

The Structural Determinants of Media Contagion

by Cameron A. Marlow

Thesis Proposal for the degree of Doctor of Philosophy
at the
Massachusetts Institute of Technology

June 2004

Thesis Advisor

Walter Bender
Senior Research Scientist
Media Laboratory
Massachusetts Institute of Technology

Thesis Reader

Keith Hampton
Assistant Professor
Dept. of Urban Studies and Planning
Massachusetts Institute of Technology

Thesis Reader

Thomas Valente
Associate Professor
Dept. of Preventive Medicine
University of Southern California

Abstract

Informal exchanges between friends, family and acquaintances play a crucial role in the dissemination of news and opinion. These casual interactions are embedded in a network of communication that spans our society, allowing information to spread from any one person to another via some set of intermediary ties. These events are part of the ongoing *media contagion* that affects us on a daily basis, allowing news, stories and opinions to be directed to interested parties.

Communication technology has the general effect of accelerating the speed of our interactions, allowing people to maintain more contacts at a lower cost. The internet in particular has provided a number of communication media through which individuals can easily and quickly stay in touch. As these interaction media change the speed and structure of our social networks, they also have the possibility to affect the process of media contagion by speeding, slowing, or changing the types of information that are infectious.

Weblogs are a recent communication technology embedded in the web that engenders the process of media contagion, wherein individuals publicly write about their lives on a recurring basis. Because weblog authors are tied by social networks of readership, contagious media events happen frequently, and in a form that is observable to researchers interested in the diffusion of information.

This proposal describes a thesis project aimed at bettering our understanding of the process of media contagion through the study of weblogs, their authors and the information that diffuses through their writing. It seeks to explicate the structural properties that determine the flow of information through weblog networks and characterize the effect of having a weblog on the communication patterns of the authors.

Introduction

Informal communications between friends, colleagues and strangers represent a fundamental mode for disseminating news and ideas to the general population. A water cooler conversation or weekly call to family is embedded in a much larger system of diffusion wherein one message can pass from an individual to a directed audience of interested parties. In this manner families can stay in touch, political movements can mobilize and major news events can reach the widest audience possible. This is the process of *media contagion*, whereby pieces of information diffuse through the social networks of individuals in our society.

Communication technology has the general effect of increasing the speed of interaction, allowing an individual to stay in touch with more people, and in turn increasing the frequency of interaction. From improvements in travel to the telephone to the internet our society exists on a trajectory of rising connectedness allowing family and friends to stay in touch and businesses to expand their markets. The World Wide Web represents a significant point on this path where informal messages take on a persistent form, recorded for longevity in an archive accessible from any point in the world.

Because a substantial amount of daily interaction now happens in a publicly-accessible persistent medium, we now have the ability to track and study the diffusion of information among individuals in an efficient manner. Previous attempts to study the diffusion of ideas required intense effort on the part of large research teams to first map the social structure of a given community and then track a single message as it spread among individuals. Without intervention, a computer can now track this process as it unfolds on the web, not for one message, but for all messages that might be diffusing.

Weblogs are a relatively new communication medium built atop the web that encapsulates informal interaction. Individual authors record personal journals of thoughts, stories, news and ideas of interest to a publicly-accessible web site. Each weblog author has a set of friends and colleagues who also maintain weblogs; the process of communication for an individual author is operationalized as (1) reading the daily set of weblogs and (2) posting those stories and ideas that are of interest on their own weblog. In this way weblogs encapsulate the process of informal diffusion that happens everyday both online and offline. Because they are public, they are fertile ground for the study of media contagion.

This thesis proposal represents the culmination of work done on the Blogdex project, now in its third year tracking diffusion among individuals within the weblog community. It is an attempt

to use the ongoing communication of weblog authors to better understand the basic informal communication that occurs as part of our daily lives. At the same time it maintains the goal of describing the difference between this very explicit medium of interaction and the other modes through which we communicate.

Specific Aims

The collection of work on the diffusion of information has highlighted the importance of structural features in the process of dissemination, as the social network of associates is the conduit for information propagation. Certain individuals are ascribed a position of power to the extent that they are able to control the passage of a given idea to the rest of the population. The following five aims outline a method to characterize both the social networks of bloggers and the information that passes between them. Furthermore they are meant to differentiate weblogs from other forms of communication, explicate the use of weblogs among the general population and ascertain the trajectory of the medium in demographic and social terms.

Aim 1: *Describe the structure of the weblog social system and model its static and dynamic qualities.* Using a corpus of weblogs collected over a three-month period, the social network of authors will be extracted from the hypertext links made between these sites. The structure will be analyzed using standard social-network measurements including node degree, clustering coefficient, centrality, betweenness, and network growth.

We hypothesize that the observed social network will be unlike previously analyzed social structures. The level of clustering and cliquishness will be significantly lower than those found in offline social networks suggesting a culture based on individuals instead of groups. While many weblog ties mimic those found offline (i.e. friends both online and offline), those ties formed exclusively online will emerge along lines of interest, not social or geographic propinquity. The resulting social structure will be more fluid and diffuse than offline interactions can produce.

Aim 2: *Create a descriptive model of the diffusion of media across the social network of bloggers.* In addition to the social-network data collected in Aim 1, all references from weblogs to outside web sites will also be maintained along with referent and time of discovery. These unique contagion events will be comparatively analyzed and categorized along dimensions of speed of diffusion, structural properties of the referent community and the path the information takes to spread among these individuals.

We hypothesize that the diffusion of a given piece of media will be influenced by three factors: contagiousness, as measured by the speed of diffusion; contextuality, as defined by the breadth of the adopting community; and perception of audience, the increasing scale through which the idea spreads. These three related dimensions all interact to create the incident community created by the media event.

Aim 3: *Describe the demographics and population characteristics of the current weblog community.* We will conduct an online social survey of weblog authors based on the collection of weblogs discovered in Aim 1. English weblogs will be randomly selected from this set (n = 3,000 respondents) and their authors emailed to take part in an online survey. The first section of the survey will focus on demographics of the authors (gender, race/ethnicity, age, education, marital status, and location) and their current involvement in the medium; these data will be compared to the most recent census figures to ascertain the representativeness of the sample and standard census measures will be employed to describe current growth trends of the medium.

We predict that the respondents will reflect a technologically-savvy population being biased towards younger, Caucasian males with at least a high-school education. We also predict that while the growth of the total number of weblogs has an exponential form, the high levels of migration, turnover and overlap will predict a much slower growth rate. The disconnect between the population of weblogs (web pages) and the population of authors (people), accompanied by the persistence of unused weblogs will add a high level of inaccuracy to any measurements based solely on total sites.

Aim 4: *Compare weblog ties and non-weblog ties of authors and describe the effect of weblogs on the networks of the author.* The second section of the survey described in Aim 3 will cover the online and offline social ties maintained by the surveyed weblog authors. Respondents will be questioned about their online and offline social ties and relation to social capital.

We hypothesize that there will be a marked difference between the social networks created online and those formed offline. Because ties among bloggers are created along lines of personal interest, they are not as heavily influenced by the physical and cultural space that typically constrains the possible social interactions offline. The diffuse ties formed online will tend to bridge authors to more diverse resources increasing the social capital embedded therein. Offline ties, on the other hand, will tend to be geographically constrained and

Background and Significance

Social Networks

The concept of media contagion is inextricably intertwined with the fact that every individual is part of a connected network of social relationships that spans the world. This analysis was first made explicit by Stanley Milgram's pioneering observation of the "small world" phenomenon, wherein any two individuals in our society could be connected by a small number of acquaintances (Milgram 1967). The structural nature of social phenomena has grown into both a theory (Wellman 1997) and a method for analyzing social interactions (Wasserman and Faust 1994). Social Network Analysis (SNA) considers society as a set of individuals (actors) and the relationships between them (ties), drawing conclusions from the properties of the networks that a particular individual or group of individuals maintains.

A number of observations relevant to media contagion have been made in the field characterizing the individuals and specific social ties that promote information diffusion. Granovetter observed that information is more likely to spread through the weak ties of an individual, namely those acquaintances with whom a person seldom communicates. These individuals are typically connected to entirely separate networks of contacts, and thus serve as a bridge to information and resources; as these contacts harbor a much greater wealth of information than our more regular contacts; this theory has been labeled "the strength of weak ties" (Granovetter 1973, 1983).

From the observation that not all ties are created equal and that some individuals have access to more resources than others comes the notion of *social capital*, a measurement of the value of a person's social ties (Coleman 1988). Since the inception of the concept, there has been considerable debate over what exactly constitutes social capital, and how it should be measured (Fischer 2001; Portes 1996; Putnam 2000; Woolcock 1998).

Social networks are typically measured in two manners: first, by taking a set of individuals and attempting to map the social relationships between them, known as *whole networks*; second, by randomly sampling a population and asking subjects to enumerate their personal networks, referred to as *ego networks* (Wasserman and Faust 1994). Because whole network data is difficult to collect, especially for large populations, the ego network approach has come to be the standard unit of measurement. Network surveys extract personal networks from respondents in a through a number of different survey techniques known as *network modules*:

- *Name generator*: This technique asks individuals to list contacts by name given some defining feature or role (e.g. people you obtain social support from, people you talked to last

Thursday, etc.) and the social ties that exist between these contacts (Fischer and McCallister 1978; Marsden 1987; van der Poel 1993)

- *Reverse Small World*: Subjects are given a set of randomly chosen names derived from some master list (e.g. phone books, company roster, etc.) and asked to denote the names with which they have some social tie (Killworth 1990).
- *Position Generator*: A list of common occupations of varying socioeconomic status (SES) are presented and subjects are asked to denote those positions for which they have a social tie (Lin 1998)

Each of these modules represents a proxy to the subject's connectedness to others in society. The name generator produces an explicit ego network, typically considered to be the respondent's strong ties. The latter two methods gauge the extent to which the subject is connected generally in society, and produce much better approximations of the weak ties they may have. For this reason, the position generator has become an accepted tool for measuring social capital (Lin 1998).

SNA has also become a popular method for describing the effect of technology on interpersonal communication. The analogy has been made that computer networks are social networks insofar as they support existing social ties and allow for the formation of new ties (Garton, Haythornthwaite, and Wellman 1997; Wellman 2001) while embedding us further within our local community (Hampton and Wellman 2003). Online interaction supports the development of weaker, more specialized ties evolving around specific areas of interest (Wellman and Gulia 1997) potentially connecting individuals to a much larger wealth of information than they would have available offline. These findings exist within a much larger debate over the effect of the internet on the psychology of the individual: one camp heralds the net as a place where social relations can evolve regardless of race, gender, creed or geography (Patton 1986) while another purports a dystopian end where individuals become increasingly disenfranchised from their friends and family. (Nie 2001; Kraut et al. 1998; Nie and Erbring 2000).

The study of weblog authors' online and offline social relations will add to this dialog by addressing a rapidly expanding online community. Because of the public nature of weblog content, we have the uncommon ability to present a representative sample of users within the context of a specific communication tool.

Diffusion of Information

One of the earliest realizations of the structural effects of diffusion comes from the work of Katz and Lazarsfeld on their two-step flow of communication. While studying the effect of media on

opinion they realized that television campaigning had little effect on its own, but when coupled to an interaction with highly-influential opinion leaders, the impact was substantial (Katz and Lazarsfeld 1955). Through a number of studies they concluded that people's opinions are mediated through their interaction with others, and that influence is determined by a small set of opinion leaders in the community.

Studies of news events during the latter half of the 20th century showed analogously that interpersonal communication is a regular part of the news-diffusion process, and that diffusion on the first day of a news event tends to be dominated by television and interpersonal communication (Deutschmann and Danielson 1960). Interpersonal communication tends to be the most important mode of dissemination in events that tend to spread to the entire population and those that spread to very small audiences while everything else tended not to be a topic of conversation (Greenberg 1964). These findings suggest that media contagion plays an important role in bringing news to a specific audience, a theme that has been reinforced by other studies in the area (Bogart 1951; Funkhouser and McCombs 1971).

Rogers has conducted a number of studies on the diffusion of innovations and analyzed the spread with respect to various structural properties (Rogers 2003). The results have shown that individuals fall into predictable *adopter categories* that have subsequently become part of the common vernacular: innovator, early adopter, late adopter and laggard. Every person is assumed to have a specific threshold at which they will decide to adopt a particular innovation; after the percentage of the population that has adopted passes this threshold, the individual will decide to adopt. Over the course of a specific diffusion event, overall adoption rates typically follow a characteristic *S-curve*, where growth begins with early adopters, gains momentum until a critical mass is achieved, and the diffusion begins to saturate the population (Rogers 2003; Valente 1995).

Analyses of these data with respect to structural properties have shown that adoption thresholds are localized, and that the percentage of adopters in an individual's personal network determines their incentive to adopt an innovation (Valente 1995). Ronald Burt has argued that thresholds are not the most important factor, but rather that adoption is dependent on imitation. When one of two *structurally equivalent* individuals (two people who share a large number of contacts) decides to adopt, the other will follow (Burt 1987). Weimann has extended the strength of weak ties to suggest that within a given social structure, those marginal individuals weakly tied to all of their contacts are commonly a source of new ideas and information (Weimann 1982). Social network researchers have validated theories of weak ties

and opinion leadership, showing that innovations, rumors, and beliefs tend to move from those marginal in a network, to the central figures, and back to the rest of the population (Valente 1995; Rogers 2003; Weimann 1982).

The study of the diffusion of information through weblogs sits on a different scale than has been used in previous diffusion studies. Whereas former work required extensive man-hours to determine the diffusion of a single idea through a small community (Valente 1995), contagion data from weblogs can be collected for millions of individuals and thousands of new examples every day.

Epidemiology

Like the study of information diffusion, epidemiology concerns itself with the empirical analysis and modeling of infectious agents. While epidemiology was an early influence on the study of social networks (Rapoport 1953), the two fields have greatly diverged in their study of similar phenomena. Because most infectious diseases are airborne and not heavily influenced by structural features of social interaction, epidemiological methods typically look at rates of infection from a population level instead of the local, structural perspective.

A few standard models are used to model diseases of different types: *susceptible-infected* (SI), for diseases which have no cure, e.g., cancer; *susceptible-infected-recovered* (SIR) for diseases where recovery incurs immunity, e.g., influenza; and *susceptible-infected-susceptible*, for infections which can be cured but garner no immunity, e.g., syphilis (Baily 1975). These methods allow epidemiologists to accurately determine various characteristics of the spread of a disease for large populations. For the SIR model of infection, given populations X of susceptibles, Y of infected and Z of removed, the following differential equations define the spread of the disease over time (Anderson and May 1992):

$$\begin{aligned}dX/dt &= -\lambda X(Y/(X+Y)) \\dY/dt &= \lambda X(Y/(X+Y)) - \delta Y \\dZ/dt &= \delta Y \\ \lambda &= c\beta\end{aligned}$$

where δ is the rate of recovery/removal and λ is the the rate of infection, defined by c , the average number of contacts an individual in the system has and β , the transmissibility of the disease. As with the study of diffusion of information, these equations conform to an S-curve given that the rate of infection is greater than 1.

The use of c has become a contested feature in the analysis of diseases that are heavily dictated by structural features, such as sexually transmitted infections. In these cases, the number of

contacts a person has varies widely, and in some cases most of the contacts are held by a small, core group of individuals. In these situations, the variance of c (i.e. the structure) can greatly dictate all of the other parameters, and is usually replaced with a distribution (Anderson and May 1992).

Self-organizing networks

A new theory of self-organizing networks has been gaining momentum in the past few years based on empirical observations in a number of different disciplines. Drawing from networks in a variety of empirical domains, these researchers have devoted their attention to modeling the static and dynamic features of large, organic networks. Their results are becoming accepted as a general theory of networks outside of the specific domain and context within which networks are usually considered.

The first discovery in this new discipline was a model for generating networks with properties similar to those observed by Milgram. Two constraints determined the model, namely that nodes in the network should be highly clustered at a local level (i.e., most nodes are densely connected to a small number of other nodes) while the entire system should have a relatively low characteristic path length (the average distance between any two nodes). Watts made the observation that by taking dense networks and rewiring only a few connections, a network can be generated that satisfied both conditions; his model has been termed the *small worlds network* in homage to Milgram (Watts 1998).

Other researchers have focused specifically at the distribution of node degree, discovering that many real-world networks do not follow the Poisson distribution predicted by a random graph; instead, many self-organized networks follow a form with a disproportionately large number of nodes having very few connections while a very small group is extremely connected. These distributions are power laws as they follow to the form $P(k) \sim k^{-\alpha}$ where α is the slope of the line when the distribution is plotted in log-log form. The observation of these networks has led to a vast array of papers on the topic, popularized recently by Albert-László Barabási. Barabási has posited that power law distributions in self-organizing networks often arise from a process of preferential attachment, where nodes with higher degree are more likely to receive new links than less connected ones (Barabási 2002).

Among the literature on power laws, a significant amount of attention has been dedicated to estimating parameter values of the exponent (α). For epidemiologists, this feature is at the crux of this technical debate because α determines the variance of distribution of node degree (or contact rate). The contact rate variance affects the value of the reproductive ratio (R_0) (i.e, the

number of secondary infections transmitted in an entirely susceptible population, when one subject is infected). When the contact variance is infinite, R_0 exceeds the epidemic threshold level and disease remains endemic and cannot be arrested (Pastor-Satorras and Vespignani 2001). Conversely a bounded/finite contact variance keeps R_0 below the epidemic threshold allowing for infectious diseases to die out of the population. Hence the values of α may have implications for the spread of infectious agents for a given population (Zoltan and Barabási 2002)

Weblogs

Despite the relative infancy of weblogs compared to other online media, their public nature has led to a number of empirical observations, both by weblog authors and academics alike (but mostly by weblog authors). Central to the topic of this thesis, three areas have become the focus of attention: the distribution of social ties throughout the community, measurement of authority among webloggers, and modeling the diffusion of information among them.

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

- *Blogrolls*, or the list of other weblogs that a given author reads regularly. These lists are maintained by webloggers as a tool to direct readers to other similar content, and as a general proxy to the social affiliations of the author.
- *Permalinks*, or references to specific pieces of content on another weblog. These ties are formed in the process of content creation, namely by referring to something that someone else in the community has said. These links can be thought of as a more implicit form of social structure.

A recent debate that has raised quite a bit of attention among weblog authors is related to the distribution of links within the community. Clay Shirky wrote a piece documenting the fact that a power law existed in the blogroll links between authors (Shirky 2003). Shirky assumed this model to claim that within the weblog ecosystem, the "rich get richer," and that the longer one has been an author, the more central they will be. Further analysis showed that permalinks, while still following a power law, produced an entirely different measure of popularity or authority. This led to the observation that popularity and influence are not necessarily correlated within this social system (Marlow 2004). Furthermore, the distribution of ties does

not follow the expected small world pattern suggested by Watts, rather the network of affiliations is a dense mesh of relations with very little clustering at any point (Marlow 2003).

The diffusion of information has also been the attention of researchers due to the fact that contagious media events are extremely common and observable within the weblog community. Early studies of this process have shown that unlike innovations, news and stories do not have distinct adopter categories across all categories, but rather that time of adoption is topically defined (Marlow 2002). Looking from the perspective of the media, three distinct categories of diffusion have been found by looking at the rate of spread: factual news, shown by a rapid growth and decay; opinion, shown by a slightly slower growth; and services, shown by a constant rate of diffusion (Adar 2004; Marlow 2003).

Research Design and Methodology

After realizing the importance of media contagion among weblogs, I constructed a system to track and analyze media events and relay this information back to the community. The impetus was to create a system that gave webloggers a global perspective on what others were talking about, such that popular information did not become stuck in a local part of the social structure. The system, named Blogdex, went live in August 2001, received a good amount of initial feedback, and has since become an integral part of the community.

Our goal is to collect data which will expand our understanding of the diffusion of ideas through the community of weblog authors and the extent to which this medium affects the social connectedness of its members. To accomplish this task we will employ two separate data sources: an automatically-discovered set of weblogs and link-diffusion data within this group over a specified period of time; and a social survey of authors randomly sampled from the first data set. This section outlines the research design for each data set along with the methodology and expected academic contributions for each of the specific aims of this proposal.

Diffusion Aggregator

Up until recently, Blogdex was an opt-in service for webloggers interested in participating. But to create the most complete data on diffusion we will attempt to automatically discover the largest possible set using automatic aggregation techniques. As weblogs are publicly accessible websites, they are freely available for download at any time. Unlike other survey techniques, the only barrier to obtaining a complete sample of the weblog community is our ability to identify them among other websites. We will employ standard crawling techniques (cite) to locate potential weblogs: starting with lists of recently updated weblogs (available from blo.gs or

weblogs.com), we will retrieve and store these websites (which can almost be assumed to be weblogs) and the sites linked to and from them.

This potential set of weblogs will be run through a heuristic based on common weblog features (words, html, javascript, links to identified weblogs, etc.); sites with a high-enough score will be identified as weblogs while others will be stored and checked again as the heuristic or features change. For those sites identified as weblogs, another crawler will be employed to keep the most-recent-possible local copy using weblog-update engines and a local-learning system to predict the update frequency. Furthermore, a trigram-based, stochastic language-identification system (cite) will be performed against the text of each weblog and the resulting language stored with the weblog in the database along with the confidence returned by the classification process.

Whenever a changed weblog is found, the site will be parsed and hypertext links extracted. These links—markers of media contagion—will be stored in a relational database along with the source weblog, time of discovery and surrounding text as context. This database will be the primary data for analysis in Specific Aim 2. Links made to other weblogs (either in the form of direct reference or permalinks to specific posts) will be stored in another database as the referent social network to be used in Specific Aim 1.

As all data collected in this system is observation of public behavior, identities of individual weblogs will be maintained pending exemption approval by MIT's IRB, COUHES. These data will be maintained on a secure server and backups made on a nightly basis to guarantee data integrity.

Survey

The second primary data set will come in the form of a general social survey of weblog authors. Subjects will be acquired by randomly selecting sites identified as weblogs by the diffusion aggregator and sent an email from correspondence information located on their site. Sites without correspondence information will be discarded from the sample and the percentage of contactable subjects noted.

Recent studies estimate that the total number of active, hosted weblogs is between 2 and 4 million (Perseus 2004). The authors of this survey admit that a majority of weblogs are abandoned shortly after creation and that many authors maintain more than one weblog. Assuming a one-to-one relationship between weblogs and authors for the lower end of their estimate, a sample size of $N=2,398$ will be needed to attain a 95% confidence level on results (at a confidence interval of 2). Given that worst-case response rates for online surveys are around

50% without incentive, we will oversample to maintain our intended sample size, placing the total number of contacted authors at $N=5,000$.

Due to the time constraints of the study, we will focus only on weblogs written in the English language. When the weblog acquisition stage of the diffusion aggregator has reached a relative point of stability, the output of the language detector will be used to select those written in English, and these data will be verified when email addresses are culled from the selected weblogs.

Non-incentivized web surveys must be constrained in length in order to achieve good response rates. As previous studies have shown, maximum response rates occur when the survey takes less than 15 minutes to complete (MacElroy and Gray 2003). Our survey will have two sections excluding the demographic questions that will occur before the survey officially begins.

The first section will address weblog history and usage, asking each respondent to enumerate each weblog that they have maintained. For every site respondents have authored, the url, creation date, last posted date, general activity, current availability and number of other authors will be requested.

The second section will probe the online and offline social networks of the subjects, using two survey instruments: first, a name generator (McCallister and Fischer 1978) will be used to explicate the strong ties of subjects, where respondents will be asked to list all formal and informal communications that occurred on a specific weekday in the past month, and the ties that exist between these contacts; second, a position generator (Lin 1999) will be employed to measure weak ties, where respondents will be given a set of common professions of varying socioeconomic status (SES) and asked whether they know a person in each position. In both cases, questions will be asked about the contacts listed, namely demographics, primary form of communication, relationship to the subject, and whether or not they were met online or through weblogs. For an example of the questions to be asked of respondents, please see the survey sample in Appendix I.

The identities of survey respondents will be strictly masked in order to maintain the utmost security and privacy of subjects. Pin numbers and passwords will be included in the initial email, which will allow subjects to take part without having to supply identifying information, and all records relating pins to the originating weblogs will be destroyed before the survey begins. All information that is necessary for identification purposes (weblog URLs, first names of social network members, etc.) will be immediately and irreversibly translated into a deidentified form using the MD5 message-digest algorithm (cite). Data will be stored on a computer running the

survey website and offloaded to a standalone machine for analysis while data will be checked daily against checksums to prevent data loss from tampering of any kind. The survey is currently seeking COUHES approval for ethical considerations.

Specific Aims

The following four sections will address the research design, evaluation methodologies and expected contributions for each of the specific aims of this proposal. The first two aims constitute the core analyses of this project while the latter two provide context for interpreting our findings.

Aim 1

Describe the structure of the weblog social system and model its static and dynamic qualities.

A number of theories have been hypothesized as to the effect of internet communication technologies on the social structure of participating individuals casting this technology in both a utopian and dystopian light. Whether we are heading towards a state of empowered, networked individualism or becoming increasingly disenfranchised from those around us remains to be proven. As growing numbers of individuals begin to use the internet as their primary means of communication and place for finding new associates, it becomes increasingly important to understand online social organization and its difference from interaction occurring offline.

Our data stands in an important position to either confirm or deny many of the claims of previous work in computer-mediated social structure. While more and more interaction is occurring through internet channels, very little personal, informal communication is available to researchers. Newsgroups, message boards and chat rooms are all publicly available resources for the study of online communities, but as group environments they engender an entirely different dynamic than the informal interaction than those between individuals. At the same time, rich, explicitly-defined social networks (such as those generated by Friendster or instant-messaging systems) remain property of companies that maintain infrastructure.

Weblogs constitute one of the largest online communities whose content is publicly available. Within this social environment both implicit and explicit social networks exist as permalinks and blogrolls respectively. These data will be collected as a byproduct of the aggregation of diffusion data, and present an opportunity to analyze and better understand the static and dynamic properties of social networks within an online community.

In addition to validating theories of online social structure, the data collected by the diffusion aggregator will be beyond the bounds of standard social-network analysis. As most existing

studies have been developed for analyzing small whole networks or survey samples of ego networks, the complexity of current algorithms is beyond the computational limits of current computers by two orders of magnitude. Due to space constraints, all algorithms must use an adjacency-list representation of the network while computational constraints will limit us to algorithms which perform a constant number of operations per node and edge in the graph.

To characterize the static social structure we will employ a number of techniques used in the recent study of emergent networks; a number of measures will be calculated for each individual in the network. The *in-degree* and *out-degree* of each node will be used to describe the distribution of social ties within the community. We expect this distribution to follow a roughly scale-free form where most individuals will maintain a relatively small number of ties to other webloggers while a few will be connected to very large number of associates. The *clustering coefficient* will be used to show how densely ties are distributed for each individual while the *characteristic path length* will be calculated to show centrality within the whole network. Based on preliminary analyses we expect the entire system to form a mesh-like structure with clustering coefficients being much lower than those observed in offline networks but still maintaining a low characteristic path length for each individual. These results would validate the belief that online ties tend to have a form resembling networked individualism where clusters and local groupings are uncommon.

Because social ties are marked by the time that they are observed, we are able to also characterize the dynamics of this system. Starting at the event horizon of the diffusion aggregator, we will take the growth of the system over a three-month period to determine some of the growth patterns of each individual's personal network. We will record each of the static network parameters at one-week intervals throughout the period of data collection and analyze the growth (rate and acceleration), density, and relative scale of social ties for individuals. We expect that new webloggers will experience a period of rapid growth initially in terms of outward links stabilizing into slow exponential growth and a similar growth pattern for inward links from the beginning. These data would suggest a preferential attachment model for the weblog system whereby individuals grow at a rate relative to their age with exceptions for those that bring offline social ties into the system.

Aim 2

Create a descriptive model of the diffusion of media across the social network of webloggers. Previous studies of the diffusion of information have suffered from a primary limitation of data collection. In order to observe the diffusion process, researchers would manually survey the

study population to obtain a network of informal ties and then subsequently question them about their knowledge of a specific idea. Changes to the social structure were not observed and most studies were limited to only one or a few examples of diffusion.

The data collected by the diffusion aggregator will be wholly different from former studies thanks to the relative scale of data collected. The population studied is many orders of magnitude larger than the next largest diffusion study while the number of examples ranges in the millions instead of just a few. We will be in a prime position to validate former theories of information and innovation diffusion that have traditionally been constrained by the limited size and scope of their data. Furthermore we hope to create a general vocabulary for describing media contagion and the categories of behavior from the perspective of individual media events.

The first stage of our analysis of these data will be to characterize media events in terms of epidemiological measures. A standard SI model of infection will be used to infer the time of critical mass, infectiousness and total population size for each event. Distributions of these variables will be observed in order to characterize the landscape of events and give context to the interpretation of individual events. The distributions will also be clustered to look for aggregate groupings, such as the various subtypes observed in (Adar 2004; Marlow 2003). The expected outcome, as observed in previous studies of aggregate event data (Marlow 2002, 2003) is that the adoption sizes for media events will follow a power law regardless of the sampling timeframe. Furthermore, we expect to validate previous findings in the different subtypes of media events, namely the classification of news, commentary and services, each with a characteristic time to critical mass, regardless of overall adoption size.

Our second analysis will be on the effect of structural properties of media contagion. Using a number of heuristics based on existing social ties and previous diffusion characteristics of each author, we will infer the path of diffusion for each media event. As with the spread of any infection, the structural properties of individuals have a profound effect on the rate of diffusion and overall population size. Using social-network measures obtained in Aim 1 (namely degree and betweenness centrality), we expect that there will be a strong correlation between the centrality of adopting weblogs and the overall population reached by the event. We also expect that the order of adoption will be a determining factor in the size of the event, namely that degree and betweenness will be mitigating factors in the overall spread of ideas.

The third examination will take the individual micro-communities of each media event and look for larger groupings that aggregate weblogs based on interest. While we expect the social network to have few areas of high clustering, we believe that each media event will find a target

population that is regular with respect to the subject matter. Clustering these data around the membership of each event, we expect to find cohesive clusters of individuals relating them to general media categories (e.g. politics, humor, or technology).

Finally, our fourth analysis will look at the aggregate characteristics of individuals in the system over time. Like many previous diffusion studies, we will explore the adoption characteristics of individuals to determine whether or not individuals fall into specific adopter categories (Rogers 2003). As shown previously (Marlow 2002), we expect to find that these categories are extremely contextual in nature, where for some set of events certain individuals will tend to be earlier adopters, while for others they will trail.

Aim 3

Describe the demographics and population characteristics of the current weblog community.

To evaluate the representativeness of this study and place the weblog medium in a more general context, we will use our survey apparatus to explore demographics and population characteristics of those authors. Standard demographic variables will be obtained before the survey begins in order to characterize dropout and nonresponse from the subject population.

Data collected from the demographic portion of the survey will be compared to US Census estimates for 2003 including race/ethnicity, age, sex, zip code, education, income and marital status. We expect that the population of weblog authors will be slightly biased towards Caucasian, younger, male individuals living in more urban locations. We also expect to find a slightly lower education level than average due to the predominance of younger subjects in the sample. Likewise, income and marital status should follow this trend.

The first section of the survey will address the use of weblogs and current activity with the medium to be measured globally using standard population metrics: growth (new weblogs); turnover (dead weblogs); in-migration (new weblogs by previous authors); out-migration (dead weblogs by previous authors); and overlap (multiple weblogs by the same author). For each uniquely identified weblog they have maintained we will attain the age of the weblog, its migration to a new site, and whether or not it is currently active. We expect that while previous estimates of weblog population have shown rapid exponential growth based on the total number of weblogs in existence, the number of active weblogs at any given point in time will also follow an exponential form far below the total. These measures will be used to determine an accurate size and growth rate for the community allowing us to predict the overall saturation at future points in time.

Both of these survey areas will be compared to the personal network characteristics and diffusion properties of the subjects obtained from Aims 1 and 2. From this comparison we will have a first look at the influence of personal characteristics on the structural properties of an online medium. We expect that age and length of use of the medium will be the two most determining factors both in size of personal networks and influence via degree centrality. These data will provide a much needed analysis on the relationship between the author's age, their time using the medium, and the number of ties that they maintain within their weblog activities.

Aim 4

Compare weblog ties and non-weblog ties of authors and describe the effect of weblogs on the networks of authors. In order to understand the role of weblogs in personal interactions we must ask the authors themselves to compare the medium to other forms of communication they may use to find new, online social contacts and support the ties they already have. This portion of the survey will also examine the extent to which weblogs are integrated into the author's social network and the role that they play in informing friends, family and acquaintances with information about their lives and interests.

The analysis will be supported by two measures of personal networks. First, a name generator will be used to explicate the communication network for the subject on a given day. The resulting ego network will reflect the author's strong ties, namely those people that are part of the intimate group of individuals with whom the author communicates on a regular basis.

From these data we can infer the different subgroups that exist among the alters, namely family, friends, business associates and the like. Our first analysis will examine the extent to which strong ties have access to, and regularly take part in a subject's weblog experience. Given that the weblog medium is typically used as an open digest of an author's thoughts, there is no barrier for these strong ties to be readers of the weblog. However, we expect that most weblog users will constrain the extent to which their strong ties have knowledge of and in some cases access to their site. In the case that clear subgroups of alters exist, we expect that this implied security will exist only within certain groups, i.e., family or business associates.

The second network survey, the position generator, will be employed to examine the weak ties of subjects. Subjects will be asked to identify individuals they know holding occupations from a standard list ranging in SES score. For each associate, respondents will be asked about their relationship and whether it was created and is maintained by their weblogging activity. As the position generator is a common method for measuring social capital, this section of the survey will allow us to contextualize the weblog community with respect to the American population in

general. Controlling for age we expect to find a bias towards individuals with higher social capital, suggesting that existing social capital influences the impetus to maintain a weblog. Furthermore, insofar as the medium is still young, we anticipate that most of the social capital of weblog authors will come from ties offline or through other online media, and that the amount of time weblogging will be a determining factor of weblog social capital formation.

The data collected from both network generators will be compared to the networks observed by the diffusion aggregator, allowing us to determine whether or not standard social network measures are effective at predicting the size and diversity of associates created within the weblog environment. We anticipate that both the size and diversity of offline ties will be strongly correlated with those found in weblogs, and that in many cases a large number of ties online came from previous offline interactions.

Contributions

The contributions of this thesis are divided between the studies of online communities, the diffusion of information and social networks, both in terms of methodology and empirical results. First, this research represents a novel approach to the analysis of the diffusion of information, yielding a data set containing millions of actors and millions of events. Compared to previous diffusion studies that relied on surveys and much smaller communities, this thesis will present an analysis of media contagion on an entirely different scale, allowing for a more comprehensive picture of the process, albeit in one medium.

Second, given the scale of data collection, this thesis will tackle a number of methodological issues attached to the analysis of large social network data sets. Many of the standard network measures employed by researchers today are computationally intractable for data sets of the magnitude of this work, and the methods contained in this thesis represent new techniques for porting older measures into this new scale. In particular, new measures of authority, community and the epidemiology of media contagion are needed for the analysis, all of which will hopefully be useful in the analysis of other sizeable data sets.

The final contribution is a clear and accurate picture of an emerging online medium, including some of the effects it is having on the lives of the authors. Given a continued exponential growth of weblogs within society, this work will be a cornerstone for the future analyses of their aggregate effect on our acquisition of media events. As we expect that many of the effects of weblogs are similar to other online media, these analyses will hopefully prove useful in the study of analogous interactions, and the results generalizable to a wider set of online social relations.

Timeline

The timeline for this thesis is arranged around an expected completion date in November, 2004. Since most of the development and evaluation measures have been tested and piloted over the past two years of work, I expect this is a suitable amount of time to complete the research and draft the thesis document. Following is a general outline of activities over the next five months:

June: Development and testing of diffusion aggregator and beginning of diffusion data collection.

July: Data collection and development of survey apparatus. I will also complete a draft of the background section of the thesis document.

August: Continued data collection, selection of survey population, and deployment of survey apparatus.

September: Data Analysis and writeup of results

October: Writing of the thesis document

November: Thesis Defense

Resources and Budget

Since the Blogdex project already has resources for its continued operation at the Media Laboratory, no additional equipment will be needed for the completion of this thesis. As the survey apparatus will be conducted in an online environment, survey development and analysis will be within my means as well. The only resources requested will be for committee travel at the time of the thesis defense.

Bibliography

- Adar, Eytan. 2004. Implicit structure and the dynamic of blogspace. Paper read at Workshop on the weblogging ecosystem: Aggregation, analysis and dynamics (at WWW2004), at New York, NY.
- Anderson, Roy, and Robert M May. 1992. *Infectious diseases of humans: Dynamics and control*. Oxford: Oxford University Press.
- Bailey, Norman T. J. 1975. *The mathematical theory of infectious diseases and its application*. 2nd ed. New York, NY: Oxford University Press.
- Barabási, Albert-László. 2002. *Linked : the new science of networks*. Cambridge, Mass.: Perseus Pub.
- Bogart, Leo. 1951. The spread of news on a local event: A case history. *Public Opinion Quarterly* 14 (4):769-772.
- Burt, Ronald S. 1980. Models of Network Structure. *Annual Review of Sociology* 6:79-141.
- . 1987. Social contagion and innovation: Cohesion versus structural equivalence. *The*

- American Journal of Sociology* 92 (6):1287-1335.
- . 1993. The social structure of competition. In *Explorations in economic sociology*, edited by R. Swedberg. New York: Russell Sage Foundation.
- Coleman, James S. 1988. Social Capital in the Creation of Human Capital. *American Journal of Sociology* 94 (Issue Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure):S95-S120.
- Dawkins, Richard. 1989. *The selfish gene*. New ed. Oxford New York: Oxford University Press.
- Deutschmann, Paul J., and Wayne A. Danielson. 1960. Diffusion of knowledge of the major news story. *Journalism Quarterly* 37:345-355.
- Dezso, Zoltán, and Albert-László Barabási. 2002. Halting viruses in scale-free networks. *Physical Review E* 65 (055103).
- Doherty, Irene, Nancy Padain, Cameron Marlow, and Sevgi Aral. Forthcoming. Determinants and consequences of sexual networks as they impact STI spread. *The Journal of Infectious Diseases*.
- Feld, Scott. 1982. Social structure determinants of similarity among associates. *American Sociological Review* 47:797-801.
- Feld, Scott L. 1982. Social structural determinants of similarity among associates. *American Sociological Review* 47 (6):797-801.
- . 1991. Why your friends have more friends than you do. *The American Journal of Sociology* 96 (6):1464-1477.
- Feld, Scott L., and Richard Elmore. 1982. Patterns of sociometric choices: Transitivity reconsidered. *Social Psychology Quarterly* 45 (2):77-85.
- Fischer, Claude. 2001. *Bowling Alone: What's the Score?* Anaheim, CA: American Sociological Association.
- Fischer, Claude S. 1982. *To dwell among friends: personal networks in town and city*. Chicago: University of Chicago Press.
- Fischer, Claude S., and Lynne McCallister. 1978. A procedure for surveying personal networks. *Sociological Methods and Research* 7:131-148.
- Freeman, Linton C. 1978. Centrality in social networks conceptual clarification. *Social Networks* 1 (3):215-239.
- Funkhouser, G. Ray, and Maxwell E. McCombs. 1971. The rise and fall of news diffusion. *Public Opinion Quarterly* 35 (1):107-113.
- Garton, Laura, Caroline A. Haythornthwaite, and Barry Wellman. 1997. Studying online social networks. *Journal of Computer Mediated Communication* 3 (1):75-105.
- Gladwell, Malcolm. 2000. *The tipping point: how little things can make a big difference*. 1st ed. Boston: Little Brown.
- Granovetter, Mark. 1973. The Strength of Weak Ties. *The American Journal of Sociology* 78 (6):1360-1380.
- . 1983. The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory* 1:201-233.
- Greenberg, Bradley S. 1964. Person to person communication in the diffusion of news events. *Journalism Quarterly* 41:489-494.
- Hampton, Keith, and Barry Wellman. 2003. Neighboring in Netville: How the internet supports community and social capital in a wired suburb. *City and Community* 2 (4):277-313.
- Katz, Elihu, and Paul F. Lazarsfeld. 1955. *Personal influence*. Glencoe, IL: Free Press.
- Killworth, Peter, Eugene Johnsen, H. Bernard Russell, Gene Ann Shelley, and Christopher McCarthy. 1990. Estimating the size of personal networks. *Social Networks* 12:289-312.
- Kraut, R., V. Lundmark, M. Patterson, S. Kiesler, T. Mukopadhyay, and W. Scherlis. 1998. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist* 53 (9):1017-1031.
- Lee, Nancy Howell. 1969. *The search for an abortionist*. Chicago, : University of Chicago Press.

- Lenhart, Amanda, John Horrigan, and Deborah Fallows. 2004. Content creation online: Pew Internet and American Life Project.
- Lin, Nan. 1999. Building a network theory of social capital. *Connections* 22 (1):28-51.
- . 2001. *Social capital : a theory of social structure and action, Structural analysis in the social sciences ; 19*. Cambridge, UK New York: Cambridge University Press.
- MacElroy, William, and Michael Gray. 2003. *IMRO online survey satisfaction research: A pilot study of salience-based respondent experience modeling 2003* Available from http://www.ijor.org/ijor_archives/articles/survey%20sat%20article%2007.09.03.pdf.
- Marlow, Cameron. 2002. Getting the scoop: Social networks for news dissemination. Paper read at Sunbelt International Social Networks Conference, at New Orleans, LA.
- . 2003. Modeling emergent communities through diffusion. Paper read at Sunbelt International Social Networks Conference, at Cancun, Mexico.
- . 2004. Audience, structure and authority in the weblog community. Paper read at 54th Annual Conference of the International Communication Association, at New Orleans, LA.
- Marsden, Peter. 1984. Measuring tie strength. *Social Forces* 63:482-501.
- Marsden, Peter V. 1987. Core discussion networks of Americans. *American Sociological Review* 52 (1):122-131.
- . 1990. Network data and measurement. *Annual Review of Sociology* 16:453-463.
- McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415-444.
- Milgram, Stanley. 1967. The small world problem. *Psychology Today* 1 (1):60-67.
- Nie, N. 2001. Sociability, interpersonal relations, and the internet: Reconciling conflicting findings. *American Behavioral Scientist* 45 (3):420-435.
- Nie, N., and L. Erbring. 2000. Internet and society: A preliminary report. Stanford, CA: Stanford Institute for the Quantitative Study of Society: Stanford University.
- Pastor-Satorras, Romulado, and Alessandro Vespignani. 2001. Epidemic spreading in scale-free networks. *Physical Review Letters* 86 (14):3200-3203.
- Patton, Phil. 1986. *Open Road*. New York: Simon and Schuster.
- Perseus Development Corporation. 2004. The blogging iceberg. Braintree, MA: Perseus Development Corporation.
- Portes, Alejandro, and Patricia Landolt. 1996. The downside of social capital. *The American Prospect* 26:18-21.
- Putnam, Robert D. 2000. *Bowling alone : the collapse and revival of American community*. New York: Simon & Schuster.
- Rapoport, Anatol. 1953. Spread of information through a population with socio-structural bias. *Bulletin of Mathematical Biophysics* 15:523-543.
- Rogers, Everett M. 2003. *Diffusion of innovations*. 5th ed. New York: Free Press.
- Shirky, Clay. 2003. *Power laws, weblogs and inequality* [Newsletter] 2003 Available from http://www.shirky.com/writings/powerlaw_weblog.html.
- Valente, Thomas W. 1995. *Network models of the diffusion of innovations, Quantitative methods in communication*. Cresskill, N.J.: Hampton Press.
- van der Poel, Mart G. M. 1993. Delineating personal support networks. *Social Networks* 15:49-70.
- Wasserman, Stanley, and Katherine Faust. 1994. *Social network analysis: methods and applications*. New York: Cambridge University Press.
- Watts, Duncan J. 1999. *Small worlds : the dynamics of networks between order and randomness, Princeton studies in complexity*. Princeton, N.J.: Princeton University Press.
- . 2003. *Six degrees : the science of a connected age*. 1st ed. New York: W.W. Norton.
- Weimann, Gabriel. 1982. On the importance of marginality: One more step in the two-step flow

- of communication. *American Sociological Review* 47 (6):764-773.
- Wellman, Barry. 1987. *Different strokes from different folks: which ties provide what kinds of social support*. Berkeley: Institute of Urban and Regional Development University of California Berkeley.
- . 1997. Structural analysis: From method and metaphor to theory and substance. In *Social structures: A network approach*, edited by B. Wellman and S. D. Berkowitz. Greenwich, CT: JAI Press.
- . 1999. *Networks in the global village : life in contemporary communities*. Boulder, Colo.: Westview Press.
- . 2001. Computer networks as social networks. *Science* 293 (14):2031-2034.
- . 2002. Little boxes, glocalization and networked individualism. In *Computational and Sociological Approaches*, edited by M. Tanabe, P. van den Besselaar and T. Ishida. Berlin: Springer.
- Wellman, Barry, and Milena Gulia. 1997. Net surfers don't ride alone: Virtual communities as communities. In *Networks in the global village*, edited by B. Wellman. Boulder, CO: Westview.
- Wellman, Barry, and Caroline A. Haythornthwaite. 2002. *The Internet in everyday life*. Edited by W. Barry and H. Caroline. Malden, MA: Blackwell Pub.
- Woolcock, Michael. 1998. Social capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society* 27:151-208.
- . 1998. Social capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society* 27:151-208.

Biographies

Cameron A. Marlow is a PhD student in the electronic publishing group at the MIT Media Laboratory, and creator Blogdex, a service that tracks the diffusion of ideas through the population of bloggers. Marlow recently held a fellowship at the Centers for Disease Control and Prevention studying the effects of the Internet on the spread of sexually transmitted diseases. He holds a B.S. in computer science from the University of Chicago and a M.S. from MIT in Media Arts and Sciences.



Keith N. Hampton is Assistant Professor of Technology, Urban and Community Sociology and holds the Class of '43 Career Development Professorship in the Department of Urban Studies and Planning at MIT. He received his Ph.D. and M.A. from the University of Toronto in sociology, and a B.A. in sociology from the University of Calgary. His research interests focus on the relationship between information and communication technologies, social relationships, and the urban environment.

Thomas W. Valente directs the Master of Public Health Program in the Department of Preventive Medicine at the USC Keck School of Medicine. He received a B.S. in mathematics from Mary Washington College, an M.S. in mass communication from San Diego State University, and a Ph.D. from the Annenberg School for Communication at the University of Southern California. Valente spent nine years at the Johns Hopkins University School of Public Health from 1991 to 2000 conducting research and teaching health communication, program evaluation, and network analysis.

Appendix I: Sample survey

1. Demographics
 - 1.1. Gender
 - 1.2. Birth year
 - 1.3. Highest level of education completed
 - 1.4. Marital status
 - 1.5. Race/ethnicity
 - 1.6. Zip code of current residence
 - 1.7. Household income
2. Weblog use
 - 2.1. How many weblogs have you ever created?
 - 2.2. For each weblog, please answer the following questions:
 - 2.2.1. When was the weblog created?
 - 2.2.2. Is it still active? If not, please specify the date you last posted to it.
 - 2.2.3. Did it ever change locations? If so, how many times?
 - 2.2.4. How often did you post to this weblog?
 - 2.2.5. How many authors did the weblog have (including yourself?)
 - 2.3. How much time did you spend last week working on your weblog?
3. Social networks: Name generator
 - 3.1. Please list the first names and last initials of each individual you communicated with on the most recent weekday. Examples types of communication: email, instant messaging, face to face, or over the telephone. Feel free to use your email client or instant messaging client to aid in this process.
 - 3.2. For each individual please answer the following questions:
 - 3.2.1. Demographics (age, gender, education, marital status, race, location)
 - 3.2.2. What form of communication do you typically use to contact them?
 - 3.2.3. What is this person's relation to you?
 - 3.2.4. Does this person read your weblog?
 - 3.2.5. List all of the other individuals this person knows (from list)
4. Social networks: Position generator
 - 4.1. Please check the box next to any of the following occupations if you know someone who holds that job.
 - 4.2. For each of the jobs checked, please answer the following questions
 - 4.2.1. What is this person's relation to you?
 - 4.2.2. What form of communication would you use to contact them?
 - 4.2.3. Does this person read your weblog?
 - 4.2.4. Does this person have a weblog?