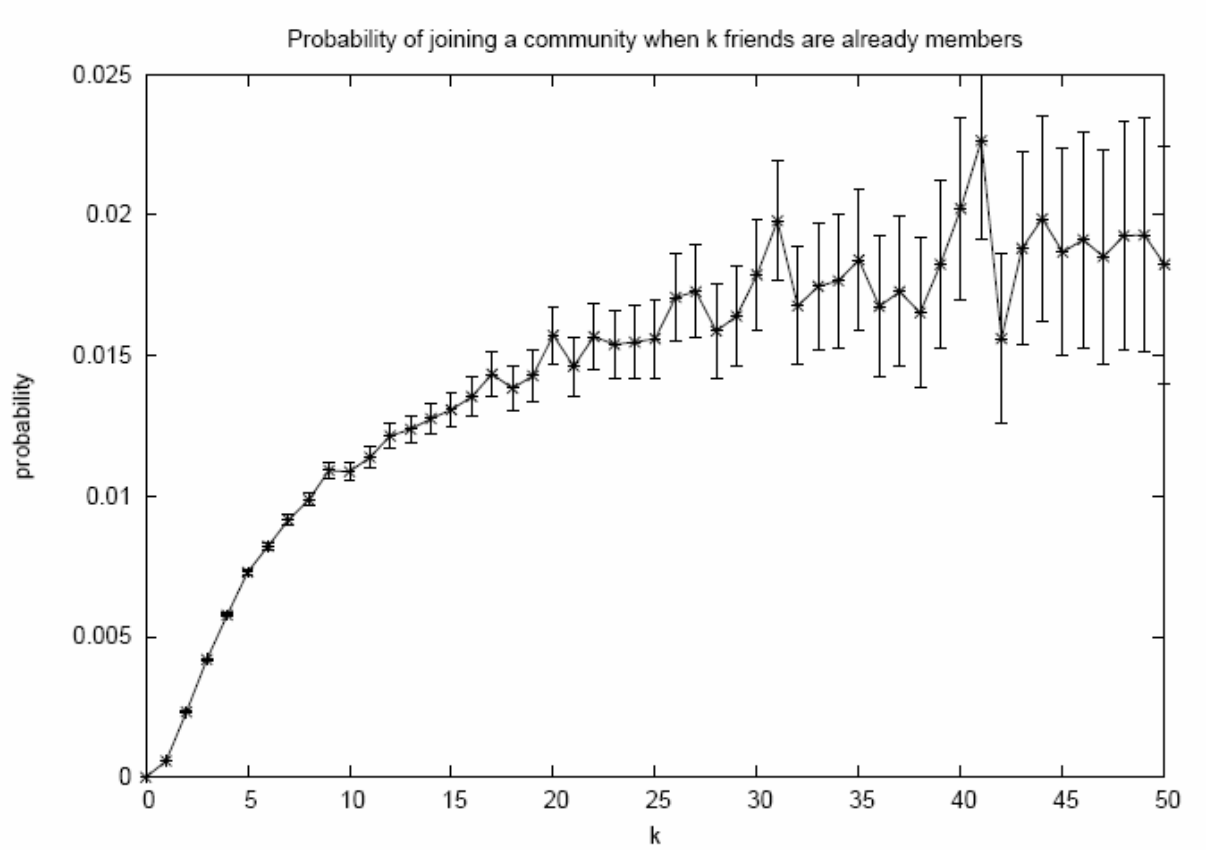

Figure 2: The probability *p* of joining a LiveJournal community as a function of the number of friends *k* already in the community. Error bars represent two standard errors. Source: Backstrom et al. (2006).

Although most diffusion studies display a characteristic sigmoidal probability function for adoption, the growth of LJ communities appears to follow a concave function, suggesting that the threshold for joining is relatively low.

Issue Salience

Political scientists have long recognized that struggles over the items that get on the agenda can be just as important as debates over the items themselves. What causes an issue to become "hot." Perhaps more importantly, how do once-hot issues fall off the agenda? These non-events have been very difficult to observe using content-driven methods of data collection. The Yesternet will make it possible for political scientists and opinion and market researchers to track not only the propagation of positions on issues but also the diffusion of opinion as to what issues ought to be debated.

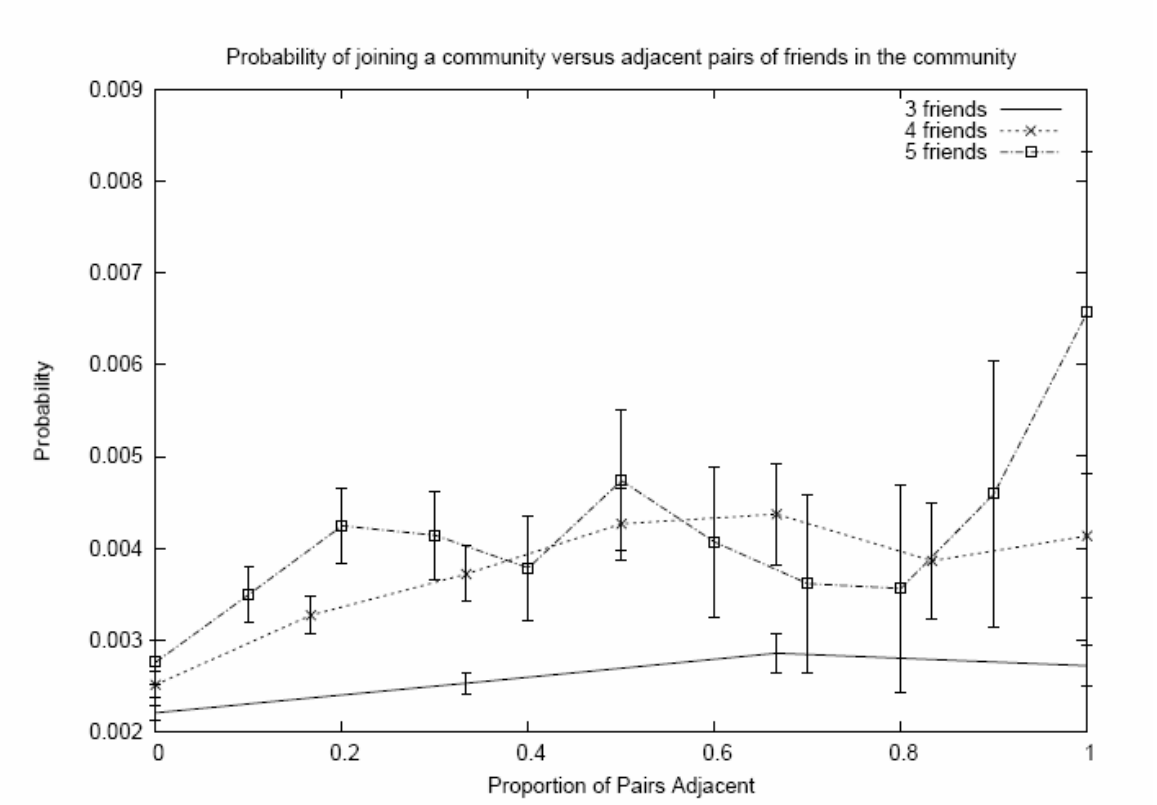

Figure 3: The probability of joining a LiveJournal community as a function of the internal connectedness of friends already in the community. Error bars represent two standard errors. Source: Backstrom et al. (2006).

Emergence and Diffusion of Norms in Online Communities

Access to a ten-year record of the Web will advance cross-disciplinary research on the study of social life in cyberspace. Online interactions are distinctive in the separation from fixed and recognizable demographic and social identities, the relaxation of spatial and cultural constraints on interaction, and the self-organization of authority (Preece and Maloney-Krichmar 2003, Wellman et al. 1996). These conditions are unique in history, providing an exceptional opportunity for research on the emergence, spread, and enforcement of norms – a seminal problem in the study of social order. Must norms be imposed top-down, through formal institutional arrangements, or can they also self-organize through bottom-up interactions among users (Postmes, Spears and Lea 2000, Preece 2004)? Path breaking studies have focused on Usenet (Kollock and Smith 1996, Smith 1999), Multi-User Domains (Reid 1999) and Slashdot.com (Halavais 2001, Poor 2005), but methods of on-line data collection have limited the analyses to the study of a small number of groups (Herring et al. 2002, Postmes et al. 2000), or a single system of norm enforcement (Halavais 2001, Poor 2005). Despite the development of new tools for analysis of Usenet data (Viegas and Smith 2004, Fiore, Teirnan and Smith 2001), our ability to track emergent norms and compare larger numbers of groups remains limited by the problem of observing a rapidly moving target, as communities come and go with changing technologies and fashions (DiMaggio et al. 2001, Wilson and Peterson 2002). The Yesternet addresses these limitations by providing an extensive and permanent record of on-line communities. When the Yesternet is fully implemented, longitudinal analyses of a vast number of online communities will make it possible to compare groups that differ in reliance on formal institutional arrangements (such as moderators), ideological appeal, or sub-cultural identity.

Network Dynamics of Polarization

Polarization is a special case of the diffusion of beliefs, characterized by the spread of competing views. Most diffusion studies assume that adopters positively influence others to adopt, leading to convergence within a population. Studies of polarization introduce the possibility for *negative* influence as well, which has the unintended effect of inducing targets to do or think the *opposite*. This not only causes beliefs to acquire a bimodal distribution, but more importantly, beliefs also tend to become correlated, creating opposing camps of like-minded people (DiMaggio, Evans, and Bryson 1996). Empirical tests, however, are not possible using conventional methods of survey analysis. Surveys are limited to issues that have already attracted the attention of scholars or pollsters, which precludes observation of the emergence and development of a contested issue. In addition, survey data does not provide access to the influence process or the networks within which influence propagates.

On-line discussion provides a unique opportunity to observe the dynamics of polarization. Pundits and academics have recognized the growing importance of news groups and blogs in providing self-organizing frameworks for on-line discussion and debate (Carpini, Cook and Jacobs 2004, Klam 2004, Price and Cappella 2002, Price et al. 2003). Yet rather than fostering greater political and ideological integration, the Internet is increasingly regarded as contributing to the polarization of beliefs, opinions, and attitudes in an era of "culture wars" (Adamic and Glance 2005; Ackland et al. 2006; Sunstein 2001; Sunstein 2004). Existing work on the effects of on-line influence is based mainly on either observational studies of a small number of groups (Ranerup 2000, Price and Cappella 2002, Price et al. 2003), or analysis of cross-sectional data taken from a large number of affiliated blogs (Adamic and Glance 2005). Such data are not sufficient to test predictions about changes over time in the distribution of beliefs, including the tendency for substantively unrelated issues to become aligned over time.

Coupled with developments in natural language processing now underway at Cornell, the Yesternet will make it possible for studies like those of Adamic and Glance and the Ackland team to include analysis of opinion dynamics, by mapping changes over time in the network structures in which these opinions are embedded. Researchers will be able to track opinion dynamics on the Web over the past decade, based on access to an extensive record of online discussions since 1996, including not only content but also network structure. Archived snapshots of message boards containing directed messages will make it possible to study the propagation of influence through dyadic interactions on a dynamic social topology. Research on the propagation of beliefs will have, for the first time, access to ongoing discussions and debates as these unfold in networks composed of identifiable cliques. We will be able to locate the structural positions of opinion leaders in networks of directed messages. By comparing the opinion dynamics of like-minded groups with those that attract members of rival camps, we will be able to study the spread of both positive and negative influence. Initially, research will be limited by the use of keywords which are better suited to track the salience of an issue than to sort out positions. Eventually, we hope to employ recent methods that apply machine learning to automatically identify the sentiment of pieces of text (Pang, Lee and Vaithyanathan 2002), and to automatically create belief-based summaries of one or more documents or text fragments (Cardie et al. 2004, Breck and Cardie 2003, Wiebe, Wilson and Cardie 2005). NLP tools developed at Cornell have already proven effective in reading movie reviews, and work is underway to extend these to political positions. The tools can be trained using on-line text of speeches by politicians or blog authors whose positions are known, and then extending their reach.

Diffusion across Organizations

Organizational Web pages typically have much more complete background data on the identity of the page owner and greater opportunity for supplementation through conventional archival sources. The Yesternet opens up new vistas for the study of organizational innovation. While organizational Web pages are generally a poor source of information on organizational policies and programs, the Web as a communications medium provides detailed histories of many kinds of organizational actions. Corporate downsizings are publicly communicated, as are patents, new product offerings, and many forms of market entry (Budros 1997). We plan to develop natural language and machine learning tools to locate and classify these sorts of communications (whether from public media, corporate press releases, annual reports, or public documents) in a systematic and reproducible fashion.

The Yesternet also permits an even bigger step on the explanatory side, and particularly in the detection of network diffusion. Much contemporary analysis is moving away from formal organizational affiliations and towards the study of individual migration. Research on top management teams shows powerful effects of prior organizational affiliations on product market entry and strategic profiles (Boeker 1997, Geletkanycz and Hambrick 1997). An organization's life chances depend on its parents, much as an individual's does (Phillips 2002). And in an analysis of a multinational bank's adoption of managerial innovations in 15 areas, the only consistent source of inter-organizational influence was the network generated by top manager mobility (Still and Strang 2004).

The Yesternet provides an opportunity to systematically generate networks based on inter-organizational migration, and to embed these in diffusion analyses. Using hand coding, members of our team have already developed procedures for tracing managerial career trajectories on the Web, via an inspection of public, company, and business media sources. These have been used to generate career histories for several thousand executives across more than 200 companies. As we move from labor-intensive human coding to semi-supervised natural language processing and machine learning techniques, we will be able to extend the scope of these studies to Web scale analysis.

Network Analysis and Modeling

Past research on diffusion has either ignored network structure or has treated the network as a fixture. A critical issue about which surprisingly little is known is the patterns by which large networks evolve over time. One premise of social science research facilitated by Web data is that the "Web" is a well-defined notion, independent of when and how it is measured. In fact, the Web exists in a constant state of flux. Not only is new content being added, but also entirely new forms of media and discourse are emerging (blogs, file sharing networks, and many others). Can we extract a common core structure for the Web that is stable across time, even as it grows, or can we make precise some of the ways in which the Web has undergone fundamental and qualitative changes in the past ten years? At a very basic level, we do not currently have a good understanding for whether the fundamental network properties of the Web in 1996 resemble those of the Web in 2005. For example, have the small-world properties, or the heavy-tailed nature of the degree distributions, remained roughly the same over all this time?

Some of our recent network studies have shown how even the most basic working assumptions about network structure are challenged by studying the evolution of these networks over long time scales. In recent work using the citation network of the e-print arXiv, we have found that the network is in fact densifying (the average number of edges per node is growing over time, rather than remaining constant as has been tacitly assumed in most models). We have also found that the diameters of many of these networks are in fact decreasing as the network grows (rather than growing slowly as a function of the number of nodes, as standard small-world models posit) (Leskovec, Kleinberg and Faloutsos 2005). We

are only beginning to explore possible reasons for these phenomena, in the form of models and explanations exogenous to the network itself. Moreover, these finding are currently based on the arXiv, which is much smaller and more homogeneous than the full Web. While the goal is to perform these measurements on the Yesternet, the arXiv is a very appealing and tractable model dataset on which to perform network studies before attempting them at Web scale. (This issue led us to use the arXiv as the basis for framing network evolution questions in the 2003 KDD Cup Competition (Gehrke, Ginsparg, and Kleinberg 2003). Following this competition, the arXiv has become a widely used testbed in the data mining community.)

Limitations of the Yesternet

Although the Yesternet will open up new avenues for research on innovation diffusion, there are also limitations and reasons for skepticism. Within the Internet Archive itself, some pages were never collected, some are lost, and others are blocked (including some discussion sites like LiveJournal). A more serious problem is that the Web is not representative of the offline world, as indicated by research on the digital divide (DiMaggio et al. 2001), a problem that is further complicated by the paucity of demographic data about page authors.

These problems are less serious in comparative studies of differences between large demographically homogenous groups or for studies using organizational pages that allow researchers to match Web data with background information obtained off-line. However, the anonymity characteristic of on-line interaction can be a serious limitation in studies at the individual level, compared to survey data with complete demographic profiles of respondents. For individual pages, limited demographic data can sometimes be identified (Lin and Halavais 2004), using page content or the geographic location of the page's IP address.

Nevertheless, substantial sampling problems clearly remain. Using Pew Internet Research data, it is possible to identify sampling bias and adjust accordingly. It is also important to run validity tests to see whether well-understood results can be replicated using conventional sampling methods. These tests are likely to reveal systematic demographic biases that will need to be taken into account in the analysis and interpretation of results. It will also be important to compare online results with case studies of face-to-face interaction to test robustness and to develop qualitative insights into the character of on-line interaction. These limitations remind us that social science using internet data should complement other research strategies rather than replace them.

V. References

Ackland, R., O'Neil M., Bimber B., Gibson, R. and S. Ward (2006), "New Methods for Studying Online Environmental Activist Networks," Paper presented to 26th International Sunbelt Social Network Conference, Vancouver.

Adamic, L. A. and N. Glance, "The Political Blogosphere and the 2004 U.S. Election: Divided They Blog." *WWW2005 Conference's 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*, 2005. http://www.hpl.hp.com/research/idl/papers/politicalblogs/AdamicGlanceBlogWWW.pdf

Aizen, J., D. Huttenlocher, J. Kleinberg and A. Novak, "Traffic-Based Feedback on the Web." *Proceedings of the National Academy of Sciences*, 101(Suppl. 1): 5254-5260, 2004.

Boeker, W., "Executive Migration and Strategic Change: The Effect of Top Manager Movement on Product-Market Entry." *Administrative Science Quarterly*, 42: 213-236, 1997.

Breck, E. and C. Cardie. "Playing the Telephone Game: Determining the Hierarchical Structure of Perspective and Speech Expressions." 20th International Conference on Computational Linguistics (COLING-04), 2004.

Budros, A., "The New Capitalism and Organizational Rationality: The Adoption of

Downsizing Programs 1979-1994." *Social Forces*, 76: 229-250, 1997.
Cardie, C., J. Wiebe, T. Wilson and D. Litman, "Low-Level Annotations and Summary Representations of Opinions for Multi-perspective Question Answering." New Directions in Question Answering, M. Maybury. Boston, AAAI Press/MIT Press: 87-98, 2004.
Carpini, M. X. D., F. L. Cook and L. R. Jacobs, "Public Deliberation, Discursive Participation, and Citizen Engagement: A Review of the Empirical Literature." *Annual Review of Political Science,* 7: 315-344, 2004.
Coleman, J. S., E. Katz and H. Menzel, *Medical Innovation: A Diffusion Study*. New York, Bobbs-Merril, 1966.
Davis, G. F., "Agents without Principles? The Spread of the Poison Pill through the Intercorporate Network." *Administrative Science Quarterly*, 36: 583-613, 1991.
DiMaggio, P., E. Hargittai, W. R. Neuman and J. Robinson, "Social Implications of the Internet." *Annual Review of Sociology*, 27: 307-336, 2001.
DiMaggio, P., J. Evans and B. Bryson, "Have Americans' Social Attitudes Become More Polarized?" *The American Journal of Sociology*, 102: 690-755, 1996.
Edelman, L. B., "Legal Environments and Organizational Governance: The Expansion of Due Process in the American Workplace." *American Journal of Sociology*, 95: 1401-1440, 1990.
Fiore, A., S. L. Teirnan and M. Smith, "Observed Behavior and Perceived Value of Authors in Usenet Newsgroups: Bridging the Gap." *CHI 2002 Conference on Human Factors in Computing Systems*, 2002. http://research.microsoft.com/~masmith/CHI 2002 - Objective Author Behavior and Perceived Value - Final.doc
Gehrke, J., P. Ginsparg and J. Kleinberg, "Overview of the 2003 KDD Cup." *ACM SIGKDD Explorations Newsletter*, 5: 2 149-151, 2003.
Geletkanycz, M. and D. Hambrick, "The External Ties of Top Executives: Implications for Strategic Choice and Performance." *Administrative Science Quarterly*, 42: 654-681, 1997.
Halavais, A., "Networks and Flows of Content on the World Wide Web." *2003 International Communication Association Conference*, 2003. http://alex.halavais.net/research/halavais-ica03a.pdf
Halavais, A., "The Slashdot Effect: Analysis of a Large-Scale Public Conversation on the World Wide Web." Unpublished Doctoral Dissertation, University of Washington, 2001. http://alex.halavais.net/research/diss.pdf
Herring, S., K. Job-Sluder, R. Scheckler and S. Barab, "Searching for Safety Online: Managing "Trolling" in a Feminist Forum." *The Information Society*, 18: 371-384, 2002.
Klam, M., "Fear and Laptops on the Campaign Trail." *New York Times Magazine*, September 26, 2004: 42, 2004.
Kleinberg, J., "Bursty and Hierarchical Structure in Streams." *Proc. 8th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, 91-101, 2002.
Kollock, P. and M. Smith, "Managing the Virtual Commons: Cooperation and Conflict in Computer Communities." *Computer-Mediated Communication: Linguistic, Social, and Cross-Cultural Perspectives*, S. Herring. Amsterdam, John Benjamins: 109-128, 1996.
Leskovec, J., J. Kleinberg and C. Faloutsos, "Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations." To appear in *Proceedings of the 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2005.
Lin, J. and A. Halavais, "Mapping the Blogosphere in America." *WWW 2004 Conference's Workshop on the Weblogging Ecosystem*, 2004. http://www.blogpulse.com/papers/www2004linhalavais.pdf
Morris, A., "Black Southern Student Sit-in Movement: An Analysis of Internal Organization." *American Sociological Review*, 46: 744-767, 1981.
Pang, B., L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques." *Proc. 2002 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), 79-86, 2002.
Phillips, D. J., "A Genealogical Approach to Organizational Life Chances: The Parent-Progeny Transfer among Silicon Valley Law Firms 1946-1996." *Administrative Science Quarterly*, 47: 474-506, 2002.
Poor, N., "Mechanisms of an Online Public Sphere: The Website Slashdot." *Journal of Computer-Mediated Communication*, 10: Article 4, 2005.
Postmes, T., R. Spears and M. Lea, "The Formation of Group Norms in Computer-Mediated Communication." *Human Communication Research*, 26: 341-371, 2000.
Preece, J. and D. Maloney-Krichmar, "Online Communities." *Handbook of Human-Computer Interaction*, J. Jacko and A. Sears. Mahwah, NJ, Lawrence Erlbaum Associates, Inc: 596-620., 2003
Preece, J., "Etiquette Online: From Nice to Necessary." *Communications of the ACM*, 47: 56-61, 2004.
Price, V. and J. N. Cappella, "Online Deliberation and Its Influence: The Electronic Dialogue Project in Campaign 2000." *IT & Society*, 1: 303-329, 2002.
Price, V., D. Goldthwaite, J. N. Cappella and A. Romantan, "Online Discussion, Civic Engagement, and Social Trust." *Working Paper, University of Pennsylvania*, 2003.http://cct.georgetown.edu/apsa/papers/Price.pdf
Ranerup, A., "On-Line Forums as an Arena for Political Discussions." *Digital Cities, Technologies, Experience, and Future Perspectives*, T. Ishida and K. Isbister. Heidelberg, Germany, Springer-Verlag: 209-223, 2000.
Rangaswamy, A. and S. Gupta, "Innovation Adoption and Diffusion in the Digital Environment: Some Research Opportunities." *eBusiness Research Center Working Paper*, 1999. http://e-commerce.mit.edu/papers/ERF/ERF46.pdf
Reid, E., "Hierarchy and Power: Social Control in Cyberspace." *Communities in Cyberspace*, M. Smith and P. Kollock. London, Routledge: 107-133, 1999.
Rogers, E. M., *Diffusion of Innovations*. New York, Free Press, 2003.
Ryan, B. and N. Gross, "The Diffusion of Hybrid Seed Corn in Two Iowa Communities." *Rural Sociology*, 8: 15-24, 1943.
Smith, M., "Invisible Crowds in Cyberspace: Measuring and Mapping the Social Structure of Usenet." *Communities in Cyberspace*, M. Smith and P. Kollock. London, Routledge: 195-219, 1999.
Still, M. C. and D. Strang, "Who Does an Elite Organization Emulate? Networks, Influence, and Benchmarking." *Presented at the 2004 Annual Meeting of the American Sociological Association*, 2004.
Sunstein, C. R., "Democracy and Filtering." *Communications of the ACM*, 47: 57-59, 2004.
Sunstein, C. R., *Republic.Com*, Princeton, NJ, Princeton University Press, 2001.
Viegas, F. and M. Smith, "Newsgroup Crowds and Authorlines: Visualizing the Activity of Individuals in Conversational Cyberspaces." *37th Hawaii Int'l Conference on System Sciences*, 2004.
Watts, D. J., *Six Degrees: The Science of a Connected Age*. New York, W. W. Norton & Company, 2003.
Wellman, B., J. Salaff, D. Dimitrova, L. Garton, M. Gulia and C. Haythornthwaite, "Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community." *Annual Review of Sociology*, 22: 213-238, 1996.
Wiebe, J., T. Wilson and C. Cardie, "Annotating Expressions of Opinions and Emotions in Language." To appear in *Language Resources and Evaluation* (formerly *Computers and the Humanities*) *1*, 2005.
Wilson, S. M. and L. C. Peterson, "The Anthropology of Online Communities." *Annual Review of Anthropology*, 31: 449-467, 2002.