The Structural Determinants of Media Contagion

by

Cameron Alexander Marlow

B.S. Computer Science
University of Chicago, 1999

M.S. Media Arts and Sciences
Massachusetts Institute of Technology, 2001

SUBMITTED TO THE PROGRAM IN MEDIA ARTS AND SCIENCES, SCHOOL OF ARCHITECTURE AND PLANNING, IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER, 2005

Author _____
Department of Media Arts and Sciences
August 2, 2005

Certified by _____
Walter Bender
Senior Research Scientist
Department of Media Arts and Sciences
Thesis Supervisor

Accepted by _____
Andrew Lippman
Chairman
Departmental Committee for Graduate Students
Department of Media Arts and Sciences

Thesis Committee

Thesis Supervisor _____

Walter Bender
Senior Research Scientist
Massachusetts Institute of Technology
Department of Media Arts and Sciences

Thesis Reader _____

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

Thesis Reader _____

Tom Valente
Associate Professor
University of Southern California
Department of Preventative Medicine

The Structural Determinants of Media Contagion

by

Cameron Alexander Marlow

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

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. Weblogs have recently emerged as a part of our media ecology, and incidentally engender this process of media contagion; because weblog authors are tied by social networks of readership, contagious media events happen frequently, and in a form that is immediately measurable.

The generally accepted notion of media diffusion is that it occurs through two channels: externally, as applied by a constant force such as the mass media, and internally through socio-structural means. Sitting between our traditional notions of mass media and the public, weblogs problematize this classical theory of mass media influence. This thesis aims to elucidate the role of weblogs in media contagion through a sociological study of this community in two parts:

First, I will address the issues of modeling the social structure of weblogs as observed through their readership network, and the various media events that occur therein. Using a large weblog corpus collected over a one-month period, I have constructed a model describing the structure of popularity and influence from the extracted readership network, and will show that this model more accurately describes the weblog network. I will also derive a typology of media events from collected examples using features of structural and non-structural diffusion.

Second, the extent to which these data are reflective of actual social processes as opposed to artifacts of data collection and aggregation will be explored. To validate the models presented in part one, I have conducted a survey of randomly selected authors to examine their social behaviors, both in weblog use and otherwise. I will characterize the range of weblog uses and practices, presenting an analysis of personal influence in the blogging community.

Thesis Supervisor: Walter Bender
Title: Senior Research Scientist

Acknowledgments

This thesis is dedicated to the memory of Redmond Lyons-Keefe, a best friend whose kindness and laughter will be dearly missed.

Thanks

This work could not have been accomplished without the help and support of the following individuals:

THE A LIST

To my committee, Walter Bender, Keith Hampton and Thomas Valente. Also my family, who made it possible for me to be at MIT in the first place.

THE AA LIST

Linda Peterson and Pat Solakoff, EP: Erik, Joanie, Scotty, Sunil, Vadim, Carla, Ouko and Larissa. UG: Scorchio, The Pud, You're Fired, Shut up, Daddy Long, the Iz Rocka, Gys, JFlower, Eddie, Tek Fu and TFY2005. Aisling, of course. All of the MIT students, faculty and expats who have helped me along the way: Jeana, Nikita, Jess, Aaron, Jeff, Max, Min, Min Suh, Jonah and Andrea. My family at the CDC, especially the weak tie, Victoria Gammino. All my crew in A-town, Zach, Tom, and Blondie. This font, Hoefler Text.

THE AAA LIST

Andy Baio, Erik Benson, danah boyd, Michael Buffington, Maciej Ceglowski, Tom Coates, Anil Dash. Nick Denton, Redrick DeLeon, Cory Doctorow, Ze Frank, Rusty Foster, Ryan Gantz, Matt Haughey, Alison Headley, Scott Heiferman, Meg Hourihan, Matt Jones, Jason Kottke, Jason Levine, Leonard Lin, Erica Lucci, Peter Merholtz, Joshua Schachter, Tim Shey, Ben and Mena Trott, David Weinberger, and anyone who writes something called BLOG.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

On November 22, 1963, NBC Correspondent Frank McGee declared, "This afternoon, wherever you were and whatever you might have been doing when you received the word of the death of President Kennedy, that is a moment that will be emblazoned in your memory and you will never forget it...as long as you live." For this particular event, wherever they were, over 50% of America were talking to someone else, because that person was relaying the horrible news. Whenever a catastrophic news event occurs, the probability that we hear about it from another person increase dramatically; this might not come as a surprise to someone who has experienced an event of this magnitude, which most people have. But what might be unexpected is the fact that as news becomes increasingly *irrelevant* to the rest of the population, your odds of finding it through interpersonal communication start to *increase.*

These were the findings of Bradley Greenberg in a study of news diffusion in 1964. In the cases where news is extremely important, people feel an urgency to inform each other (Greenberg, 1964). Perhaps this form of socialization is a way of dealing with tragic news, forcing us into a situation where we cope as a group, or maybe our inability to believe such shocking news compels us to seek out more information from whomever is nearby. There are probably a number of explanations for behavior in the wake of flashbulb news, but very little of this lends to our understanding of why tiny news stories with extremely specialized audiences also reach their audience through word-of-mouth.

The justification provided by Greenberg is that the mass media is efficient at delivering news that most of its audience is interested in. But when a story becomes very specific, it loses general context and falls below the fold. To some percentage of the population though, this story might have substantial personal significance. Assume I live in a large city, and that thumbing through the local news I notice a story about a school robbery of some computer equipment. 99 percent of the city's population might not even finish reading the headline, but what if one of my friends teaches at this particular school?

What if my brother just bought a computer from a shady character on a nearby street? In these cases, I would make an effort to pass this news on to the relevant people.

Informal communications between friends, colleagues and strangers are a fundamental mode for disseminating news and ideas. 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.

Collective communications

Imagine a world where people could choose to have all of their personal, written communications be publicly available in a persistent and searchable repository for all eternity-call this the "collective communications repository" (CCR). This includes conversations about personal affairs, political opinions, current events, and nearly every other imaginable topic that might arise. Our first thought when encountering this scenario is to dwell on a host of negative side-effects: that our innermost thoughts about friends and family might be read by strangers, that these potential onlookers might make us censor our feelings, that we might be confronted at a much later date for something we said years ago, and so on.

But there are three important upsides that are less immediately apparent: first, by having my conversations online, I invite people with similar interests, opinions, problems, desires—in other words, like-minded individuals—to introduce themselves to me. I could be sitting next to my potential soul-mate on the subway, but because of my lack of time or attention, I might never talk to them. Having my interactions available publicly increases the chances of serendipitous encounters like this one actually being realized.

Second, if many of my friends and acquaintances also made their social lives public, I could stay up-to-date with their lives on my own time. With busy lives and schedules, we often find it hard to spend as much time as we would like with our friends. As a result, sometimes we miss the important events and conversations that make us feel connected to each other; as a result, the strength of our relationships can sometimes fade. If I could catch up with my friends virtually, namely by paying attention to the goings-on in their lives

from the comfort of my home, the bounds of time-constrained, collocated interaction could be broken.

The third benefit is less intuitive, and happens at a systemic level above these personal interactions. If we assume that a large percentage of my social contacts' are in this CCR, and that we are keeping up with each other in this manner, we enable media contagion to arise in ways not otherwise possible. Returning to the news story of a local school robbery, let us assume that I forward this on to my teacher friend because she works at the school in question. Although she would normally call all of her fellow teachers at the school and pass the news on, assuming she is on vacation, this might not happen in time. However, if her communications are being followed by friends through this public repository, the obscure news story will be passed on to them as soon as they check.

While the CCR may sound absurd, there is a recent web-based communication medium that comes very close to replicating the scenario. The medium is a personal publishing technology known as a *weblog*: written mostly by individuals, weblogs are regularly-updated, personal journals of thoughts, stories, news and ideas of interest to the author in a publicly-accessible web site. Each weblog author has a set of friends and colleagues who also maintain weblogs; in addition to talking about my life and experiences, as a weblogger I will also respond to the things that my friends are talking about.

The main difference between the CCR and the community of webloggers is that weblogs are a broadcast medium, not a public record of all communications. Although I may use many other media to communicate with my friends, some of this may happen explicitly within our weblog writing. For example, in the case of the school robbery, I might send the news story to my friend via email, but because the story is of interest to her, she will write about it on her weblog. From there, any of her friends that are also teachers would find out about it quickly; if her readers find the story personally engaging, they too might right about it on their weblogs.

Putting all of the questions of negative side-effects aside, it is easy to see that weblogs share each of the three properties of the CCR I have described:

1. *Weblogs allow for serendipitous social relationships.* Because weblogs are public, on the web, and indexed by major search engines, someone I do not have any prior social contact with could read something I have written and strike up a conversation with me.

2. *Friends can stay in touch on their own time.* Like email, reading weblogs does not require both parties to be paying attention at the same time. But unlike email, weblogs are not necessarily a two-way medium, and I can keep up with my friends without having to respond.

3. *Weblogs engender media contagion.* Because all of this interaction happens between connected individuals, ideas, stories and opinions can easily spread to those weblog authors and readers who are engaged by them.

This thesis explores these processes by addressing two underlying questions: First, what does the social structure of weblog authors look like, and how does it relate to their offline social ties? Second, how can media contagion be described, and to what extent does the social structure play a role in this process? Before positing any hypotheses, let me first address these two areas in more detail.

SOCIAL STRUCTURE

At the writing of this thesis, it is estimated that 67% of American residents use the internet regularly (Pew Internet and American Life Project, 2005). For these individuals the internet provides a number of different informational services, but the most popular use is social (Wellman, Quan-Haase, Witte, and Hampton, 2001). Just as the telephone allowed us to keep in contact with distant friends and family while also supporting our local relationships (Fischer, 1982), the internet allows for additional means to stay in touch with our existing relations. However, the internet provides two new types of relationships that were otherwise not possible: *online ties*, or acquaintances cultivated online based on shared interests (Rheingold, 1994; Sproull and Kiesler, 1991), and *latent ties*, or relationships that are available but have yet to be activated by some form of social contact (Haythornthwaite, 2002).

The number of weblogs has grown exponentially from their inception around 1998, and current estimates put the total at 36 million worldwide (Perseus Development, 2005) in over 30 different languages. Most research into the computer mediated communication (CMC) relies on self-reported statistics within surveys to understand the relationships that exist within a given medium. Unlike other popular forms of online interaction, weblog authors make their social ties explicit through links to the other weblogs they read. For this reason, they represent a unique opportunity for studying the social relations of an online community.

Unfortunately, webloggers do not tell us what these readership links actually mean: are the other people friends, acquaintances, or individuals the author

has never met? Are they local, and if so, how often do they meet each other? These questions cannot be answered without explicitly asking the authors.

MEDIA CONTAGION

Observing media contagion from the perspective of the information, one can draw analogies to the propagation as a sort of disease spreading through a network of organisms. In some cases, an external source might be the cause of infection, such as a water source, air ventilation or processed foods. In the case of diffusing media, the comparable external force would be mass media, such as radio, television or newspapers. Infections can also spread directly from person to person, which is the case I have been describing as media contagion thus far. In many cases however, the diffusion of any piece of media can be internal (person-to-person), external, or some combination of the two.

Previous research in the diffusion of information has focused primarily on *innovations*, or ideas or practices that are adopted by individuals with some associated cost (Rogers, 1962; Valente, 1995). Innovations are prime examples for diffusion research because they occur at a slow enough rate that as to be tracked effectively by a team of researchers. Because of the high cost associated with collecting these data, only a few examples of diffusion have been collected over the past 70 years (Valente and Rogers, 1995), and these cases have been reanalyzed extensively (Valente, 1995; Rogers, 2000; Burt, 1987).

Every day tens of thousands of media events occur within the weblog community (Marlow, 2003), each one associated with an observable social structure. In the course of a day, weblog authors produce more diffusion data than all previous research into innovations, and in a forum that is observable by both humans and computers. By automating the process of diffusion monitoring with web robots, the weblog community could serve as an invaluable resource for diffusion data.

Looking ahead

This chapter is meant to serve as a cursory introduction to media contagion, the weblog community, and the questions being addressed by this thesis. The rest of this document is arranged as follows:

Chapter 2 provides an extensive review of the background literature upon which this thesis is based. There are three major areas of that I will be using extensively in my design and analysis: *social networks*, or the system of thought

that describes social phenomena through its structural properties; *diffusion studies*, or the amalgamation of a variety of work in the diffusion of information, including news, innovations and opinion; finally, *computer-mediated communication*, or the study of the effect of computer-based media on our social behaviors.

Chapter 3 outlines the design and methodology I have chosen to explore the social relationships of webloggers and the media contagion that arises therein. This work involves two basic research endeavors: (1) a comprehensive, automated system to track the weblog readership network and media contagion events, and (2) a general social survey to understand the greater social context of the medium and validate the findings of the first endeavor.

Chapter 4 shows the results and detailed analysis of the thesis.

Chapter 5 provides a summary of the important findings and conclusions obtained throughout the thesis.

Weblogs represent a novel form of interpersonal communication: all interactions are public, persistent over time, searchable, and broadcast instead of person-to-person. Surely they are not meant to replace any of the types of communication that we already use, but rather they serve as an extension of our personal interests into public sphere. Russell Neuman predicted this evolution almost 15 years ago:

> We are witnessing the evolution of a universal, interconnected network of audio, video and electronic text communication that will blur the distinction between interpersonal and mass communications and between public and private communications. (Neuman, 1991, p. 12)

More than their novelty, weblogs represent an important chance to study the structure of a social system from an omniscient perspective. Data that would otherwise be impossible to collect, or involve large ethical questions, are provided free of cost to researchers. They also represent the first opportunity to study the process of media contagion in detail, and to understand how small, local interactions can give rise to large, emergent media events.

Chapter 2

Background

The chapter is arranged around three basic areas of research that are themselves interdisciplinary pursuits: Social networks, a subfield of sociology concerning itself with the structural analysis of social relations; computer mediated communications, a multi-disciplinary look at the effect of online interactions on communication behavior; finally, diffusion studies, an amalgam of communications studies focusing on how innovation and news move through society. In addition I will provide a short history and research for the emerging medium of weblogs.

2.1 SOCIAL NETWORKS

While social networks have a long and storied past, they came into the public's attention and became a cohesive field after 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 of behavior (Wellman, 1997) and a methodology for analyzing social interactions (Wasserman and Faust, 1994).

ACTORS

TIES

ALTERS

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 maintains. Friends, family, and acquaintances can all be seen generically as social ties, or *alters* of that person. Each alter can provide any number of different resources, as well as indirect access to resources contained in

Theory

When people speak of social network analysis, they typically refer to the field as a ***methodology*** for analyzing social behaviors which can be used in conjunction with other, non-structural approaches to social behavior. This common simplification ignores one of the motivating force for using these methods, namely the theoretical underpinnings that give relevance to SNA. Wellman (1997) has provided a history of structural analysis, and gives a simple description of how the structural approach differs from others:

1. Behavior is interpreted by structural constraints on activity, rather than inner forces within an individual.
2. Analyses focus on the relations between people instead of trying to sort individuals into categories.
3. Individuals should be interpreted with respect to the *whole* structure of their network, not only as individual pairwise relationships
4. It cannot be assumed that networks break down into discrete groups, and must first be interpreted solely as a network.
5. These structural measures should supplement and sometimes replace statistical measures that require independent units of analysis.

There are a number of different theoretical findings of social organization based on the structural approach—far too many for the scope of this thesis. However, the concepts of ***homophily***, ***tie strength***, and ***social capital*** will be relevant to my analysis, so I will introduce them in detail here.

HOMOPHILY

One concept that demonstrates the importance of structural analysis is ***homophily***, the observation that people who are socially related tend to have similar characteristics: age, ethnicity, class, and so on. A cursory look at a group might lead one to believe that these properties define a sort of category membership, for instance an "older kids' club." But a closer look at these similarities often reveals that the structure predicts the category, not the other way around.

FOCI OF ACTIVITY

Feld (1982) observed that most non-familial social ties arise from various ***foci of activity***: work projects, clubs, and neighborhood groups are all popular locations in which personal relationships are formed. Feld found that the structure of these activities attracts people of similar characteristics; in the case of work relations, social ties tend to be twice as homogeneous as would be expected by chance. In a similar way, geographic propinquity often causes

of similarity, as neighborhoods, suburbs and towns can become sinks of similar people (McPherson, Smith-Lovin, and Cook, 2001).

Another source of homophily that is observed in race similarities is *inbreeding homophily*, where existing homogenous relations give rise to more relations of the same type. Shrum, Cheek, and Hunter (1988) found that in the third grade, cross-race relationships were extremely unlikely, and that until the majority race reached 90%, the racial groups remained cohesive and separate† . Inbreeding homophily is also often a factor in class and socioeconomic similarities among ties (McPherson et al., 2001).

INBREEDING HOMOPHILY

Above this breakpoint, the two usually integrate.

TIE STRENGTH

An important distinction within SNA is the classification of relations based on *tie strength*. The abstract tie that connects two actors in a network can indicate a number of different meanings: it could represent an old, familial tie, a new acquaintance, or even a negative relationship. The measurement of tie strength has been ascribed to a number of factors, including the length of a relationship, time spent together, intensity, amount of communication, and the breadth of topics discussed. Marsden (1984) has compared tie strength operationalized both by the length of a relationship and the intensity and found that the best measure is in the "closeness" of to people; he has found that the best way to distinguish tie strength is to simply to ask if an individual "feels especially close" to the alter.

One of the most important theoretical developments in the field of social networks was Granovetter's work on the *strength of weak ties* (Granovetter, 1973, 1983). It was long thought that increased tie strength implied more access to resources through that tie, as stronger ties would be more willing to lend their support. Granovetter provided evidence that this assumption was not always the case; the stronger one's relationship with another person is, the more likely it is that they have the same friends and acquaintances, and in turn access to the same social and physical capital. Weaker ties, on the other hand, connect an individual to other people and networks that are not immediately available to the individual. In this respect, referred to by Granovetter as a *bridging tie*, weak ties can provide more information than would be available through stronger ties.

STRENGTH OF WEAK TIES

BRIDGING TIE

A corollary to the *strength of weak ties* argument is the concept of *structural holes*, first described by Burt (1992). A structural hole is simply the same concept of the bridging tie illustrated by Granovetter, except rephrased in terms of competition; structural holes are simply places in a network where

STRUCTURAL HOLES

one individual brokers the information that flows between two groups. With this observation, individuals who wish to be in a structurally advantageous position can intentionally create structural holes by not introducing alters who must otherwise communicate through him or her.

While much attention is paid to weak ties in social network literature, the strong tie is equally important for different reasons. Weak ties may provide information or resources not available through close friends and family, but strong ties provide emotional support, higher levels of intimacy, more self-disclosure, general reciprocity, and more frequent interaction (Wellman and Gulia, 1999a; Granovetter, 1983; Fischer, 1982). In some cases, such as job attainment, strong ties can be just as instrumental in providing hard-to-find resources as weak ones (Lin, 1981).

SOCIAL CAPITAL

Social capital, a concept pioneered by Bourdieu (1985) but reintroduced and popularized by Coleman (1988) is one of the more popular concepts to arise from sociology in the past few decades. Simply defined, "whereas economic capital is in people's bank accounts, human capital is inside their heads, social capital inheres in the structure of their relationships" (Portes, 1998). Early work on social capital operationalized the concept in terms of structural features: strong ties (Lin, 1981), weak ties and structural holes (Burt, 1992), or density (Bourdieu, 1985), each measured in accordance to some end goal that motivated the definition, such as job attainment or access to information and other resources.

Robert Putnam repurposed the term to have an entirely different meaning, equating it with civic engagement, or "features of social organizations, such as networks, norms, and trust, that facilitate action and cooperation for mutual benefit" (Putnam, 1993, p. 35). Instead of using social capital as a measure of *individual wealth*, Putnam considered it as a *collective* measure and while both definitions made intuitive sense, they were in direct conflict with each other. Theoretical models have since been developed to integrate these two views (Woolcock, 1998; Portes, 1998; Lin, 2001); a simple model proposed by Quan-Haase and Wellman (2004) describes three types of social capital that can be independently measured: network capital, participatory capital and community commitment, the first of which relates the original definition, and the latter two individualized forms of Putnam's interpretation.

INDIVIDUAL WEALTH

COLLECTIVE WEALTH

The debate over social capital is an important one because it represents much of the focus of social network research; once we can place a value on the

networks of individuals and communities, we can begin to explain the methods by which people can take advantage of their capital. If this resource exists in the same space as human and financial capital, and is easily measurable, it should be fungible to the same extent that other capital is. However, this topic is still a matter of heated debate.

Measurement

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*.

WHOLE NETWORKS

EGO NETWORKS

NETWORK MODULES

The oldest and most widely used technique for measuring a subject's social network is the *name generator*. First used by Fischer and McCallister (1978), this technique asks individuals to enumerate a set of individuals with whom they have some kind of social contact or support defined by the survey. Subjects are then given a matrix within which they define the social ties that exists between each of these alters, along with the strength of the tie and type of relation.

NAME GENERATOR

The results of this method vary significantly based on the question used to extract the network. A number of different approaches to generating names have been tested (Van Sonderson, Ormel, Brilman, and Van Linden van den Heuvell, 1990), but the most popular approach is the one used to identify "core"-networks which asks the question "with whom do you talk about personal matters." (Marsden, 1987) Because of the time required to enumerate an entire ego network, this method is usually constrained to the strong ties in an individual's ego network.

Another approach to enumerating relationships is to give subjects a set of randomly chosen names derived from some master list (e.g. phone books, company roster, etc.) and ask them to denote the names with which they have some social tie. This instrument is known as the *reverse small world* generator, and it is generally accepted as a means of acquiring the size of a person's extended social network (Killworth, Johnsen, Russell, Shelley, and McCarthy,

REVERSE SMALL WORLD

1990). Those individuals with a large number of weak ties will be able to identify many names on the list while less connected subjects will recognize fewer. Unfortunately, in order to get an accurate distribution, long lists must be asked of each subject; in the case that phone books are used to generate names, this is even more the case.

POSITION GENERATOR

Unlike the previous two instruments, two instruments have been created to look more specifically at access to resources as opposed to access to names. The first of these methods is known as a *position generator*; in this instrument, subects are presented with a list of common occupations of varying socioeconomic status (SES) and asked to denote those positions for which they have a social tie (Lin, 1999a). This instrument gathers information about a subject's access to potential resources, often considered a proxy to social capital.

RESOURCE GENERATOR

More recently some have grown wary of the focus on occupation and focused even more directly on the resources that might be extracted from one's extended network of social ties. Snijders (1999) have devised a *resource generator* which asks subjects to denote various types of social support that they have access to. This list includes a broad range of resources available from varying levels of scarcity, and assesses the value of a person's social capital by the relative value of resources they have access to. Van der Gaag, Snijders, and Flap (2005)have compared the resource and position generators in a general social survey of the Netherlands, and found that the resource generator while the resource generator is a better predictor of a subject's *instrumental* social capital, or immediate ability to extract resources. The position generator, however, was found to be as good a measure of outcomes related to social capital, and recommended for the fact that it is much more economical in terms of survey length.

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 three 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, 1999a; Van der Gaag et al., 2005).

Analysis

Once network data has been gathered, there are a number of different features that can be measured to translate the abstract graph into a richer understanding of the network. All of the following measures can be observed for both undirected and directed graphs; given that this entire thesis assumes the directionality of social ties, I will consider only the case of directed graphs.

RECIPROCITY AND TRANSITIVITY

In a directed graph, an edge (i, j) is said to be *reciprocal* if the opposite edge (j, i) also exists. For three nodes we say that the graph is transitive if the edges (i, j) and (j, k) imply a third edge (i, k) also exists. Both of these measures can be calculated as a proportion over an entire graph, giving a measure of how likely an edge in the graph is reciprocal, or how likely two edges are transitive.

DEGREE

The number of edges incident to a given node is defined as the *degree* of that node. A node with a high degree will have many incident edges while a low-degree node will have few. With directed graphs there are two separate degrees: *in-degree*, the number of edges coming into a node and the *out-degree* is the number of edges emanating from it.

IN-DEGREE

OUT-DEGREE

The degree of a node is a proxy to how connected it is locally, but in order to understand how that node relates to the rest of the network, one must consider the degrees of all other nodes. The *degree distribution* shows the relationship between degrees, typically expressed as the probability that a given node will have a given degree.

DEGREE DISTRIBUTION

DENSITY

In a given graph, there is a maximum number of edges that can exist, namely if every node is connected to every other node. For a graph with n nodes, a completely full, directed graph will have $n(n-1)$ nodes. The density (δ) of a graph is then simply the number of edges E divided by the total possible:

$$\delta = \frac{E}{n(n-1)} \tag{2.1}$$

In addition to looking at the density of an entire graph, we can also calculate the density of any subgraph therein. Most commonly this is the induced subgraph of one node (the ego), all of its alters, and any edges connecting the

PERSONAL NETWORK DENSITY

CLUSTERING COEFFICIENT

ego and its alters. This density of this subgraph is typically called the *personal network density*. A dense ego graph suggests the actor is part of a closely-tied network of alters, while a sparse graph implies a more loosely connected group. The average personal network density for the whole graph has recently been termed the *clustering coefficient* by Watts and Strogatz (1998).

PATH LENGTH

The most famous structural property of a network is that which Milgram used to describe the distance between any two people, or "6 degrees of separation." In social network analysis, this value can be calculated by looking at the paths between two nodes. Since there are a near infinite number of *possible* paths connecting any two nodes in a graph, measures of path length typically depend on the *geodesic*, or shortest path. The "degrees of separation" measure of a network is termed its *characteristic path length*, and is simply the average geodesic for the graph:

GEODESIC

CHARACTERISTIC PATH LENGTH

$$L(p) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{n} g(i,j)}{n} \tag{2.2}$$

Where $g(i,j)$ is the length of the geodesic between nodes i and j. I will return to the discussion of characteristic path length later in the section on self-organized networks.

CENTRALITY

Centrality is one of the more useful structural properties of a network as it relates an individual actor to the rest of the nodes, and can also be used as a measure of efficiency across for the entire network. Freeman (1979) is the earliest and most referenced measure of centrality, providing three distinct measures of centrality, each exposing a different property of the network.

DEGREE CENTRALITY

The most basic type of centrality is *degree centrality*, which is simply the in-degree for each node in the graph:

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k) \tag{2.3}$$

where $a(p_i, p_k)$ is the adjacency matrix, and thus 1 if an edge exists between p_i and p_k, and 0 otherwise. Because this value is dependent on the size of the graph, Freeman also provides a normalized degree centrality so that these values can be compared across graphs:

$$C'_D(p_k) = \frac{\sum i = 1na(p_i, p_k)}{n-1} \tag{2.4}$$

This value is equivalent to the proportion of the graph that is connected to a given node. The largest limitation of degree centrality is simply that it does not provide any information beyond one degree of connectivity. To remedy this situation, Freeman provides two more measures, the first of which is *betweenness centrality*, a measure of the probability that the given node lies on the shortest path between any two other nodes in the graph:

BETWEENNESS CENTRALITY

$$C_B(p_k) = \sum_{i<}^{n} \sum_{j}^{n} b_{ij}(p_k) \tag{2.5}$$

where $b_{ij}(p_k)$ is the probability that p_k lies on a geodesic connecting nodes i and j. Similar to degree centrality, we can normalize C_B to make it comparable between networks:

$$C'_B(p_k) = \frac{2C_B(p_k)}{n^2 - 3n + 2}. \tag{2.6}$$

Finally, Freeman defines his third measure of centrality, *closeness centrality*, or the average distance from a given node to all other nodes in the graph. This measure is calculated by summing the geodesics connecting the given node to other points and taking this value's inverse:

CLOSENESS CENTRALITY

$$C_C(p_k) = \frac{1}{\sum_{i=1}^{n} d(p_i, p_k)}, \tag{2.7}$$

where $d(p_i, p_k)$ is the geodesic from p_i to p_k. The relative value of closeness centrality can be found simply by multiplying by the value by $(n-1)$.

While measures of betweenness are extremely useful in putting a node into the context of its place relative to the rest of the graph, with the exception of degree centrality, each of these measures is extremely computationally expensive. For instance, in the standard implementations of betweenness centrality, the algorithm requires $O(V^3)$ time, or on the order of V^3 steps to calculate and $O(V^2)$ space during these operations. This means that for a graph of 100,000 nodes, one would need at least a gigabyte of memory and about 100 days on todays fastest computers.

This is a severe limitation for many of the large graphs that are emerging today. To remedy the situation, some have attempted to cut the running time down by optimizing the algorithm. Brandes (2001) has been able to reduce the

running time to $O(VE)$ for V nodes and E edges, a significant improvement over the old algorithm. In addition, his implementation only requires $O(V + E)$ space, which makes the problem much more tractable for data in the thousands of nodes. However, for the cases stated above, assuming a relatively sparse graph, one should still expect computation time on the order of days or weeks.

PAGERANK

Other measures of centrality have been developed outside of the domain of SNA that have similarities to those mentioned above. Most notably, the algorithm behind Google's ranking, ***PageRank***, is a measure of flow across a network (Brin and Page, 1998); given a random walk on a graph, the PageRank for a node is equivalent to the probability that the walker is at that node at any given point in time. PageRank differs from closeness and betweenness centrality in that it does not deal in ***shortest paths***, but rather across all possible paths; this discrepancy is important to observe, but for large networks this is an adequate measure of centrality that is calculable.

STRUCTURAL HOLES

Structural holes, as described by Burt (1992), measure the extent to which an individual bridges various groups, or controls the communication between these groups. He defines this value via another measure, ***structural equivalence*** which he developed earlier as the overlap between two individuals' networks (Burt, 1987):

$$SE_{ij} = \frac{\sum_{k=1}^{n} o_{ij}(k)}{n}; i \neq j \neq k \tag{2.8}$$

Where i and j are two members of the network, n is the number of nodes and $o_{ij}(k)$ is 1 if both i and j have ties to k and 0 otherwise. Structural holes arise when an individual has little overlap with another person, or in other words, low structural equivalence. Because we are interested in individual measures, we can define the structural "holiness" of a person as the inverse of their structural equivalence with all other actors (Burt, 1992):

$$H_i = \frac{1}{\sum_{j=1}^{n} SE_{ij}}; i \neq j \tag{2.9}$$

Since the SE_{ij} values are already normalized, and this measure is over the entire network, H_i will be comparable across nodes. Like betweenness, H_i should be a measure of how much an individual controls the information

spread to the entire network. But as with betweenness centrality, the measurement of structural holes is $O(V^3)$ running time.

CONNECTIVITY

A final measure of a graph is the extent to which it is connected. Many of the methods described assume a connected graph, or one for which every node can be reached from some other node in the graph. If some nodes are not connected, typically the largest connected subgraph, or the *largest connected component* is used for analysis. In an undirected graph, sometimes certain nodes will be accessible from one node but not from another; if all nodes must be connected, then the subgraph with only edges connected to all other nodes, or the *strongly connected component* can be used.

LARGEST CONNECTED COMPONENT

STRONGLY CONNECTED COMPONENT

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 approaches described by SNA.

WATTS-STROGATZ SMALL WORLDS

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 (Watts and Strogatz, 1998).

The Watts-Strogatz model for generating small-worlds networks starts with a regularly-ordered ring lattice which has a high characteristic path length and high clustering coefficient. Taking a vertex and the edge that connects it to its nearest neighbor in a clockwise sense, with some probability ρ the edge is deleted and rewired to another randomly chosen node in the network. As ρ approaches 1, the graph starts to take on properties of a random graph. The

Watts-Strogatz small world network has a value of ρ in between 0 and 1, where the graph has a high clustering coefficient and low characteristic path length.

SCALE-FREE NETWORKS

Other researchers have focued 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) k-α 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).

The condition for preferential attachment as described by Barabási assumes that for a new node coming into the graph, the probability that this node will make a link to vertex i is dependent on the connectivity k_i as defined by $\Pi(k_i) = k_i / \sum_j k_j$. In order to obtain a power-law distribution for the degree distribution, two further conditions must be satisfied: first, the network must continue to grow linearly at each time step, and second, the links must not disappear Barabási and Albert (1999). These two conditions have made many criticize preferential attachment as a comprehensive generative model for all scaling.

2.2 COMPUTER MEDIATED COMMUNICATION

A considerable amount of this thesis is dedicated to understanding the way that internet communication media affect the social relationships of users. This research sits at the intersection of sociology and computer communications, but unfortunately no community of research exists that directly addresses this overlap. Instead, most work resides in the field of *Computer Mediated Communication* (CMC), which also includes psychological, social-psychological, linguistic and behavioral aspects of computer communication. Broadly defined, CMC is the study of "any form of communication between two or more individual people who interact and influence each other via separate computers through the internet or other

network" (Wikipedia, 2005a). A similar and related field is that of Computer Supported Collaborative Work (CSCW) which focuses more directly on making CMC more efficient within the workplace.

The sociological subset of this work covers a wide range of topics and involves a substantial amount of debate. This line of research is extremely young, and presents a number of methodological challenges, but at the same time can provide rich data for the analysis of social networks (Garton, Haythornthwaite, and Wellman, 1999; Wellman, 2001). I will break down the prior work into a few areas based on the foci of attention within the community: first, online social relations as those relationships formed online and that would not have existed otherwise; second, the extent to which online communication supports and engenders offline relationships; finally, the overall effect that the internet has on social capital.

Online Relationships

Early research on the social nature of the internet focused heavily on social ties formed online; Rheingold (1994) presented one of the first accounts of a "virtual community" based entirely on these online ties. In his description of the community known as "The Well," Rheingold showed that people without prior contact were coming together around mutual interests and personal interest, providing conversation, information, and social support. As opposed to offline ties, these relationships are often more specialized, revolving around one or a few interests (Wellman and Gulia, 1999b).

The constraints of online communication can have a profound effect on the formation of these relationships: without physical co-presence, a number of social cues are lost, including body movements, facial expressions, and most notably physical appearance (Wynn and Katz, 1997). These early accounts of online ties tend to spin this as liberating, removing the interaction from the physical form, allowing individuals to cultivate virtual identities that differ from their offline persons (Donath, 1997; Turkle, 1995). Online social relations can thus evolve regardless of race, gender, creed or geography (Patton, 1986; Barlow, 1996), or as articulated by Rheingold:

> Because we cannot see one another in cyberspace, gender, age, national origin, and physical appearance are not apparent unless a person wants to make such characteristics public. People whose physical handicaps make it difficult to form new friendships find that virtual communities treat them as they always wanted to be treated--as thinkers and transmitters of ideas and feeling beings, not

> carnal vessels with a certain appearance and way of walking and talking (or not walking and not talking). (Rheingold, 1994)

Theories of social context cues presented the loss of physical presence in another light, showing that purely online relationships tended to be more deviant and impersonal than the offline counterpart (Sproull and Kiesler, 1986). However, online relationships do not stay online forever; Parks and Floyd (1996) observed that with continued social interest, newsgroup users tended to increase the multiplexity (the number of simultaneous communication media) of their communication, and eventually meet face-to-face.

Supporting offline ties

In addition to providing a venue for meeting new people, online communication tools have become another tool used to support our offline relationships including family, friends, and acquaintances. As with previous communication technologies, such as the phone, fax, or mobile phone, the internet can be seen simply as one that displaces previous modes of interaction (DiMaggio, Hargittai, Neuman, and Robinson, 2001). However, Haythornthwaite (2000) has observed that as the strength of a tie increases, the multiplexity of that tie also increases; since the internet provides more tools for communicating, it is natural to expect that it will be utilized for keeping up with friends and family.

In addition to helping maintain long-distance ties, Hampton and Wellman (2003) have shown that online interactions can strengthen local relationships as well. Residents of a wired suburb studied by Hampton showed that those who participated in an online community network tended to be more aware of each other and issues in the community.

Social capital

The overall effect of the internet on our social relationships can be assessed through the lens of social capital; if an individual's overall social capital tends to decrease with internet usage, one can assume an overall negative effect on their lives, and vice versa if an increase in social capital is observed. Early work on virtual communities suggested an overall net-gain could be expected,

as new, online ties could only increase an individual's connectivity to people and resources.

The first serious debate on this topic was raised by the study published by Kraut, Lundmark, Patterson, Kiesler, Mukopadhyay, and Scherlis (1998), reporting the controversial finding of an *anti-social* potential of the internet. In the study, subjects who reported higher rates of internet use were associated with less contact with household family members, declines in their social circles, and increased depression and loneliness. Nie and Erbring (2000) found similar results, and while he found that higher internet usage led to more weak ties, these users were also likely to have less communication with friends and family, and a lowered attendance of social events. Putnam (2000) has also isolated the internet within his conception of social capital, claiming that after demographic controls, internet users were no different than non-internet users in terms of civic engagement.

After a deluge of negative research, the concept of social capital as affected by the internet has been revisited a number of times in different contexts, both from Putnam's definition and others. The largest criticism of the initial studies stems from the time they were conducted, and the types of people that participated; the internet has evolved quite a lot over the past few years, and the demographics and experience of people using it are substantially different than those represented in the early studies.

Lin (2001) has shown a positive association between email use and social capital, while Quan-Haase and Wellman (2004) has shown relationships between internet use and community participation. More importantly, this work shows that the association between internet use and social captial is not simple and linear, but involves a number of different variables.

Computational analysis

An area of methodological advancement that has large potential is the computational analysis of social interaction. A large portion of online interaction happens in a form that is either immediately observable, or persistent in a fashion that allows later analysis. Most social CMC research tends to focus on surveys or ethnographic analysis of a sampled population, but by operationalizing the forms of social interaction that are available to computers, researchers can take into account the full context.

The work in this area has typically approached data acquisition from

perspective of the individual; using pre-existing archives or by watching a person over time, large sets of personal interactions can be culled, and structural analysis tools applied to the resulting ego-networks. Since many people keep extensive email archives, these have been a popular source of social data (boyd, danah, 2002; Haythornthwaite, 2000), with in- and out-links being determined by emails received and sent to other individuals. Eagle (2005) has devised a data collection system for mobile phones that records a number of different communication measures including incoming and outgoing calls, SMS messages, and location, but the analysis potential of this apparatus has yet to be fully realized.

The other approach to automatic social analysis comes from the systemic point of view; in the case that we can observe an entire closed system of interaction, we can collect whole-network data for a community. Because the collection time for whole-network data do not scale with survey methods, it is rare to see networks above a few hundred nodes; however, with computer aggregation we can observe many orders of magnitude larger data sets.

Smith (1999) provides an early approach to this methodology, looking holistically at the conversations occurring on Usenet. He has devised a number of different measures of social exchange, a user typography, and global characteristics of the entire system from interactions taking place over a few months. Similarly, a number of projects have attempted to infer social relationships from links on the web at large (Adamic and Adar, 2003; Gibson, Kleinberg, and Raghavan, 1998; Flake, Lawrence, and Lee Giles, 2000). While these data are much further removed from explicit social interaction, they provide perspective on the process of collecting data and allow us to start working on the hurdles posed by the analysis of large data sets.

2.3 DIFFUSION STUDIES

Epidemiology, sociology, communications, consumer research, social psychology and social networks, among others

The history of diffusion research is not an easy one to tell, as it spans many disciplines† while never having reached a level of cohesion to be sustained as a field itself. Most generally stated, the connecting goal of diffusion research would be stated as "the study of how and why abstract entities spread among populations of individuals," where *abstract entities* can be interpreted to mean simply information in the broadest sense, social qualities, such as influence, prestige, or authority, or in an even broader sense, cultural features.

DIFFUSION

Before addressing the history of this topic, I would first like to clarify some of the terminology I will be using. By *diffusion* I am referring to the process by

which an entity spreads through a population. This is very similar to other social group-effects of collective behavior (Meyersohn and Katz, 1957; Park, 1967), such as rioting, swarming, or fashion; what differentiates diffusion from these other behaviors is the focus on the conscious decisions of an individual as opposed to a process of imitation or differentiation. *Contagion* is a subset of diffusion characterized by person-to-person dissemination. This is at contrast with other types of propagation, such as the constant force applied by a broadcast mechanism like the mass media.

CONTAGION

Innovation

The most cohesive and lasting subject matter among diffusion research is the study of innovations. While studying the penetration of hybrid corn into a community of Iowa farmers, Ryan and Gross noticed that as this new type of corn started to be adopted by certain farmers, others remained resistant for some period of time (Ryan and Gross, 1943). As their study continued, they watched the entire community adopt the seed as new adopters influenced the next wave of adoption. Looking at the cumulative adoption pattern over time, they saw the familiar logistic growth described by epidemiologists as a slow-growth epidemic. This form was described by Ryan and Gross with the following function:

$$y(t)\frac{Ne^{Nat}}{N-1+e^{Nat}} \tag{2.10}$$

Where $y(t)$ is the cumulative adoption, N is the total population size, a is the adoption rate, and t is time. This form was subsequently applied to a number of different adoption patterns many of which are summarized in Rogers (1962). Despite the popularity of this idea, data on the topic was very expensive to collect, and as a result only a few specific data sets are available.

One of the more important concepts to be defined within diffusion research is the notion of *adopter categories* as measured by the time of adoption for a particular innovation Rogers (1962). These categories are typically defined by the number of standard deviations an individual is from the mean adoption time; more than a standard deviation early is an *early adopter*, then early majority, late majority, and those one standard deviation behind are *laggard* (Rogers, 1962).

ADOPTER CATEGORIES

It should be evident that the normal distribution describing adopter categories is simply the derivative of the S-Curve of cumulative adoption.

FIGURE 2.1. S-Curve of Cumulative Adoption

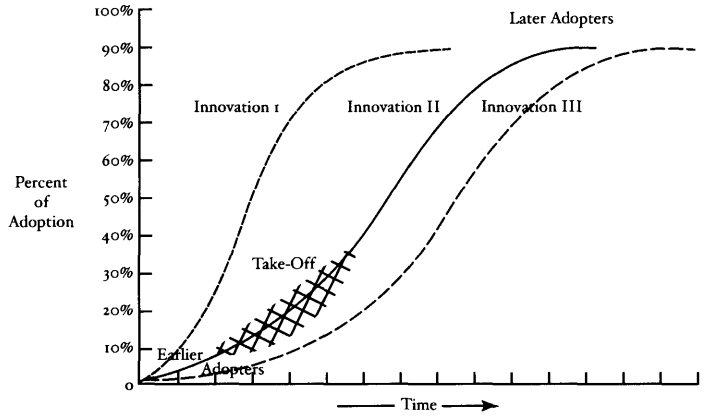
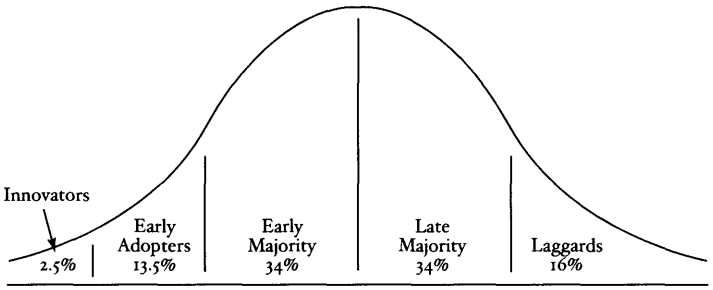


FIGURE 2.2. Adopter Categories



Influence

Some of the earliest observations of the importance of person-to-person communication came from studies examining the effects of mass media in the election of 1940. Researchers Elihu Katz and Paul Lazarsfeld conducted studies to determine the effect of campaign information on the opinions and vote decisions of subjects. What they found was that the radio, television and printed word actually had very little effect on the eventual vote decision of an individual. Instead, they discovered that political influence tended to come from other individuals:

> People tend to vote, it seems, the way their associates vote: wives like husbands, club members with their clubs, works with fellow employees, etc. Furthermore, looked at in this way, the data implied (although they were not completely adequate for this purpose) that

> there were people who exerted a disproportionately great influence on the vote intentions of their fellows (Katz and Lazarsfeld, 1955).

OPINION LEADERS

These highly influential individuals, dubbed *opinion leaders* existed in almost every strata of society and in every occupational area. As it turned out, the mass media did have an effect on the outcome of the vote decisions of individuals, but it was mostly among those who identified themselves as opinion leaders. They named this phenomenon the *Two-step flow of communication* as influence is seen to flow first from the mass media to opinion leaders, and then second from these individuals to the rest of society.

TWO-STEP FLOW OF COMMUNICATION

Katz and Lazarsfeld further divided this process into two specific influence phenomena: the sharing of opinions and attitudes within small groups and the effect of larger communication networks garnered by person-to-person communication. This distinction, between localized, group interactions and connected, external relations is one that is restated in a number of different disciplines.

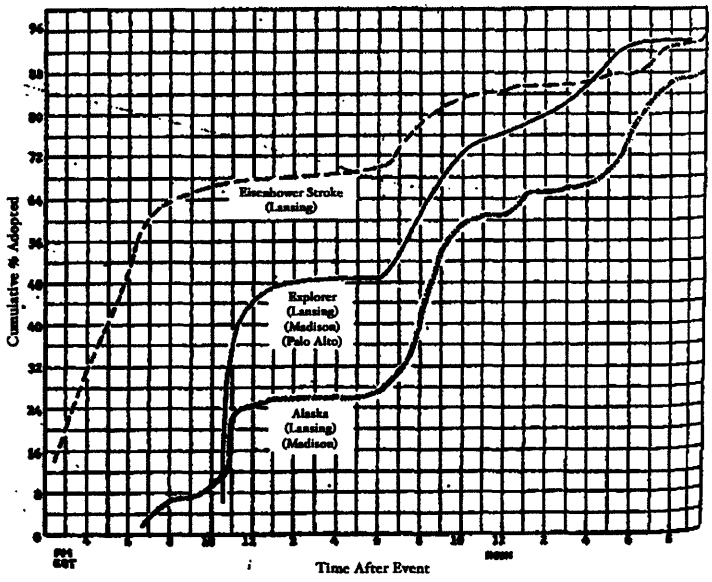
RELATIONAL DIFFUSION

Valente uses the term *relational diffusion* to describe the phenomenon of opinion leadership and group membership outlined by Katz and Lazarsfeld (Valente, 1995). Under this terminology, the diffusion of a particular event is directly related to social contact with another individual. However, these analyses are restricted to models based on status, rank, or membership that come from externally defined sources. Opinion leadership and group membership, while probable factors in diffusion are not made to be explicitly structural in the context of most communications research. Later work will operationalize this terminology with purely structural features (Granovetter, 1978; Burt, 1987; Valente, 1995)

News

Traditionally, when a news story "breaks," the flow of communication is thought to be from reporters to new organizations, and then from these news organizations to the general population through various media outlets of television, newspapers, radio and so on. In the wake of Katz and Lazarsfeld's research on personal influence, the field of journalism became interested in first step of the two-step flow, namely the transmission of mass-media to individuals. While much was known about the process by which news moved from reporters to news agencies, little was known about how individuals in the general population actually received this information.

FIGURE 2.3. Cumulative diffusion for 3 different news stories (Deutschmann and Danielson, 1960)



Paul Deutschmann and Wayne Danielson were the first to explore the realm of news diffusion from a population level. Using random phone interviews around various media events, they asked subjects of their awareness of a particular story, the source they obtained this information from, and the various other media they used after becoming aware. Figure 2.3 shows the diffusion of three news stories during the time of their study. The top curve represents the cumulative diffusion for a story about President Eisenhower's stroke, the second about the Explorer I satellite, and the third about the statehood of Alaska.

While each of these curves has a distinctive shape, a few properties should be observed. First, between the hours of the late night and early morning, very little diffusion about events occurred. The reason each of these three curves move together is dictated by the 10pm and 8am schedules of broadcast and print news services. Also, after the initial diffusion period of about one day, all of these diffusion events tend to grow linearly.

The findings of this study were primarily that most individuals acquired their news from one or more mass media outlets: the newspaper, television broadcasts, or the radio; most of the time the first source was either television or radio, while supplementary information was usually obtained from the next day's newspaper. While they were looking for the effects of opinion

leadership on the actual spread of a news story, instead they found that the primary first source was nearly always mass media; in the case that interpersonal communication was involved, it usually came in the form of supplemental relaying (adding facts to a story that had already broken). Given the discrepancy in personal communication between stories, they stipulated that word of mouth "may be smaller when the story is of lesser value."

This initial study prompted a number of followup investigations into the role of various media in the diffusion of news. Most notably, Bradley Greenberg used similar techniques to explore the specific role of personal communication in news diffusion Greenberg (1964). Greenberg hypothesized a second type of news that would receive considerable word-of-mouth dissemination; in addition to large news stories, he expected to find that "events which go unnoticed by the majority may be deliberately chosen—selectively perceived—by the few because those events have some functional importance."

Greenberg used a much larger sample of news stories, 18 in total, in his sampling. Awareness of these stories ranged from 100% (news of the Kennedy Assassination) down to 14% (a story about a racial disturbance in a nearby school). His hypothesis was confirmed, showing that the Kennedy assassination had a large amount of interpersonal first-sources (50%), down to less than 3% for stories of medium awareness. When the story had less than 33% saturation, the number of interpersonal first-sources rose to over 10%. Mass media was still the most important channel of diffusion, but two distinct classes of personal diffusion had emerged.

Mixed Models

The study of consumer goods led to an interest in modeling the effects seen both by news diffusion and work on innovations. A mixed-model of diffusion was created by combining one element of logistic, internal growth with another derived from external, exponential growth. Table 2.3, taken from Valente (1993), shows the relationship between the various components of the mixed model.

The first column shows the equation derived from Ryan and Gross (1943), which is termed the "internal" component as it relates to the interpersonal effects of diffusion while the second column shows the "external" parameter. The mixed model has two parameters, one derived from the external diffusion, a, and the other from internal effects, b. Using nonlinear regression

TABLE 2.1. Mathematical Models of Information Diffusion

	Internal	External	Mixed
Cumulative Function	$\frac{Ne^{Nat}}{N-1+e^{Nat}}$	$N(1-e^{-at})$	$\frac{N-\frac{a(N-N_0)}{a+bN_0}e^{-(a+bN)t}}{1+\frac{a(N-N_0)}{a+bN_0}e^{-(a+bN)t}}$
Derivative	$a*y(t)[N-y(t)]$	$a*[N-y(t)]$	$[a+b(y(t))][N-y(t)]$
Diffusion of	Adoption	Awareness	Adoption and awareness
Type of Communication	Interpersonal	Mass Media	Interpersonal and Mass Media

Note: N_0 is the number of initial adopters; N is the population size; a and b are model parameters

Gallant (1987), Valente fit this model to data from hybrid corn (Ryan and Gross, 1943), Eisenhower's stroke (Deutschmann and Danielson, 1960), and two subsets of the medical innovation data (Coleman, Katz, and Menzel, 1966). From the derived parameters of a and b, the data clearly supported a mostly-interpersonal growth for hybrid corn, mostly mass media for the news story, and mixed values for both medical innovation samples.

Structural models

The most recent research in this area takes many of the features described above and attempts to explain them in terms of structural features of the network across which diffusion occurs. The standard approach in this paradigm is to take a structural feature, such as personal network density, centrality, or structural equivalence, and correlate this value with the adoption time of an individual. In cases of study comparison, these measures can also be calculated as whole-network measures and compared to attributes of the diffusion.

WEAK TIES AND CENTRALITY

The first structural feature to be studied was the extent to which weak ties played a role in the diffusion process. Based on Granovetter's work on weak ties, Weimann studied the role of weakly-tied outliers in the diffusion of rumors on an Israeli kibbutz Weimann (1982). These results were confirmed by Granovetter (1978), where he posited that marginal individuals are integral to diffusion as they can bridge large components of a network. However, as

Valente (1995) shows, this observation does not seem to hold for any of the other diffusion data.

POSITIONAL EQUIVALENCE

One structural theory of diffusion is that the operational position that an individual holds within a network is related to their likelihood of adopting a given innovation. An example of this phenomenon might be that if a senior doctor decided to start prescribing a new antidepressant, other senior doctors would do likewise. The same would be true for internists as well, but only if other internists were adopting. This idea of roles within the network, as described by Boorman and White (1976) was also tested by Valente and found to be unrelated to the any of the classic network data.

STRUCTURAL EQUIVALENCE

COHESION

Another approach to describing diffusion is that of *cohesion*, where the strength of a tie between two people (for some measure of tie strength) determines the likelihood that innovations will diffuse across across that tie. For instance, if your best friend started using a mobile phone, you might be influenced to buy one as well. This model is slightly more rich than positional equivalence because it provides a relationship between any two people in the network, not just those with the same role.

While this might make intuitive sense, Ronald Burt has provided convincing evidence that in the case of medical innovation, this is not the case. His model of structural equivalence, where influence is related to the overlap between two people's social ties, is opposed to any notion of tie strength. Structural equivalence is similar to positional equivalence in that it relates two individuals on the role they serve in the network, except that it also specifies the relationship between individuals in different roles. An example of structural equivalence would be two individuals in an office who both have the same boss, many of the same professional ties.

The diagram shown in Figure 2.4 depicts a case of contagion through structural equivalence. The red node depicts one actor that has already adopted. Despite the fact that none of the middle blue node's alters have adopted, because she shares the same ties as the first adopter, she will be highly likely to follow.

In the context of diffusion, structural equivalence suggests a sort of competition; when two actors are equal in the network, adoption will occur across their tie for competitive reasons. One important difference from

FIGURE 2.4. Contagion through structural equivalence

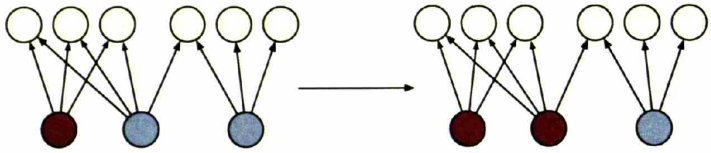
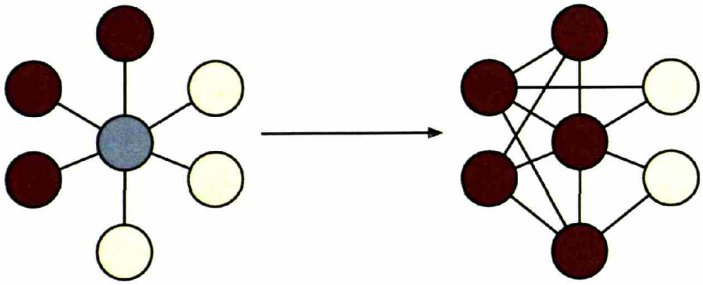


FIGURE 2.5. Contagion through thresholds



cohesion is simply that a tie between two people does not have to exist for them to influence each other's adoption. In a thorough and methodical paper, Burt shows that structural equivalence actually predicts adoption time for network members, while cohesion does not do so to any significant end (Burt, 1987).

THRESHOLD MODELS

A structural model that follows the adoption theory put forth by Rogers (1962) is based on the notion that everyone has a different threshold for adopting a particular innovation, and further that this threshold is related to the percentage of one's alters that have already adopted. For instance, if I have a high threshold to adopting a cell phone, it might take all of my friends and family having one before I finally decide to do so.

Figure 2.5 is a graphical depiction of contagion through thresholds. The blue node is the actor in question, red nodes having already adopted, and yellow yet to adopt. In the first frame, 50% of the actor's network has adopted, and assuming that the actor has a threshold of 0.55 for the given innovation, adding one more alter crosses this threshold, and the actor adopts.

Valente (1995) proposed this model and determined the likelihood that diffusion could be explained by thresholds in the cases of farming practices, medical innovation, and korean family planning. He found that in the case of

family planning, thresholds were likely the cause of diffusion, less so for farming practices, and negligible in the case of medical diffusion.

Scale-Free Diffusion

Following the recent 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 a technical debate because α is 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 (Dezso and Barabási, 2002)

2.4 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. 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. This section will serve as a primer for the various terms used to describe weblogs, and also a review of the literature that has been written about them to date.

Definition

Weblogs have been defined and redefined a number of times in their history, mostly by the authors themselves. One of the first definitional writings on the term comes from Cameron Barrett's weblog *Camworld*, where in 1998 he wrote the piece "Anatomy of a Weblog," which described the type of website that he himself was writing:

> A few months back, I heard the term weblog for the first time. I'm not sure who coined it or where it came from, so I can't properly credit it. Typically, a weblog is a small web site, usually maintained by one person that is updated on a regular basis and has a high concentration of repeat visitors. Weblogs often are highly focused around a singular subject, an underlying theme or unifying concept. (Barrett, 1998)

The term caught on, and a small handful of people began addressing each other as "webloggers." Rebecca Blood provided an early account of the growth of this medium in her post "Weblogs: A history and perspective." She shifted the focus from an author-centric publishing metaphor to one that was much more dialogic:

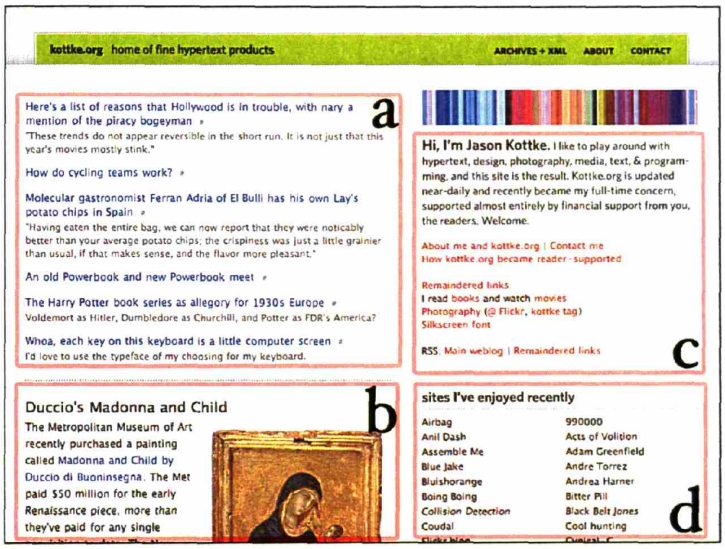
> While weblogs had always included a mix of links, commentary, and personal notes, in the post-Blogger explosion increasing numbers of weblogs eschewed this focus on the web-at-large in favor of a sort of short-form journal. These blogs, often updated several times a day, were instead a record of the blogger's thoughts: something noticed on the way to work, notes about the weekend, a quick reflection on some subject or another. (Blood, 2000)

Both of these writings came at a time when the number of people engaging in the medium were still a very identifiable group. The medium continued to spread exponentially, and after the September 11th attacks on America, a number of political pundits adopted the form and gained considerable attention. Blogs have been legitimated by continued press, and even the mass media's use of the term to describe some of their publications. The number and range of weblogs has even prompted the Oxford English Dictionary to define the term:

> web•log [noun]: a Web site on which an individual or a group of users produces an ongoing narrative. ORIGIN 1990s: from *web* in the sense [World Wide Web] and *log* in the sense [record of incidents]

What, exactly, a blog is has become an intense debate among academics, journalists, politicians and webloggers themselves, all with their own agenda in interpreting the term. I have presented this description to make the point that I will not be attempting to specify what this word means generally, for that I defer to the most unbiased and updated description I know of, the Wikipedia entry on the term (Wikipedia, 2005b). However, for the purpose of this thesis I will need to describe a working terminology that covers a large subset of these websites. It is by no means meant to be comprehensive, but these features will be important to understanding my analyses.

FIGURE 2.6. Weblog anatomy: (a) links; (b) posts; (c) meta-information; (d) blogroll



Anatomy

Weblogs come in a million shapes and colors, and picking any one to use as an example will tend to leave some popular features out. The design and structure of weblogs does tend to follow an aesthetic form that has been copied and iterated upon, so most weblogs have some subset of a common set of features. Important to my analysis are four major components, as shown in Figure 2.6. The weblogs I have chosen to describe is Jason Kottke's weblog, Kottke.org which displays the features I would like to highlight.

LINKS

The section labeled (a) in the figure is an example of the richness of hypertext links in weblog content. As described by early definitions (Barrett, 1998; Blood, 2000), weblogs have roots in public bookmark lists where authors pointed their audiences to things they were currently reading. As this form merged with a more journalistic form, many links were made to enrich the content, either by contextualizing its meaning or referencing other weblogs or writings.

POSTS

The primary form of authorship within the weblog world is the "post," shown as (b) in the diagram. These discrete units of writing are usually displayed in

an anti-chronologically sorted list (newest posts on top). Posts can be short or long, personal or impersonal, and cover just about any topic imaginable. The features that do define posts though is that they are *dynamic* and *temporal.* A weblog has been "updated" which this primary content changes. Some weblogs are updated very frequently while others go long spans of time without any changes, often when the author is distracted with other things (such as a Ph.D. student writing their dissertation).

COMMENTS

Not shown in the diagram, but part of nearly every weblog is a form of reader feedback known as "comments." Usually as a part of every post there is a form that allows the reader to respond to the author in a public manner. In some communities, comments are a regular part of the medium, and in others that are more author-centric, instead of posting a comment other webloggers will respond on their own site instead. To facilitate this type of distributed conversation, some weblog tools have a sort of automatic-commenting system called Trackback or Pingback (cite) that serves to tell one weblog that another has written about it.

META-INFORMATION

Along with the content of a blog, authors typically provide some amount of personal information † which may include their email address, a biography, other websites they maintain, and alternative ways of interfacing with their weblog.

This is not always the case, though; some authors choose to remain anonymous.

An important evolution in the form of the weblog was the separation of the writing from its aesthetic form on the web. Most weblogs now offer a computer-readable form called RSS (for Really Simple Syndication or Rich Site Summary) which allows readers to download the content in programs that resemble an email client. For many readers and authors this has made the process of keeping up to date with weblogs a trivial task; instead of having to check weblogs to see if they have been updated, RSS clients do this regularly and notify the user when content has been changed.

BLOGROLLS AND PERMALINKS

Early in the development of weblogging, there were no central directories, indexes, or search engines that allowed them to find each other. To help navigate this social space, bloggers started to list the other weblogs that they read on the side of their weblog, as shown in area (d) of the diagram. This list was named the *blogroll*, and soon became a standard part of most weblogs.

BLOGROLL

There was a time when weblogs were extremely temporal, and the only way to reference something another weblogger said was to quote their text on your own weblog. Otherwise there would be no concurrency in the interaction. Some infrastructure was developed to solve this problem, allowing authors to link to specific posts within each others' sites. These links were meant to persist even after the content was updated and no longer available at the top of the site. Because of their enduring nature, these became known as *permalinks*; they are now an integral part of the weblogging form, enabling authors to have a coherent dialog over time.

PERMALINKS

These two types of links, *blogrolls* and *permalinks* are the social currency of the weblog community. They are distinct in a number of ways, most notably in their longevity. Blogroll links are updated infrequently and represent a static form of affiliation; in some senses they can be seen as either a proxy to real readership, and in others as a sort of badge of affiliation. Permalinks, despite their name, are actually quite transient. In the moment they connect two weblogs thematically or conversationally, but as time progresses these references are replaced with new posts and new links.

Because of this important distinction, the social meaning of both links is entirely different. Blogroll links, or *static links*, represent some sort of affiliation, while links occurring inline with the content of a weblog (i.e. permalinks), or *dynamic links* simply mean "I read this post" in the most general form. For conceptual clarification, I will use the terms *static* and *dynamic* links in the rest of this text.

STATIC LINKS

DYNAMIC LINKS

Ping Servers

The accuracy of timing is extremely important to the task of modeling diffusion; higher accuracy on the timing of events allows for a better recreation of the series of events. To this extent, our goal is to know when a particular weblog was updated as closely to the actual time as we possibly can. Given that we have no a priori information about weblog updates, we can still find the update frequency of a given weblog by sampling it over a period of time. A number of signal processing techniques can be used to isolate this characteristic frequency with a minimal amount of sampling. Thankfully, none of this is necessary because in most cases weblogs will *tell us* when they have changed their content.

Towards the end of 1999, it became apparent that as a distributed set of tools, weblogs would need a way to let various weblog applications know that their

content had changed. The first of such tools was Weblogs.com created by Dave Winer and Userland software (Winer, 2000); these systems, collectively called *ping servers* are contacted automatically by the software used to publish a weblog (referred to as a "ping"). In some cases this ping data is kept private and used to proprietary ends, and in other cases it is republished for anyone to use. By far the most popular public ping service is *blo.gs* (Winstead, 2005) first created in 2001 by Jim Winstead and then acquired by Yahoo! Inc. in 2005.

PING SERVERS

BLO.GS

Research

Despite their young age, weblogs have become the focus of many different academic disciplines, and engaged a lot of public discourse around their social and cultural impact. While most of the research is outside the scope of this thesis, a few research projects have focused on very similar topics. Namely, a few models and analyses have been made in the domains of social structure and information diffusion.

SOCIAL STRUCTURE

The nature of weblog social interaction is conducive to study simply because many of the forms of social interaction are made in an explicit manner in a public forum. The idea that these links form a network of readership and social relations has been utilized by a number of different research projects (Marlow, 2002, 2003; Adar, Adamic, Zhang, and Lukose, 2004; Herring, Scheidt, Bonus, and Wright, 2004; Herring, Kouper, Paolillo, Scheidt, Tyworth, and Welsch, 2005). Typically static links are taken to represent a form of readership, while dynamic links imply discourse or interaction around a particular topic (Herring et al., 2005).

In addition to explicit social links, some researchers have used other signs of implicit reference as a sign of readership. Similarity in explicit social ties (structural equivalence), link similarity, textual similarity, and timing in updates were used to infer social relations by Adar et al. (2004) and a mixture of social links and information similarity by Gruhl, Liben-Nowell, Guha, and Tomkins (2004).

Based on a subset of weblogs collected from weblog directories (such as the "weblogs" category on Yahoo!), Kumar and colleagues have looked at the whole-network properties of their sample over a long period of time. They extracted a sample of roughly 20,000 weblogs, and also crawled the archives of these weblogs to obtain a historical archive. The sample that allowed them to

extract these data had prerequisites that cannot be explained away in terms of bias. They found a graph of about 70k edges, with more than ten times this number given the multiplicity of links between some weblogs. They found dense subgraphs that, along with a "bursty" linking behavior, allowed them to extract embedded communities.

Herring et al. (2005) have recently conducted a general analytic survey of the structure of the weblog community using both quantitative and qualitative methods . Using a sample obtained from the blo.gs ping service, four random weblogs were selected, and from those weblogs an ego-network of alters identified. This set of 5,517 weblogs was manually identified and analyzed using standard social network measures. They found a range of different types of social interaction, from one-directional affiliation to repeated, reciprocal referencing between authors. From this sample they concluded that the majority of weblogs are disconnected, while densely connected exist in fewer areas. Their findings suggest that contrary to the bursty nature described by Kumar, Raghavan, Novak, and Tomkins (2003), few weblogs actually engage in regular, reciprocal dialog. This could point to irreproducible results put forth by Kumar, or it could reflect a change in the style of weblogging today; because the community is growing so fast, it is hard to say.

In a less academic setting, a recent debate that has drawn considerable attention among weblog authors. Clay Shirky has written a piece documenting the existence of a power law among static links between authors (Shirky, 2003). Based the work of Barabási, Shirky assumed that this scaling emerged from the expression of preferential attachment; the earlier a weblogger started their site, the more likely they were to have a high in-degree. Because of the structure of Barabási's model, there would be little likelihood that an incoming author would be able to upset this distribution, and over time the "rich get richer."

Further analysis has shown that dynamic links, 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 interrelated 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).

Herring et al. (2005) have looked at this more closely, but their sample consisted of a qualitative sample of 24 weblogs.

Most of these studies of weblog structure have made the assumption that linking and topic similarity are in some way "social," imply "ties," but none have presented a broad analysis of the true meaning of these relations† At this

point we can refer to weblog interconnections as a "readership network," but real social relations need to be empirically confirmed.

INFORMATION DIFFUSION

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. Using the external links as signifiers of incidence of a particular idea, research has shown the diffusion of information around the readership network by tracking these links. Early studies of this process have shown that unlike innovations, the news and stories referred to by links do not produce distinct adopter categories, but rather that adoption for a given weblog is both contextually and topically defined (Marlow, 2002). From the perspective of the media, three distinct categories of diffusion have been found by looking at cumulative adoption patterns: 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 (Marlow, 2003; Adar et al., 2004).

In addition to using links as a proxy to information, Gruhl et al. (2004) have analyzed the diffusion of ideas through common spikes in natural language. They found that weblogs, normalized to their respective time zones, tend to have regular posting patterns (assumedly related to the cycle of authors lives). Based on the flow of information during the day, and the relationship between concurrently used words, they have induced both the network and path of transmission for "topical" news events. They have observed a "fanout" of diffusion, wherein information starts at one node and spreads with decreasing probability of transmission to all connected nodes. While they present such background as threshold models of diffusion (Valente, 1995; Granovetter, 1978) and scale-free epidemic spreading (Pastor-Satorras and Vespignani, 2001), they do not confirm any model parameters for the diffusion, but rather validate their method of data collection.

Chapter 3

Design and Methodology

As stated in the introduction, the primary goal of this thesis is to answer two questions about media contagion among weblogs:

1. What does the social structure of weblog authors look like, and how does it relate to their offline social ties?
2. How can media contagion be described, and to what extent does the social structure play a role in this process?

There are two apparatuses that I have used to answer these questions: first, a *weblog aggregator* that automatically collects information about the social structures of webloggers and the various media contagion events that occur within. Second, a *general social survey* was employed to probe deeper into the motivations and social character of weblog authors. This chapter will outline all of my choices in the methodology and design of these two instruments, along with explicit hypotheses related to the questions listed above. Before I address the design, I should first take a step back and define the sample frame for this thesis.

3.1 SAMPLING WEBLOGS

In choosing a sampling methodology, it is easiest to think in terms of the *target population*, or the population of people one wishes to study and the *frame population*, the method by which members of this population are identified and recruited.

TARGET POPULATION

FRAME POPULATION

As with any study of online populations, specifying a target population for weblogs is not a trivial task. How big is the blogosphere? Should self-contained communities be included in the same sample as those those that are more open? What exactly is a weblog and where should the line be drawn within the expanses of gray area that surround this definition?

Likewise, the construction of a frame population is equally difficult. What do we use to acquire a list of weblogs? How do we differentiate possible non-weblogs from the list we have obtained? This section will serve to answer these questions.

Target Population

As discussed in the Chapter 2, there has been quite a bit of speculation over what exactly defines a weblog. Some individuals look historically and arrive at an extremely narrow definition while others take a more inclusive approach that includes almost the entire web. Given all of the arguments in this heated discussion, the approach I usually take is to let the author decide: if an individual believes they are writing a weblog, then indeed that is a weblog. This said, there are some necessary exclusions from the group.

Large, self-contained weblog communities such as MSN Spaces (Microsoft, 2004), LiveJournal (LiveJournal, 2001) and Xanga (Xanga.com, 2003) are all hosted on a few servers, and the sheer volume of their updates exceeds my rights to crawl them (see Appendix A for a description of constraints on crawling). While LiveJournal does provide alternative means of acquiring their content, it does not include the rich information listed on the front page of these sites. None of these services are obtainable without violating conduct which would probably get my aggregator banned; for this reason I have not included them in my sample.

These three services constitute a large amount of weblogs within the United States, but it has been suggested that much of the social interaction is inward-looking, i.e. within the individual service. I have considered the connections from my sample to these other communities to assess the amount of overlap that they have. Furthermore, none of these bloggers were excluded from the survey portion of the thesis.

In summary, the target population I have chosen is the population of weblog authors, sans those who use one of the services that are outside the constraints of allowed aggregation. I have assumed that this will not have a large impact on the results as these communities are purported to be distinct. The connectivity from my sample into each of these sites is used to determine how much overlap is missed by this decision.

Frame Population

Since there is no global system for tracking the existence of weblogs, I must rely on various sources of weblog data to acquire my study population. There are a number of mitigating factors that help decide which frame population to use; each source comes with an associated *bias*, and I hope to minimize these effects as much as possible. Related to this matter is the of *size*; many sources provide overlapping data, and I wish to choose those that produce the largest possible sample, which in turn helps minimize any selection bias. Finally, there is a matter of *precision*, which affects not only our knowledge of a weblog's existence, but also the frequency and timing of its updates.

The following list shows the four principal means I have identified for weblog acquisition.

DIRECTORIES

There are a number of online directories that allow authors to self-categorize their weblogs into a catalog, very much like the Yellow Pages. For each directory there is an inherent level of self-selection based on the diffusion of these directories into the weblog population. The largest of these directories, Eatonweb Portal (Eaton, 1999), Bloghop (Bloghop, 2000) and the Globe of Blogs (Globe of Blogs, 2001), only contain tens of thousands of weblogs each† , which is many orders of magnitude smaller than the expected population size. Additionally, directories tend to be out of date and biased heavily towards those individuals who both know about the directory and have chosen to join.

The largest of these directories, Bloghop, contains just under 30,000 weblogs.

SPIDERING

The most inclusive approach to obtaining a sample of weblogs is to start with a small set of known weblogs and progressively spider the web in the same way search engines obtain their indexes (McBryan, 1994). This method requires us to have some way of automatically discerning weblogs from other web pages, typically using heuristic or statistical approaches. The BlogCensus project used this approach with a heuristic derived from common weblog features to identify weblogs (Ceglowski, 2002). While the population derived by spidering is by far the most complete, it incorporates the biases of the algorithm used to identify weblogs, can take months to acquire a sample, and can include a number of dead or unused weblogs.

APPLICATIONS

While weblogs are created by a number of applications, the most popular tools account for an overwhelming majority of the weblogs (Perseus Development, 2004). Because these services maintain statistics about their users, they can be a good source of population information. Nearly all of these provide the size of their user base, the number of weblogs created, and some provide some basic demographics data.

Choosing to focus on individual applications has the advantage that the total sample population is specified and there are typically easy measures for gathering contact information for each individual weblog author (such as demographics or email address). The disadvantages are that the resulting data contains the biases inherent to each application, and while the population may be well specified, it may be impossible to actually gather all of the weblogs. As noted in Chapter 2, the only broad-based weblog surveys focused on individual weblog applications (Perseus Development, 2004, 2005).

PING SERVERS

The final method for obtaining a sample is to look at those sites that identify themselves through the use of ping servers. The ping server has become the standard method for weblogs to make other systems aware of their existence. This approach has the advantage that many weblog authoring tools are set to use ping services by default and as such provide a similar census to the previous approach, but across many applications. The most important difference between ping servers and all other methods of acquisition is that ping servers provide accurate timing information for when weblogs have been updated. With all other methods, the weblogs must be polled regularly to determine when they have been changed, which will seldom be as accurate as the push-method employed by pinging. Additionally, ping servers guarantee a set of *recent* weblogs, and remove all dead or otherwise unused sites that are part of other censuses. The largest disadvantage is again that the census is biased towards those sites that choose to use a given ping service.

Choosing a source

Given the possible options for arriving at our population of weblogs, the most comprehensive method is spidering, while the most accurate are ping servers. With enough resources and time, a comparison between the two methods would be in order, but with estimates of the size of the hosted weblog world

(Perseus Development, 2005), it is beyond the scope of this thesis to attempt such a task.

There are a number of ping servers that exist, but the largest by far is *blo.gs*. Because blo.gs syndicates pings from most other public services, it simplifies my task of weblog acquisition considerably. For the course of my study, I use all weblogs that have pinged blo.gs in my frame population.

3.2 WEBLOG AGGREGATOR

Because weblog content is in a public, computer-accessible environment, I have the ability to track and analyze the behaviors of many authors. Instead of studying some subset of the population, I can let a computer (or a few computers) aggregate the population described above. Whenever a weblog is updated, a system can fetch that weblog, store its contents to disk, then analyze and index various features that may be of interest.

I have constructed a weblog aggregation system to continuously monitor updates to weblogs and collect data about their behaviors over time. The architecture of this system is provided in Appendix A. For the rest of this section I will assume the following data: a set of weblogs, their links to other weblogs, and their links to other websites, all tagged with observation times. A number of different measures are calculated for these data, both from the perspective of the weblog and from the perspective of outside sites that are diffusing.

The important details of this system are simply that the content of a weblog is fetched and stored with the time that the system determined it had been updated. From this content, all of the external links are extracted, and indexed according to their type: links made to other weblogs in the sample are stored as social links, while everything else is stored as a diffusion event. The social links are further distinguished as either *dynamic*, or linking to specific content on the referenced blog, or *static*, linking to the front page.

The weblogs that are obtained from blo.gs are not constrained to America or even English-speaking authors. Any number of languages may be used in the writing of the aggregated sites, and this should not affect the structural analysis or modeling of media contagion; however, the survey was conducted in English, so I needed to provide some facility for selecting English blogs. As described in Appendix A, the language identification system described by Ceglowski (2005) has been implemented. Whenever new weblogs are

detected, this detection is performed, and when a match is found, the language is stored with the weblog. These data are also useful in characterizing the sample I have obtained from blo.gs.

In the following two sections I will discuss how these data, namely links on weblogs index by time, are converted into a readership network and diffusion events, and the analyses used to model these data.

Readership Network

When I first approached these links between weblogs, I referred to them as the "weblog social network," but upon further consideration I realized that without asking the bloggers what exactly these links mean, the fact that they are social is just an assumption. For this reason, I refer to them as the *readership network.*

The first stage of analysis involves a structural analysis of the readership network using the various measures provided by SNA and work in emergent networks. As noted in Chapter 2, these social links come in two varieties: static (links made to the front page of a weblog) and dynamic (links made to specific posts). These networks are analyzed separately due to the differing motivations that produce them.

As with any high-variability web data, the *true* first step in data analysis is removing spam. Because blo.gs is an open service, there is no barrier for authors of spam sites to ping the service as if they were a weblog. In some cases, these rogues also use blog software, and even construct content in a form that appears to be blog posts. Unfortunately for them, their behavior is irregular in a number of observable ways that allows for their removal. Spikes in the measures of in- and out-degree are usually the most obvious signs of funny business, and I have used them to remove large amounts of spam.

The following measures of connectivity, degree, density and path length are all meant to provide a more detailed picture of the weblog community. It would be extremely useful to simply visualize these data, but that problem is not tractable on todays computers, or even those 5 years from now. Sampling this network would remove the overall context, and looking only at individual nodes would take forever, but we can get a similar feel simply by looking at various calculable measures.

CONNECTIVITY

After the data set has been cleaned of all obvious abnormalities, the next step is to convert it into a data set that is amenable to most network measures. Two induced subgraphs, forming the largest connected component and the largest strongly connected components (Cormen, 2001) must be calculated. My first hypothesis is that *both* networks are almost entirely connected:

> HYPOTHESIS 1: Both the static and dynamic readership networks will have a negligible number of nodes not connected to the largest connected component, and most of these outliers will arise from sampling error.

By sampling error, I explicitly mean that either the sites are *not* weblogs, or that they are part of a group that is not likely to use the blo.gs service (such as particular foreign languages, services, or cases on the edge of the blog definition). These errors reflect both noise in the data and the error introduced by the selection of my frame population.

DEGREE

The next piece of the puzzle comes from the distribution of the degree across the set of actors; one should expect that the distribution of attention across the weblog network will not be uniform. Both networks, both static and dynamic, are composed of the same nodes, but with different edges connecting them. If these two networks are generated by the same underlying mechanism, then I anticipate a strong similarity between the two.

> HYPOTHESIS 2: The static and dynamic readership networks will exhibit different structures, revealing different constructive mechanisms.

The degree distributions are an important piece of the understanding a large network such as the one I expect to observe. If we assume that in-degree implies authority, then Hypothesis 2 would suggest that the basis for authority in a continually moving world of attention (dynamic) is different than one that accrues much more slowly (static). To explore this phenomenon, I compared the top ranked weblogs in each network, looking for overlap and qualitative differences; if one network is dependent upon another, it should be evident in the top weblogs.

In a random network, one would expect degree to be described by a Poisson distribution (Bollobás, 1985), but recent work with large social networks such as this have suggested that degree is much more variable, and in many cases follows a power law (Barabási and Albert, 1999; Lilijeros, Edling, Amaral, Stanley, and Åberg, 2001). Previous work on weblogs has also revealed power-law distributions for both in- and out-degree (Marlow, 2003; Kumar et al., 2003; Adar et al., 2004; Gruhl et al., 2004), and a similar result here should not be surprising in any way. While some have speculated that the generative model that gives rise to these networks is preferential attachment (Shirky, 2003), I anticipate a different underlying cause related to features other than the age of the weblog.

> HYPOTHESIS 3: The power laws observed in both readership networks will reflect a generative model *other* than preferential attachment.

There are a number of features that scaling might be contingent upon: frequency of posting, quality of posting, connections outside the network, and any number of demographic variables. In this part of the analysis I have looked to see if there is any relation between other observed variables (update frequency in particular).

DENSITY AND PATH LENGTH

The final descriptive measures of the weblog networks are those observed by (Watts and Strogatz, 1998), namely the personal network density, characteristic path length, and clustering coefficients.

> HYPOTHESIS 4: Both of the weblog networks will exhibit small-worlds properties, namely having a short characteristic path length and high clustering coefficient

In addition to having power-law distributed degree distributions, many large social networks have shown what Watts and Strogatz (1998) observed: densely knit local clusters still connected to all other nodes by a very short path. This "small-worlds" network is characterized by a high clustering coefficient and a short characteristic path length. Because of the previously observed degree distributions, a majority of the network has a very low in-degree while few have high in-degree. One of two scenarios should emerge at the local level: either dense clusters or near-disconnected isolates, the latter of which seems less likely given the social nature of the medium.

GENERATIVE MODELS

Given the results of the preceding network measures, I have attempted to construct a model that best fits the structure that I have observed. Because the time window on my data, it is not possible to consider the dynamics of this network, but the model should take into account the current state, how it could have arrived at these various properties. Because the model was not immediately evident, I will discuss the issues surrounding model construction in Chapter 4

Diffusion

In addition to the links that confer social structure, external links are also signifiers of the diffusion of information. Because of the propensity of bloggers to contextualize their writing with links, and the unambiguous nature of the URL, links to various websites represent an accurate measure of media contagion.

Weblogs are constantly being exposed to a mix of external and internal forces of media contagion. In the former, the mass media and other broadcast information is constantly being consumed by a large percentage of this group, and at any time this content might cause a given blogger to write something. Likewise, bloggers are reading each other, and given the distilled nature of weblog content and specialization introduced by each person, the amount of relevant content should be higher. For this reason I expect weblog authors are often linking to the things around them.

To study the diffusion, I use the following representation: each incidence of a diffusion event is stored as a source, destination, and time, with the source being the weblog, and the destination the content they are linking to. To reconstruct a diffusion event, I take a given destination and find all of the sources. Sorting by the date, I have a list of the specified weblogs in the order the links were observed. From here the cumulative adoption or various structural features can be reconstructed.

As with the readership network, a certain amount of data refinement is necessary before it can be analyzed. First, I am interested only in the links I know have been posted within the timeframe of my study. The first time a weblog is crawled, it contains any number of links that may or may not have been posted during that update; to gather only "new" links, I consider only those links posted after the first sighting. Second, the number of examples

over the course of the month was on the order of millions, which presented an overwhelming amount of data. To make the analysis tractable, I used those links that had been found on 10 or more weblogs.

CUMULATIVE ADOPTION

Most of the analysis of diffusion uses the cumulative adoption curve to assess the quality of a model. If the model produces a curve similar to the observed adoption pattern for the various model parameters, then the model is assumed to be predictive. The first step in analyzing the diffusion data is to take the cumulative adoption curves and see whether or not there is any regularity among them; Adar et al. (2004) used K-Means clustering as an initial approximation, which I have done as well.

> HYPOTHESIS 5: Diffusion events will fit the mixed model described by Valente (1993) with some events being external, some internal, and most a mixture of the two.

I have also fit each of the diffusion examples to the mixed-model using a nonlinear, least-squares regression to obtain the model parameters for a (exponential, external diffusion) and b (logistic, internal diffusion). To assess the quality and interpretation of these fits, I have performed a qualitative examination based on the distribution that arises.

This is as far as the analysis can be taken without starting to look at the structure involved in the diffusion, which is the next stage of evaluation.

STRUCTURAL APPROACHES

A number of different structural approaches have been used to model innovations; among the more successful, Burt (1987) used structural equivalence, showing major wins over the previous model of cohesion and Valente (1995) employed individual thresholds of adoption. In both cases, a local network parameter (structural equivalence or an individual's threshold) was used to determine the point of adoption for each individual over the course of the diffusion. In the case of weblogs, I am not sure that threshold models are applicable, and structural equivalence, while potentially useful, is an intractable measure (see Chapter 2).

In the case of most innovations, the resulting adopting population is near saturation, implying that nearly everyone eventually chooses to adopt. In the case of my data though, each example spreads to a small subset of a very large graph; we can think of each media event as one of these fully-saturated

communities, and the entire network as the superset of a number of these different groups.

I began by observing how many individual links can possibly be explained by contagion. I looked at the set of weblogs in a given event in order and for each weblog checked if it was connected to any of others that have already made a link; I will term this *perceived contagion.* In the case of threshold models, nearly 100% of the adoption should be explained in this way. While structural adoption does not require a direct tie, the chances are very low that two actors with high structural equivalence will not know each other. If either of these two models explain weblog diffusion, the data must also have a high level of *perceived contagion.*

PERCEIVED CONTAGION

Because the results from the cumulative adoption analysis have shown that some events are not explained by internal growth, and for these there should be little or no perceived contagion.

> HYPOTHESIS 6: The mixed-model parameters should be highly correlated with perceived contagion.

If the mixed-model is applicable to these data, the parameter b should be related to the most basic form of contagion; likewise, high values of a should be associated with less connection between the given weblogs.

ADOPTER CATEGORIES

The final area of media event analysis relates the entire set of events to the individual weblogs who have linked to them. A given weblog can link to a given site at different points during its diffusion; the time at which a weblog posts the link can be seen as a sign of influence, that their link determines the spread of the idea, or awareness, that their link happens to be first. Unlike innovations, the term "early adopter" doesn't make intuitive sense, as linking to a site is more determined by how soon you found it than your willingness to write about it.

> HYPOTHESIS 7: Certain weblogs will emerge as *media leaders* and this status will be determined by in-degree and centrality.

The concept that certain people will have a higher propensity to adopt innovations is related more to the idea that some individuals will have better access to information than others. If they continually come across ideas earlier than the majority, this should be evident by their linking patterns over time. To construct the measure of timeliness, I used the first and second

deviations within the entire diffusion event, corresponding to the "innovator" and "early adopter" categories in Rogers (1962). Each time a weblog links in a timely fashion, I have added one point, and the sum of their points across the entire sample period determines their *media leadership*.

MEDIA LEADERSHIP

Given the literature on adoption, it is likely that those individuals with structural advantages, namely centrality, will have quicker access to information than those on the margins.

3.3 SURVEY

In the process of studying the weblog aggregation data, much of the analysis is based in inferring social behaviors from data without addressing the authors directly. To gather more information about the uses and motivations of weblog authors, I have employed a general social survey. The survey consisted of five sections: demographics, linking behavior, weblog use, general communication use, and social capital, each of which are described in detail. A copy of the survey can be found in appendix C.

Web surveys

In comparison to phone or mail surveys, web surveys have been both criticized and applauded as a method of data collection. On one hand, they have the *potential* to reach every person who has access to the internet; on the other, the sheer volume of web surveys makes it hard for subjects to discern which ones are worthwhile, and which ones are not, leading to the eventual side-effects of over-surveying (Couper, 2000). I will outline some of the potential pitfalls of this new methodology before describing the survey design.

COVERAGE

After the sample has been selected, the coverage must be justified by showing its representivity of the target population. The most common method of handling this question is by comparing some known demographic quantity within the target population to the set of subjects obtained by a given frame. As Couper (2000) points out, demographics do not tell the entire story, and a better question is "whether the two populations are similar on the substantive variables of interest." In many cases, demographics can mislead us into thinking that we are covering our bases when in fact we may be selecting for a

specific psychological behavior (more outgoing, more communicative, more likely to take surveys, etc.).

NONRESPONSE

Even after a survey has been engineered to take care of all of the coverage issues, error or bias can arise from those individuals who decide not to complete the survey for a reason that differentiates them from those who do. In some cases, nonresponse can also arise from subjects who start the survey and decide to quit before finishing. The most common problem with nonresponse in web surveys actually stems from the undefined nature of the frame; for this reason, most online surveys do not even identify the nonresponse bias. To combat this issue, some surveys have combined email, phone and web surveys of the same population to quantify this discrepancy (Dillman, 2000).

Minimizing nonresponse in mail surveys has come down to a number of different tactics that have been shown to motivate potential subjects: official-looking letterhead, personal messages, follow-up contacts, and even the inclusion of a pen with mail surveys (Dillman, 1978). While this literature is rich and heavily debated for phone and mail surveys, the same amount of work has not been addressed for their online counterpart. Couper (2000) suggests that as these issues are secondary to those of coverage, nonresponse will likely become a more researched topic in the near future. However, it cannot hurt to replicate the features of mail surveys, especially the official look and personal attention.

INCENTIVES

The effect of incentives in online surveys has been inconclusive. Bosnjak and Tuten (2003) have even seen a decrease in response rate when incentives are involved, especially when the incentive is given before the survey takes place. It appears that for the case of web surveys, perhaps due to oversurveying, or spam in disguise as a survey, people are no more likely to choose a survey because of the reward. For this reason, the survey was non-incetivised, and I expected a response rate around 30% (Bosnjak and Tuten, 2003; Dillman, 2000).

Survey Sample

The survey was piloted with a small group of about 20 weblogs who could be trusted to keep the survey secret. My fear was having the pilot spread virally beyond this group and losing my opportunity to have an unbiased sample for the actual survey. After tuning questions based on both qualitative and quantitative responses to the initial survey, it was targeted at two sample populations. The first was a *random sample* of the population collected through the aggregation process. In order to notify these bloggers of their involvement, I could only consider those sites that were in English, and had an email on their weblog.

RANDOM SAMPLE

Assuming about a million weblogs, 75% in english, and 10% having associated emails† , we can expect a potential pool of about 75,000 subjects (assuming the two constraints overlap normally). With a target population of 750,000, confidence level of 95% and interval of 3%, our sample size should be at least 1,066 subjects to achieve the desired representivity. However, this assumes a 100% response rate; given the 20-30% response suggested by Kypri and Gallagher (2003), these figures imply a sample of 5,330 subjects. For simplicity's sake, I have decided to use 5,000 and with an expected response rate of over 21.3%.

These statistics represent preliminary projections from the beginning of the aggregation period.

Each of the 5,000 email addresses were inspected to determine whether or not it appeared to be legitimate. In the cases where multiple email addresses were found on one page, I chose either the apparently related address (i.e. had a similar name to the URL of the weblog), or the one that appeared first.

Internet marketers spend quite a bit of energy determining which days are the best days to elicit a response from people via email. While many speculate that Tuesday is the best, the verdict is still out (EROI Inc., 2004; ExactTarget Inc., 2004). However, there is consensus that the weekend is a time when many people are not at their computers. I chose Monday afternoon for the mass emailing. A sample copy of the email has been attached in Appendix **??**.

The survey was also be open to the public, in addition to the random sample, so my second group came from a *self-selected sample* of subjects who found the survey through various channels. The method I chose to spread the word was to diffuse the survey as a media event: starting with my weblog, I announced the survey. I then successively contacted my friends with weblogs and asked them to spread the word. Finally, when that did not provide all of the respondents I was looking for, I created a badge of completion that allowed subjects to tell others that they have taken the survey.

SELF-SELECTED SAMPLE

Technology

Instead of working with a standard form-based web survey apparatus, I decided to use a technique known as *AJAX*, or "asynchronous JavaScript and XML" (Garrett, 2005). AJAX is a combination of technologies that allows for more natural, live applications to be developed within a web browser, without having to send and load content while the user waits. Many companies have embraced this method over the past few months, and it is quickly becoming a standard for web application design. As with any web technology, AJAX has the drawback is that it is not ubiquitous within web browsers; however, this bias is expected to be negligible given the saturation of compatible browsers.

AJAX

Because of the "live" element of the survey apparatus, I am able to record answers to individual questions as they are answered. In the standard form-based approach to surveys, answers are only recorded when the user explicitly sends them. In the case of long surveys with low completion rate, entire pages of answers can be lost Dillman (2000). In our case we know not only every answer, but the number of times it was changed, and the period within which these changes happened. This allows us to better profile our subjects use of the apparatus.

Demographics

Because the sex, age and location of weblog authors is an unknown quantity, I decided to put the demographics section at the front of the survey. Although this is usually considered taboo in survey methodology, and a potential source of nonresponse (Dillman, 1978), I am able to accurately gauge the profile of those subjects that partially completed the survey. Because my random sample was based on the detection of English language, not location, I was unable to ask all standard demographic questions to all respondents. The section is thusly split into two parts, the first was asked of all respondents (age, gender, education and country of residence), and the second was asked only of those individuals who selected the United States as their country of residents (zip code, marital status and race).

> HYPOTHESIS 8: Subjects will be more male, educated and younger than expected from census data

Previous marketing surveys (Perseus Development, 2004, 2005) have shown a slightly more female, younger and well-educated population of weblog

authors. Given that these surveys excluded independent weblogs, which include a large number of political blogs (assumed to be male-dominant), I anticipated the gender bias would shift towards the male.

As mentioned earlier, demographics do not tell the entire story (Couper, 2000). Despite my representivity, or lack thereof, the sample can still be biased by a number of factors, most importantly related to weblog use. Since no other gauge of the general population's weblog behaviors exist at this time, demographics are the best tool I have to justify the sample.

Links

The second section is the least conventional (and least robust) of all of my survey areas. Much of the analysis of our diffusion data depends on assumptions made about the links weblog authors were making, both social and informational. For this reason we wanted to ask each subject *specifically* about links they had made recently. Thanks again to the live quality of the survey application, I am able to gather information from their weblog in real time and ask them questions about individual links.

After subjects submitted the address of their weblog, the content was fetched and links extracted. A set of 5 links were selected from the set of external links made on the weblog at the time of the survey. For each of these links, subjects were asked to classify the link into a number of different social categories (weblog, weblog post, personal homepage) or "other" for any other link. In the case that the parsing incorrectly selected an internal link, "part of my weblog" was added as an additional choice. Subjects were asked subsequent questions about the link based on the type specified.

SOCIAL

When authors specified that the link was made to another *person's* web page, the questions that followed were about the relationship the author had with that person. The goal of this section was to provide some background to what a social link really means in the context of a person's weblog. The first question asked the author to specify the type of relationship the author had with the alter: friend, family, acquaintance, or "don't know them personally." We define *friend* in the same respect as Marsden (1984), "someone you feel especially close to."

The subject was then asked questions about the alter and their weblog: when

they had last read the site, when they had last posted a comment on it, and when (if ever) they had met the author in person.

> HYPOTHESIS 9: A large percentage (> 50%) of social links will reflect weblogs that are no longer read, and most reflecting little or no personal relationship

This hypothesis is based on the expectation that bloggers do not actively update their list of static links, and thus blogrolls contain a number of "dead" social links. Likewise, given the predicted degree distribution, I expected that most of these links by random selection to reflect large weblogs, and given the constraints of a single person's communications, that these would typically be individuals the subject has never met, emailed, or spoken with.

INFORMATIONAL

When authors classified a given the link as non-social, two basic questions were asked: where the author found it, and why they posted it. Possible sources for the link included another person, a search engine, a mass-media news source, general surfing ("stumbled upon it"), and none of these. The author was asked to pick only one reason for posting the link among personal, newsworthy, important, amusing, informative and "no reason." Each of these categories is based on the subjective responses of pilot subjects.

The inclusion of these two questions is an attempt to recreate early news diffusion studies (Greenberg, 1964; Deutschmann and Danielson, 1960). By filtering those links that are potentially contagious ("newsworthy" or "important"), I hoped to see some relationship to the source from which they acquired the information. Because these questions were only be piloted on a few people, I do not posit any firm hypotheses as to the outcome of the response.

Weblog Use

In this section, subjects were asked to detail their experience with weblogs in general, and also about the weblog they currently maintain. The first part of this section is devoted to their general weblog usage which included the length they have been weblogging and four measures of investment: number of weblogs they have authored in the past year, posted comments on in the past year, amount of time they spend using weblogs daily, and number of weblogs they read on a given day.

The second part of the section asked in more detail about their experience with the site they consider their "primary" weblog. These questions can be broken down into two categories: categorical, or those that help determine the type of weblog and usage, the way in which the author, other authors, and readers utilize the weblog. I will address each of these groups separately.

CATEGORICAL

One of the primary objectives of this survey is to characterize the different varieties of weblogs that exist. Much of the design of these questions is based on the following hypothesis:

> HYPOTHESIS 10: Three primary types of weblogs will emerge: professsional, editorial, and personal

Professional blogs are those written in the workplace, about work topics, and probably with the intent to increase one's reputation. Editorial blogs are those that are built on top of regular comments on current events, and often focused on politics. Finally, personal blogs are of the journal style outlined in Chapter 2, and attended to social ends. The first question in this area asked the subject to enumerate all of the primary motivations they have for maintaining their weblog; these include 11 motivations culled from the pilot, along with "none of the above." I performed a factor analysis of the various combinations that subjects chose to come up with a basic weblog typology.

I also asked the subject to approximate the percentage of posts they make about the three topics listed above (professional, editorial, and personal). These three were not meant to be mutually exclusive (i.e. one could have one post that is both personal and editorial). I validated the dimensions resulting from the factor analysis through correlations with these frequencies.

USAGE

The subject was also asked to detail their usage of the weblog, as well as approximate other individuals' use as well. The total number of authors was used to differentiate large, community weblogs from those that are individual, which is an important distinction in other sections. The amount of time they invest was posed, as well as the total number of comments it receives and the estimated audience size.

In addition, the age of the weblog and the number of times it has changed location were also provided in order to more accurately predict population statistics.

HYPOTHESIS 11: The investment a weblog author makes will be closely related with the number of comments and self-reported audience size

One of the more important questions in this thesis is what determines the audience of a given weblog. Certainly age promotes awareness, but I also posit that investment—as determined by a number of different measures—results in more readers. This is based on the observation that the best way to acquire readers is through interaction with other authors, which takes explicit investment on the part of the subject.

Communication Use

The last two parts of the survey explored the subject's general social and communication behaviors. In addition to understanding the data I have collected via aggregation, I wanted to look at the types of people attracted to weblogging, and the effect weblogs had on their behavior. Does having a weblog imply you are a hyper-communicator or have higher social capital? Do personal bloggers have different communication patterns than editorial or professional? Do more communicative webloggers have bigger weblog networks? This section addressed these questions by detailing the quantity and nature of communication among different media: email, phone, text messaging (SMS), and instant messaging (IM).

For each technology, subjects were asked to approximate the number of individuals they communicated with in the past week, as well as the topics communicated: personal, family, friends, professional, or religious (I will note that personal matters are not mutually exclusive from familial or friend-related matters).

HYPOTHESIS 12: Messaging communication (IM and SMS) will be opposed to longer formats (phone and email), and associated with younger subjects

The emerging literature on instant messaging and SMS has suggested that currently teenagers in America are heavily using messaging formats, and doing so over other communication modalities (Schiano, Chen, J., and Gretarsdottir, 2002; Isaacs, Walendowski, Schiano, and Kamm, 2002; Grinter and Eldridge, 2003; Grinter and Palen, 2002). Because the subjects were expected to cross this generational line, I anticipated a large separation in these media.

Instant messaging is unique among popular communication media as the user is constantly shown a list of their friends they have previously articulated (boyd, danah, 2004). While phones have virtual phonebooks, and email clients have address books, the size and breadth of these contacts is not presented to the user continuously as it is with IM. For this reason, Subjects were asked to estimate the size of their "buddy lists," and approximate the percentage of this list that are friends, family, professional contacts, and people they meet with regularly.

Social Capital

In the final section I hope to extract some information about the greater social network of the subjects, most notably focusing on weak ties. Given my survey time-constraints, the best means of gathering this information would either be a position generator (Lin and Dumin, 1986; Lin, 2001), measuring an individual's access to individuals of varying occupational prestige, or a resource generator (Snijders, 1999; Van der Gaag et al., 2005), measuring a subject's access to specific resources through their social ties.

Both of these survey instruments are established measures of social capital, and their relative accuracy is still a topic of debate. Because the resource generator phrases the questions in terms of actually acquiring resources, it naturally favored individuals and resources that are nearby, as opposed to the possible access to those resources. Since I am interested in measuring the overall size and range of a subject's weak ties in terms of both support and access to information, the natural apparatus for this section is the position generator. Additionally, the position generator has been shown to be a better choice when time is an issue (Van der Gaag et al., 2005).

I used the instrument in Van der Gaag et al. (2005) which provides internationally standardized measures of occupational prestige in the form of ISEI socioeconomic index measures Ganzenboom and Treiman (2003). I adapted the occupation names slightly to be more recognizable to an American audience (e.g. I changed the profession *lorry driver* to *truck driver* and *estate agent* to *real-estate agent*). For each of the 30 occupations, I have asked the subject if they know such an individual, their relation to this person (acquaintance, friend or family) and whether the tie was established online or offline.

In addition to the standard instrument measures, I added one new component to evaluate tie formation, namely whether the introduction happened online

or offline. By online I mean through any communication medium that uses the internet (email, instant messaging, bulletin boards, etc.), while offline is either face-to-face or on the phone. This variable is being added to see both whether any of the other sections (including demographics, communication use and weblog use) are related in any way to having met more of these positions online. Because this measure is untested in a complete survey, I will not make any predictions as to its outcome.

Chapter 4

Results

During the months of May and June, the weblog aggregator observed the weblog community and collected data on individual behaviors. During the second and third weeks of June, the weblog survey was presented to both a random sample of authors and also to anyone who wished to participate. This section will detail the results and analyses of these two studies.

4.1 AGGREGATOR

The aggregator started collecting data on May 16th, 2005 at around 7am. My goal was to collect a full month's worth of diffusion data, but events out of my control cut that short by a few days. Sometime during the month of May, Blo.gs was sold to Yahoo! Inc., unbeknownst to anyone in the weblog community. Some representatives of Yahoo contacted me to let me know this would be happening, but the service was down between June 14th and June 16th.

Over the course of the 37 day period, over 15 million links were extracted from about 1 million weblogs. The updates observed are shown in Figure 4.1(a), along with the drop-out of the blo.gs service towards the end of the data-collection period. When weblogs are crawled initially, all links contained on the front page of the weblog are added to the database, including many that existed before the current update. This mass of relatively static links will be indexed the first time a weblog is crawled, and afterwards a much smaller set of new links will be found. This process of "getting to know" a weblog explains the severe peak and drop-off that occurs at the beginning of the data collection.

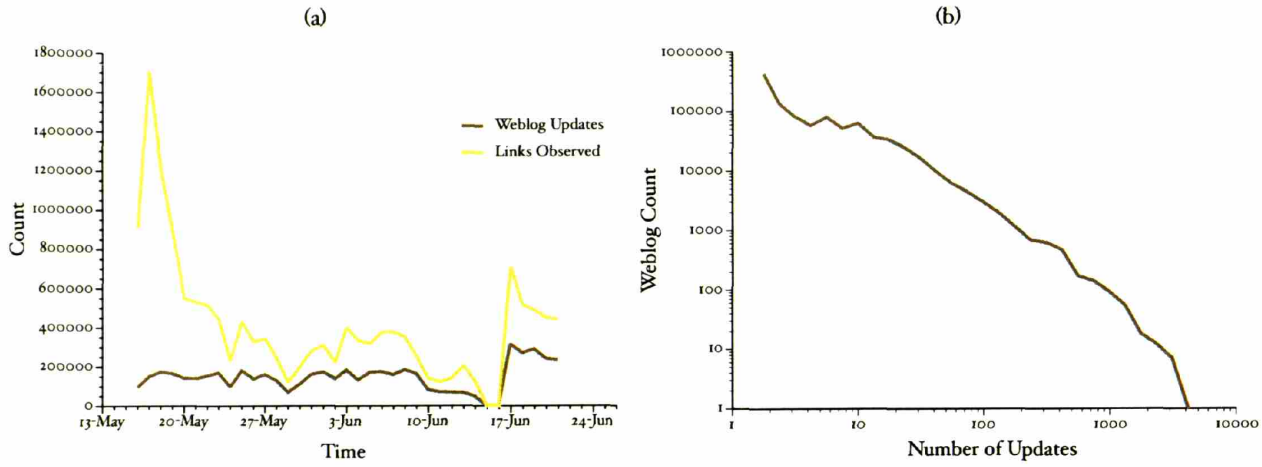


FIGURE 4.1. Weblog updates May 16-June 21, 2005. (a) shows the updates to the aggregator over time and (b) shows the updates over this entire period per weblog

Language

Of the 1,034,498 weblogs identified, the language of 647,556 could not be detected. Table 4.1 shows the distribution of languages for the 386,942 that were classified. The first percentage is with respect to the sample we have collected, or in other words the normalized value of each language. Using statistics collected from a number of primary sources, the market research company Global Research has assembled a list of estimated population figures based on native language (Global Reach, 2005). Using these figures, I calculated the difference in the expected number of weblogs and the observed as Δ %. This list exposes some of the biases of existing in the sample obtained from blo.gs. First, in some countries blogging is centralized around one or a few services, and as there is little need for outside aggregation of this material, these services tend not to involve themselves with ping tools. For example, Korea has an estimated 11 million bloggers (Lee, 2004), but nearly all of them exist on one centralized service, Cyworld (Cyworld, 2005). The few individuals who choose to be independent of this system may interface with a ping monitor, but it is extremely rare.

Second, some countries have their own ping services that do not interface with blo.gs. Japan for instance has a number of different ping monitors, many of them similar to blo.gs, but at this time they do not interact. This scenario only exists in a few countries, where the blogging population is large and distributed enough to necessitate such a system. Such ping servers exist in

Language	Detected	%	Δ %	Language	Detected	%	Δ %
English	272,597	70.45	33.59	Estonian	323	0.08	
Japanese	38,339	1.54	9.91	Arabic	273	0.07	-1.61
Spanish	13,377	3.46	-5.53	Romanian	231	0.06	-0.49
Portugese	13,363	3.45	0.41	Greek	229	0.06	-0.24
French	11,330	2.93	-1.30	Korean	228	0.06	-3.85
German	9,356	2.42	-4.48	Hungarian	208	0.05	-0.26
Chinese	6,996	1.81	-11.92	Tamil	184	0.05	
Italian	5,060	1.31	-2.49	Latin	180	0.05	
Indonesian	3,423	0.88	-0.89	Russian	149	0.04	-0.77
Farsi	2,949	0.76	0.19	Bulgarian	114	0.03	
Dutch	2,475	0.64	-1.10	Vietnamese	103	0.03	-0.70
Icelandic	1,121	0.29	0.26	Hebrew	84	0.02	-0.45
Swedish	954	0.25	-0.71	Slovak	71	0.02	-0.21
Turkish	639	0.17	-0.68	Serbian	67	0.02	
Finnish	624	0.16	-0.19	Lithuanian	56	0.01	
Cebuano	565	0.15		Slovene	38	0.01	-0.09
Norwegian	383	0.10	-0.16	Hindi	23	0.01	
Polish	372	0.10	-1.10	Other	126	0.03	
Danish	332	0.09	-0.28				

TABLE 4.1. Detected Languages. The distribution of detected languages, percentage of the total that had a defined language, and the difference from the expected percentage based on online population (Global Reach, 2005).

Japan and France while others in Sweden, Brazil, Germany and Poland are either completely or mostly inactive.

This list includes some surprising activity in a few languages. Compared to internet market research statistics (Global Reach, 2005), the largest anomalies among this list are Portuguese and Farsi, which are far above their projected online populations. The Portugese speaking population reflects a large presence of weblogs in Brazil, while Farsi constitutes a growing population of Iranian authors.

Readership Network

One of the most important pieces of data to be extracted from the aggregator data is the network which connects the authors of weblogs. This structure is simply the subset of links I have identified that point to other weblogs. These links are sorted into two types: *static links*, or links to the front page of another weblog and *dynamic links*, or links directly into the content of another

blog. As discussed in chapter 3, these two readership networks will be considered independently.

Because blo.gs is an open system with a published programmatic interface, it is susceptible to a number of different types of specious activity. There are many fraudulent uses of weblogs, most of which are aimed at the individual weblogs of legitimate authors. To distinguish this outstanding problem from what is typically referred to as *comment spam* or *blog spam* I will be using the following categories: First, web sites trying to promote their content can simply ping Blo.gs, despite not being a weblog at all (*ping spam*); second, deceitful parties can place their content on free weblog hosting services or using open source software that pings blo.gs automatically (*hosted spam*); finally, a new type of duplicitous content has emerged in the form of well-crafted weblogs written about the news or personal events that can seem completely legitimate; without looking at a few of the outbound links, there would be no reason to suspect anything was wrong (*blog forgery*)

PING SPAM

HOSTED SPAM

BLOG FORGERY

Without checking every site individually, it will be impossible to completely remove spam from my data set. However, because spam authors tend to operate using standard methods that create observable abnormalities, I will first attempt to diminish their impact as much as possible through a number of steps of refinement.

DATA REFINEMENT

Figure 4.1(b) shows this distribution of updates from the perspective of individual weblogs; this curve represents the log-binned distribution of updates over the sample period. The largest number of updates came from a weblog with over 3,000 in 34 days, or just over 88 updates per day. This amazing accomplishment suggests one of two explanations: either these updates are automated, or there is more than one person at work in changing the content of this weblog.

Looking at the top updaters in the data set reveals that, in fact, a majority of these high-transaction weblogs are indeed ping or hosted spam. The first non-spam weblog in this list is the site Linkfilter.net, a collaboratively written community blog, ranking in as the 30th must updated (with 1,909 updates). My first method for dealing with spam is to use this list as a filter, deleting the top updated sites which can be easily identified as fraudulent. This technique does not cover a broad range of spammers, but it removes a large amount of inaccurate links in a short amount of time. I have deleted these weblogs, which consumed a full 85 of the top 100 updaters.

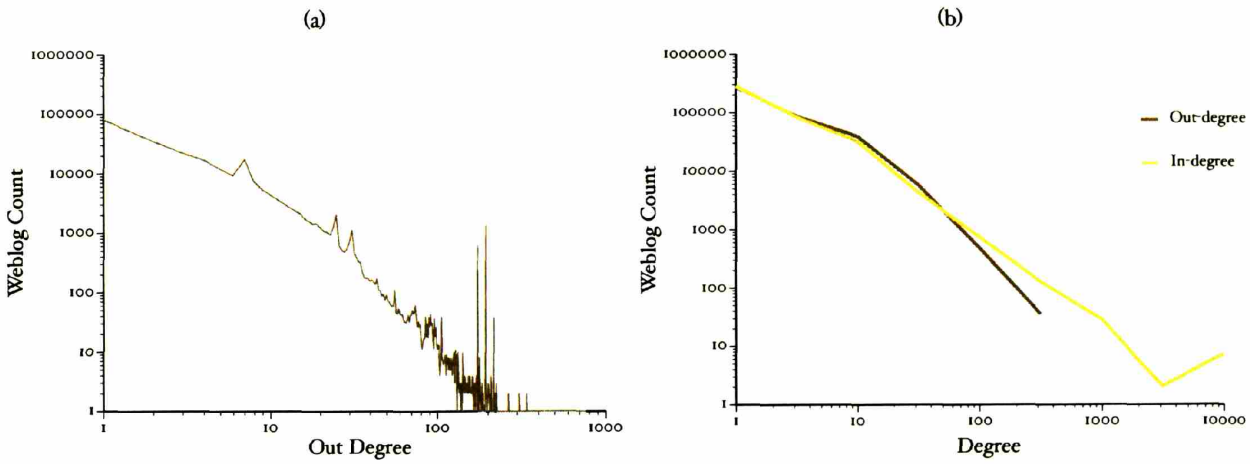


FIGURE 4.2. These diagrams show the degree distributions for the readership network. (a) is the initial observed outdegree while (b) shows the log-binned degrees after spam removal

My initial readership network contained around 425,000 weblogs with at least one link out, and about 500,000 including those with no out-links and at least one in-link from another weblog. Figure ?? shows the initial out-degree of the readership network plotted on a log-log scale. There are a number of large spikes off of what would otherwise be a fairly normal power-law distribution, most notably around the degrees of 25, 31, 174, 195 and 218. For instance, there are 1,377 weblogs in the readership network with an out-degree of 195, an incongruous amount considering that the surrounding out-degrees in the 190 rage have only one or two weblogs. Closer inspection reveals the fact that these weblogs have been automatically generated, and are not weblogs at all, but farms of hosted spam.

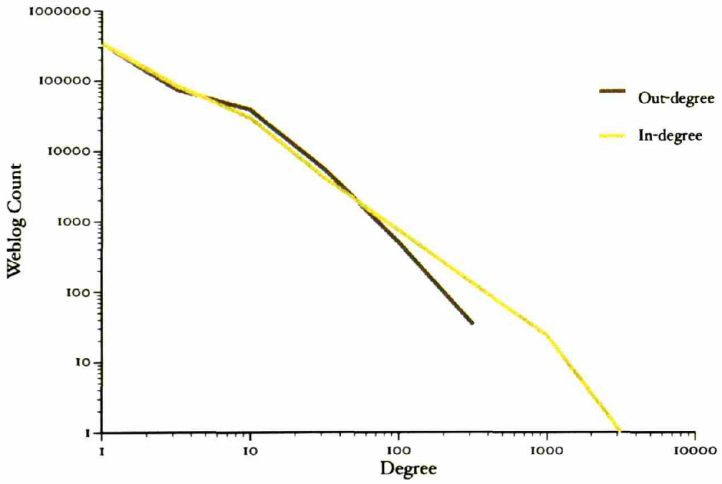
By removing these weblogs from the readership network, we achieve the more believable distributions shown in figure 4.2(b). However, a striking feature of this diagram is the massive spike at the tail of the distribution.This is an amazing feat, and again it should be interpreted with extreme suspicion. Looking at the list of blogs with top in-degree, we can see this break very clearly:

The first 7 entries are 3 times larger than the next site, Slashdot, which is unmistakably one of the more popular sites on the internet. Investigating these rank-leaders reveals that each is written by an author of the popular, open-source weblog software Wordpress (Wordpress, 2005). Their dominance in the readership network is not determined by their popularity or influence, but rather by the success of their software: each new installation of

TABLE 4.2. Top weblogs ranked by in-degree

Rank	URL	In-degree
1.	http://photomatt.net/	13,782
2.	http://zed1.com/journalized/	13,523
3.	http://www.alexking.org/	13,460
4.	http://dougal.gunters.org/	13,200
5.	http://blogs.linux.ie/xeer/	12,175
6.	http://zengun.org/weblog/	12,146
7.	http://blog.carthik.net/index.php	12,052
8.	http://slashdot.org/	3,889
9.	http://www.boingboing.net	3,105
10.	http://www.drudgereport.com/	2,281

FIGURE 4.3. Readership degree distribution after the removal of spam and artificial social links



Wordpress comes pre-configured to contain a set of self-indulgent links to the authors. † Unless a new Wordpress user manually removes these links, they remain as virtual advertising for a group of people the user has probably never met, read, or possibly even heard about.

For an discussion of the controversy surrounding these artificial links, see Kottke (2004).

As interesting as these artificial ties are, they are still artificial and should be removed from this analysis. Of the 13,782 individuals who link to at least one of these blogs, 13,216 link to 4 or more, with 11,816 linking to all seven. Starting with the premise that linking to many of these blogs is unlikely outside of the default installation, I have removed ***only these links*** from any site containing four or more. The resulting degree distribution looks much more as one would expect it to, shown in Figure **??**.

The final step in cleaning the readership network is to determine its connectedness and validate my hypothesis that the majority of nodes will be contained in one connected component. To arrive at the components in a digraph, the graph is first converted to its undirected form and then searched using breadth-first-search (Cormen, 2001, p. 532). Starting with an initial network of 385,350 nodes and 1,970,402 edges, the results can be seen in Table 4.3.

TABLE 4.3. Connected components. This list shows the connected components found in the readership network. The largest component contains over 89% of the total links, and singleton and dyadic outliers about 3%.

Size	$\sum$	Size	$\sum$	Size	$\sum$	Size	$\sum$
343,743	1	71	1	30	1	14	16
1,388	1	68	1	29	3	13	12
330	1	67	1	28	5	12	18
246	1	45	1	27	2	11	20
221	1	44	1	25	1	10	31
123	1	43	1	24	2	9	45
122	1	42	1	23	3	8	71
109	1	40	2	22	3	7	112
86	1	39	3	21	4	6	180
79	1	38	1	20	4	5	326
78	1	37	1	19	3	4	755
75	1	36	2	18	3	3	2143
74	1	35	1	17	5	2	9899
73	1	34	1	16	4	1	1739
72	1	33	2	15	10		

As can be seen from list list, a large majority of the entire set are contained in the largest connected component of 343,743 nodes. A thorough inspection of the other large components ($\sum > 10$) reveals a network of spam; typically using one or more free blog hosts, these sites often look and feel just like a weblog, their only differentiation being either the content or the links. In some cases, they even post content that resembles young, journal-style writing (such as the sites contained in one of the components of size 25).

The first non-spam component is a group of Dutch bloggers in a component of 23, and as the size diminishes, more legitimate clusters of blogs start to appear. A majority of these sites are authored in foreign languages, which suggests a pocket of authors in another country who use tools that ping blo.gs. I assume these weblogs are tied to networks of same-language bloggers who do not appear in my database simply because their localized blog software is

not setup to use ping servers. While a path probably exists from our main component to these authors, without a spidering approach to weblog acquisition, or a global ping server, they will remain isolated.

Without a connection to the main component, many of these outlying clusters of blogs will not provide accurate data for various measures, and they must be discarded for algorithms that require connected graphs. For this reason I will use the largest connected subnetwork shown above (consisting of 343,743 weblogs and 1,885,721 ties) as the readership network.

DEGREE

The degree distribution of our readership network can be a measure of how popularity, attention, and influence is divided up amongst our blog authors. The meaning of this measure is determined by the meaning of these links. Do bloggers link to people they read? Or is it someone they admire? How much influence does that person have over their thinking and writing?† Most of these questions cannot be answered merely with my extracted readership network, but we can draw few distinctions.

These questions are also addressed by the survey in Section 4.2.

Social links among weblogs come in two easily discernible forms, as stated in Chapter 2. *Static* links on a weblog are typically made on the edges of a site, either as a sign of readership, support, interest, or marker of a social relationship. *Dynamic* links are made when a blogger links to another author's specific writing, usually signifying a response or acknowledgment of their interest. Dynamic links tend to occur inline with the text of a weblog, and as the weblog is updated, they fall off of the front page; explicit links tend to remain regardless of how often the content is changed.

Of the links collected over the sample period, 1,399,749 static readership links were observed, and 541,234 dynamic, making the ratio about 3:1. Given the short time frame of the study, I had expected this ratio to be much higher, especially since, accounting for aggregation over time, my last look at these data would suggest something in the range of 10:1 (Marlow, 2004). Correlations between the static and dynamic degrees are shown in Table 4.4.

As I noted in Chapter 3, these correlations are unlikely given that we are observing true power laws. But given that these are bounded networks with a finite variance, we do see some relationship between the two, and in the case of In-degree, the relationship is quite strong. This implies that for variations around the mean of the distribution (which will be a low in-degree), the relationship will be so strong it will overcome the exceptionally large variance.

TABLE 4.4. Degree relationship. Following are the correlations between in-degree and out-degree measured by static and dynamic links

.

	In-Static	In-Dynamic	Out-Static
In-Dynamic	0.825		
Out-Static	0.120	0.063	
Out-Dynamic	0.077	0.066	0.259

$p < 0.001$ for all measures

The weblogs with the highest degree will have very little effect on this measure, so I will address them more specifically.

TABLE 4.5. Top Weblogs by Dynamic and Static Degree

Rank	Static	$\sum$	Dynamic	$\sum$
1	slashdot.org	3,893	engadget.com	1,963
2	boingboing.net	3,102	boingboing.net	1,930
3	drudgereport.com	2,280	binarybonsai.com	1,482
4	postsecret.blogspot.com	2211	dailykos.com	1,391
5	dailykos.com	2,060	slashdot.org	1,015
6	fark.com	1,980	huffingtonpost.com/theblog	842
7	dooce.com	1,858	gizmodo.com	745
8	engadget.com	1,671	arstechnica.com	725
9	talkingpointsmemo.com	1,547	goodpic.com/mt	687
10	gizmodo.com	1,490	michellemalkin.com	667
11	globeofblogs.com	1,481	powerlineblog.com	625
12	atrios.blogspot.com	1,457	radio.weblogs.com/0001011	609
13	wonkette.com	1,283	littlegreenfootballs.com/weblog	604
14	kottke.org	1,132	kottke.org	600
15	wilwheaton.net	1,125	alternet.org	586
16	powerlineblog.com	1,116	drudgereport.com	430
17	andrewsullivan.com	1,050	truthlaidbear.com	394
18	darthside.blogspot.com	965	milkandcookies.com	375
19	metafilter.com	933	metafilter.com	367
20	michellemalkin.com	932	talkingpointsmemo.com	367
21	xiaxue.blogspot.com	880	www.democracynow.org	362
22	gawker.com	867	atrios.blogspot.com	358
23	riverbendblog.blogspot.com	812	americablog.blogspot.com	357
24	volokh.com	770	www.sixapart.com	354
25	alistapart.com	763	www.volokh.com	341

The weblogs in Table 4.5 represent the cream of the crop for both static and dynamic links. A quick glance at this list reveals a number of these sites

appear on both lists, and many of the missing sites appear in the top 50 or 100 if not in the top 25. It is important to remember that we are dealing with an exponential falloff in degree, so while a difference of only 10 places in the rank might seem small, it can imply an order of magnitude difference in the overall distribution.

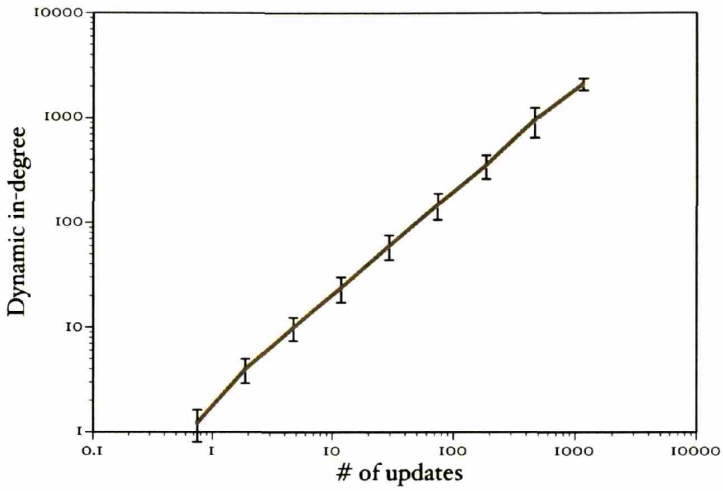
While there is certainly a trend between these two measures, this relationship deviates drastically as you move away from the top few sites. If I were to continue the list in table 4.5 into the hundreds or thousands, it would start to include many sites that weren't even ranked in both lists. Although the sample is different, and a year has passed, this distribution is no different than one observed in Marlow (2004). Comparing this list to the top weblogs in that same paper, and barring a potentially large bias in the older data, there is a considerable amount of change in the sets. Two of the blogs on the static list, *postsecret.blogspot.com* and *darthside.blogspot.com* are an art project and a humorous site respectively, and their status is reflective of a fleeting surge of popularity.

In both lists, a number of community weblogs are ranked at the top, Slashdot.org, Ars Technica, Alternet, Milk and Cookies, and Metafilter, while many of the rest are either professionally written by multiple authors, as with Drudge Report, Gizmodo, Engadget, BoingBoing, and a number of the political pudits. When the dust settles, there are very few sites at the top that fit our common perception of weblogs, namely that they are written by an individual. However, these are marginal cases, and depending on who you ask (including the authors), you will probably get different answers as to whether or not they should be considered a blog. Without drawing any lines, it is interesting to note that community blogs fill a role that few individuals could probably fill.

The fact that these community sites can provide something than an individual cannot is not surprising; the efforts of many people should be able to exceed one. But what property exactly is it that determines popularity, either from a dynamic or static perspective? My first assumption would be the quality of the information provided, and its general applicability to a wide range of interests. But one of the surprising characteristics of these top sites is the sheer volume of information that they produce. The top three sites across both lists—Slashdot, BoingBoing and Engadget—had 396, 791 and 615 updates respectively over the sample period. For BoingBoing and Engadget that amounts to over *20 posts per day*, and each from only a small number of writers.

Figure 4.4 shows the relationship between the number of updates made over

FIGURE 4.4. Updates vs. Dynamic in-degree



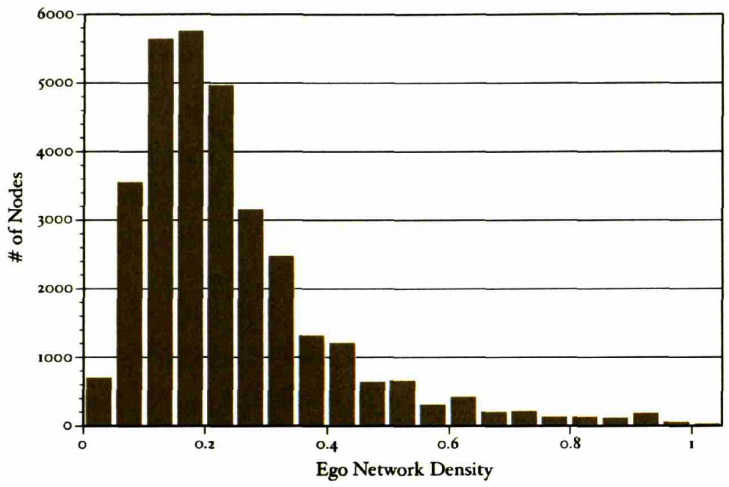
the course of the aggregation and in-degree from dynamic links. The graph shows the average number of updates over this period as a cumulative amount; for every in-degree, the value represents the average number of posts above that degree. Without any gauge of quality, it is clear that there is a strong relationship between the frequency of posting and a weblog's in-degree, and more-so for dynamic links than static. Intuitively this contradicts the notion that these power laws exist because of preferential attachment; if your degree purely related to the time at which you joined the network, then why is there such a clear relationship to the amount of information provided by these top sites? The fact that latecomers such as Engadget and Gizmodo can have such dominance also throws a wrench into the Barabási/Shirky argument. These data would suggest that there is more than just one force determining the growth of these networks. I will return to this in the next section when I propose an alternative model for network growth.

DENSITY

The density of the readership network is an important sign of how localized the interactions are. Given the stated number of edges and nodes, the total network density for the static links is $1.18e-5$, and for the dynamic network $4.58e-6$. The surprising quality of this network is that when we sum up all of the individual personal network densities, the average personal network density (clustering coefficient) of any one weblog is 0.34.

Because the degree distribution of this graph is a power law, a majority of the nodes will have very few neighbors, so this high clustering should not be that

FIGURE 4.5. Ego network densities for nodes with outdegree > 4 (N = 65,485)



startling. Figure 4.5 shows the network density for nodes with degree greater than 4. Even for the nodes of intermediate degree, the ego density is quite high. This fits into the model proposed by Watts-Strogatz, except for the fact that the characteristic path length of our graph is so high. One would expect with such high degree nodes as those at the top of our degree distribution that the distance between any two nodes would be small, but apparently there are subsets of weblogs that do not link to these hubs.

CENTRALITY

Centrality is an important measure, especially in a network with as much variation in density as the one I have described thus far. Because the data I have collected is in the form of a whole network, it has the potential to tell us which individuals are in positions of power based on the structure surrounding them.

Freeman outlines three measures of centrality in the first paper on the topic: *point centrality*, or simply the in-degree of each node, *closeness centrality*, or the average distance from a given node to all others in the network, and finally *betweenness centrality*, the probability that a given node lies on the shortest path between two others Freeman (1979). For my analysis, point centrality has already been addressed, but either of closeness or betweenness centrality would be extremely useful in relating an individual weblog to the rest of the network.

As noted in Chapter 2, measures of centrality are complex, and for large

enough graphs, intractable on any modern computers. Given our network of 300,000+ nodes with millions of edges, even the most optimized algorithms are intractable† (Brandes, 2001). Due to these constraints, I will have to seek other means for identifying central figures.

I started running the state-of-the-art betweenness measure and realized shortly after programming it that it would not be finished before my defense.

Since the innovation of Google's PageRank technology, it has become a standard for finding authority in large graphs without encountering scaling issues. Because it can be implemented with a few simple matrix multiplications, graphs represented as sparse matrices can be used without needing the space for the full adjacency matrix. The algorithm answers the question, given a constant random walk across the graph, what is the probability that the walk will be at the given node at any time? This notion is very close to (but not identical to) betweenness centrality, except it considers all paths between nodes as opposed to only the shortest path.

Dynamic Affinity

The observations made in the previous section suggest that the structure of readership within the blog community is not determined solely by the time at which the author joined the network. The difference between authority as determined by indegree derived from both static and dynamic links is a sign that more than one factor is at play in determining the distribution of links among authors. After accepting that preferential attachment is not the only force affecting network growth, the question remains, what forces do determine weblog structure?

To answer this question, I present a generative model that is capable of describing the data that I have observed empirically. Since rank within the static-link network appears to be somewhat age dependent, I will assume for now that this structure is dependent on preferential attachment, as it makes intuitive sense that a network needs to adopt leaders early on in its growth. But the constant appearance of new, strong nodes in the dynamic network must be determined by something else.

The primary observation I have made thus far is that the degree of a weblog within the dynamic network seems to be defined by their ability to continually produce information. If this production stops, the importance of that weblog will fall off over time. Instead of a model based on the compounding of links over time, I will now present a model based on *Dynamic Affinity*, a measure which is dependent on a model of attention.

The condition for preferential attachment as described by Barabási assumes that for a new node coming into the graph, the probability that this node will make a link to vertex i is dependent on the connectivity k_i as defined by $\Pi(k_i) = k_i / \sum_j k_j$. The condition for ***dynamic affinity*** is based purely on the update frequency of the node. As I have shown earlier, the probability that a weblog is updated in a given time period follows a power law. For now let us assume that this is related to the properties of our population, and not to the network itself.

At every time step t we will assume first that every edge will disappear with a fixed probability ρ, and that a constant number of links will be generated from each node to another node, with the probability of the edge being based on the affinity A_i. In the simplest case, if we set ρ to be 1, we rewire the network at every step, and assume that A_i is based on the power-law-distributed number of updates, it is clear that the graph will observe the same power law.

This may seem like a tautological statement, namely that one power law applied to a graph leads to another power law of the same form. The purpose of this exercise though is to show that given a power-law-distributed network, if we change the parameters of the ***nodes***, the network will reorder to fit the new distribution. Under preferential attachment, the number of links a node has is directly related to when it entered the network; in fact, in the original paper Barabási states that "growth and preferential attachment, are needed for the development of the stationary power-law distribution," and without network growth, the distribution will diverge. If we assume another power law describes the affinity of nodes towards each other, and a constant rate of link decay, we can achieve a dynamically stable, fixed-size network.

This leads us to the question, why does update frequency follow a power-law distribution? There are a lot of possible explanations, and as Zipf has shown us, we can find power laws everywhere (Zipf, 1949). If update frequency is inversely proportional to income, Pareto shows us that scaling will emerge (Pareto, 1896). For now I will take this as an empirical observation, but I will return to the question in the discussion of diffusion in the next section.

Media Contagion

With some footing in the various building blocks of weblog structure, I am now ready to address the issue of media contagion across this network. We have a number of measures of structure, defined both by the static and dynamic links formed by preferential attachment and dynamic affinity

respectively. The question now is: what about this structure determines the spread of a particular piece of media? How much of this diffusion happens by contagion, and how much of it is a product of external forces (such as the mass media)?

DATA REFINEMENT

As with the readership network, a few different techniques were necessary to turn the data into an acceptable form. I will walk through these steps to describe how I arrived at my final data set.

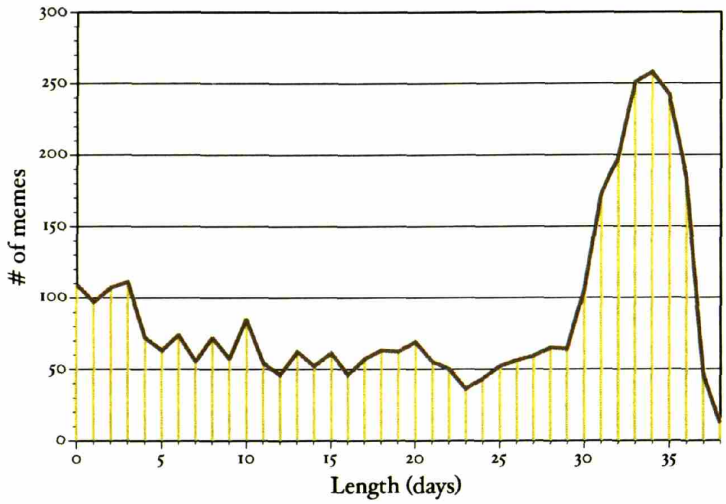
Over the course of the 5 week observation period, about 13 million total links were observed. Not all of the links are what we might consider "diffusing" links however; ignoring the links that are found the *first time* a weblog is seen, we know that any further links we observe there at future times were most probably updates made during the study. These "fresh" links reduce the data set by nearly one order of magnitude, down to 2.25 million. These represent individual references from a weblog to some other website; the total number of unique destinations is only 1.2 million in size.

Not surprisingly, the popularity of a given link over the course of this month is determined by a power law. To observe various properties of diffusion we will need a sufficient number of sample points to model this process; in the case of a link found on just a few blogs, it is hard to tell whether or not it is indeed the effect of diffusion. Because the distribution of popularity is scale free, any point at which the data is trimmed is going to be just as significant as any other. By cutting the data at 5, 10, or 20 links I am potentially removing some interesting class of diffusion, with the tradeoff being less noise and a much more tractable data set.

As a starting gauge, I have chosen those links that have been observed on at least 10 blogs during the course of the study. This cut-point reduces the data to only 3,549 examples, a number that is much more reasonable for computational purposes. In the even that some interesting class of diffusion is found around the lower limit of this size, I can always lower the minimum.

Another way to remove some unwanted noise is to look at the total diffusion time for each link. Figure 4.6 shows the distribution of diffusion time for all of the links observed. This curious distribution suggests that the probability of a meme taking more than M_T days to diffuse more or less diminishes as M_T increases except for the case where D is greater than 30. The low points around 5 and 12 days are most probably reflected by the weekly cycle that most determines meme diffusion:

FIGURE 4.6. Distribution of meme diffusion times



CUMULATIVE ADOPTION

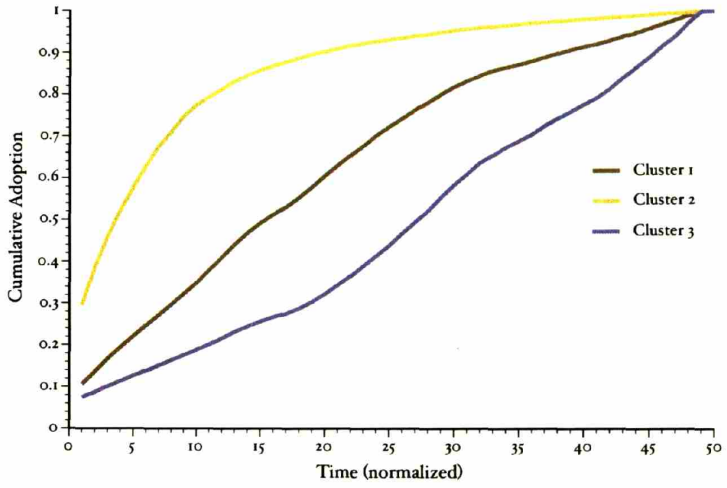
My first attempt in characterizing these data was to look at the properties of the cumulative adoption rate over time. This is the basis for looking at the growth of an idea or innovaTtion(Rogers, 1962; Valente, 1995), and as a means of explorative analysis on this large number of examples, I normalized the time and size of cumulative adoption and looked for groupings using a K-means approach of grouping the data Hartigan and Wong (1978). The largest distance between clusters came with three mean vectors, which are shown in Figure 4.7.

These three mean vectors appear to be exponential, logistic, and a mix of the two. This fits with the theories that diffusion can be affected by external (exponential) and internal (logistic) effects. In the former, we assume some constant pressure to all nodes in the network, while in the latter we assume only interpersonal transmission. In the case that both types of diffusion are possible, we can use a mixed model to approximate the various components of diffusion (Bass, 1969; Valente, 1993).

$$\frac{N - \frac{a(N-N_0)}{a+bN_0}e^{-(a+bN)t}}{1 + \frac{a(N-N_0)}{a+bN_0}e^{-(a+bN)t}} \tag{4.1}$$

In this equation a is the mass-media or external influence parameter, b is the interpersonal, or internal influence parameter, N is the sample size, N_0 is the initial number of adopters and t is time. The assumption is that a meme with

FIGURE 4.7. Mean vectors from K-Means clustering



high a will appear exponential in growth, and have very little structural influence, while memes with high b will have a long, logistic growth period. A combination of a and b will give something that has properties of both, and in some cases look more or less like linear growth.

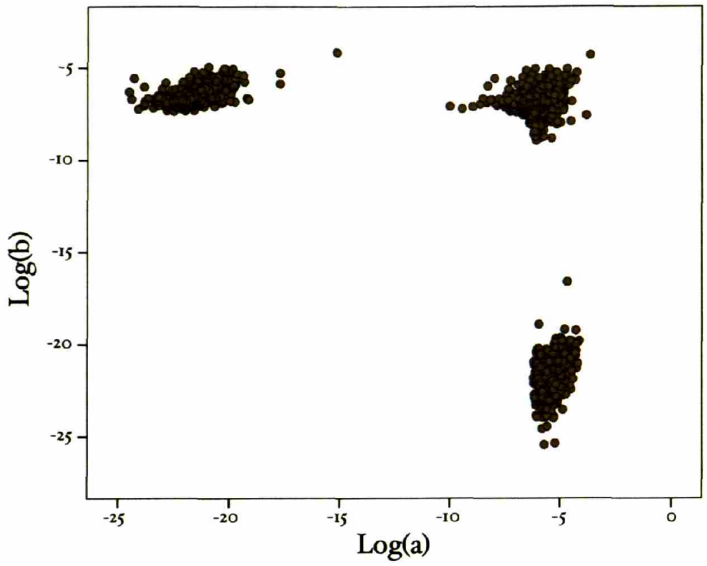
To test this model against my data, I use the simple regression technique *nonlinear least-squares* (Gallant, 1987) to find the values of a and b that fit the actual data as closely as possible to something generated by the mixed model. Unlike linear regression which has a very predictable behavior, non-linear regression are not as deterministic, and in some cases can find radically different answers depending on the starting values. In order to be sure that the regression is working properly, I will both verify that the output values make sense (i.e. the curves fit the data well), and also that the search space is not full of local minima† .

Local minima are places where the regression model appears to have found an optimal solution, when in fact it is suboptimal with respect to the global minimum.

The distribution of the fit values for a and b are shown in Figure 4.8. The distance between the clusters shown is amplified by the fact that the plot is shown in log-log scale, but the contrast between these groups is startling. For some number of memes, growth is defined by all-a and no b, while others are almost entirely b without any influence from a. A third and slightly larger group is defined by some mixture of a and b. While this seems like an important observation, I first need to confirm that the regression is indeed performing as it should.

Figure 4.1 shows the output of the regression for three different cases, each showing the actual and predicted values of cumulative adoption rate. The first

FIGURE 4.8. Fit values for a (mass media) and b (interpersonal) coefficients of growth



shows the growth of the site `blogebrity.com`, a site with a high a value and little b, the second shows `idolonfox.com`, with high b and low a, and finally the third shows `storewars.com`, an example with large values for both a and b.

I have chosen these three examples for a specific reason. The first, a website called `blogebrity.com` was a popularity-based ranking for weblogs, a sort of billboard charts of the a-list. This site was released during the time of the study, and as the data shows, it spread quite rapidly. In the regressed fit of this data, I found a to be nearly two orders of magnitude larger than b, although the fit seems to ignore the slight take-off in the beginning. This example is notable because during the time of the study, it received *no* external press, and thus all of the diffusion should be explained by structural features.

The second example, Fox's website for the popular American Idol television show (`idolonfox.com`) shows a growth that appears to be logistic at first. Compared with the first example, the growth is much slower at first, and thus has a higher b value. However, the cause of this sudden spike in popularity was the announcement of the show's winning contestant. The growth at the beginning of the curve is probably just regular chatter about the show that would appear linear if not for this announcement. This example exposes a flaw in the analysis, namely that we cannot assume that all of these media events start from a baseline at the beginning of the study; with more historical data, this example would have been a blip in the entire growth of the diffusion.

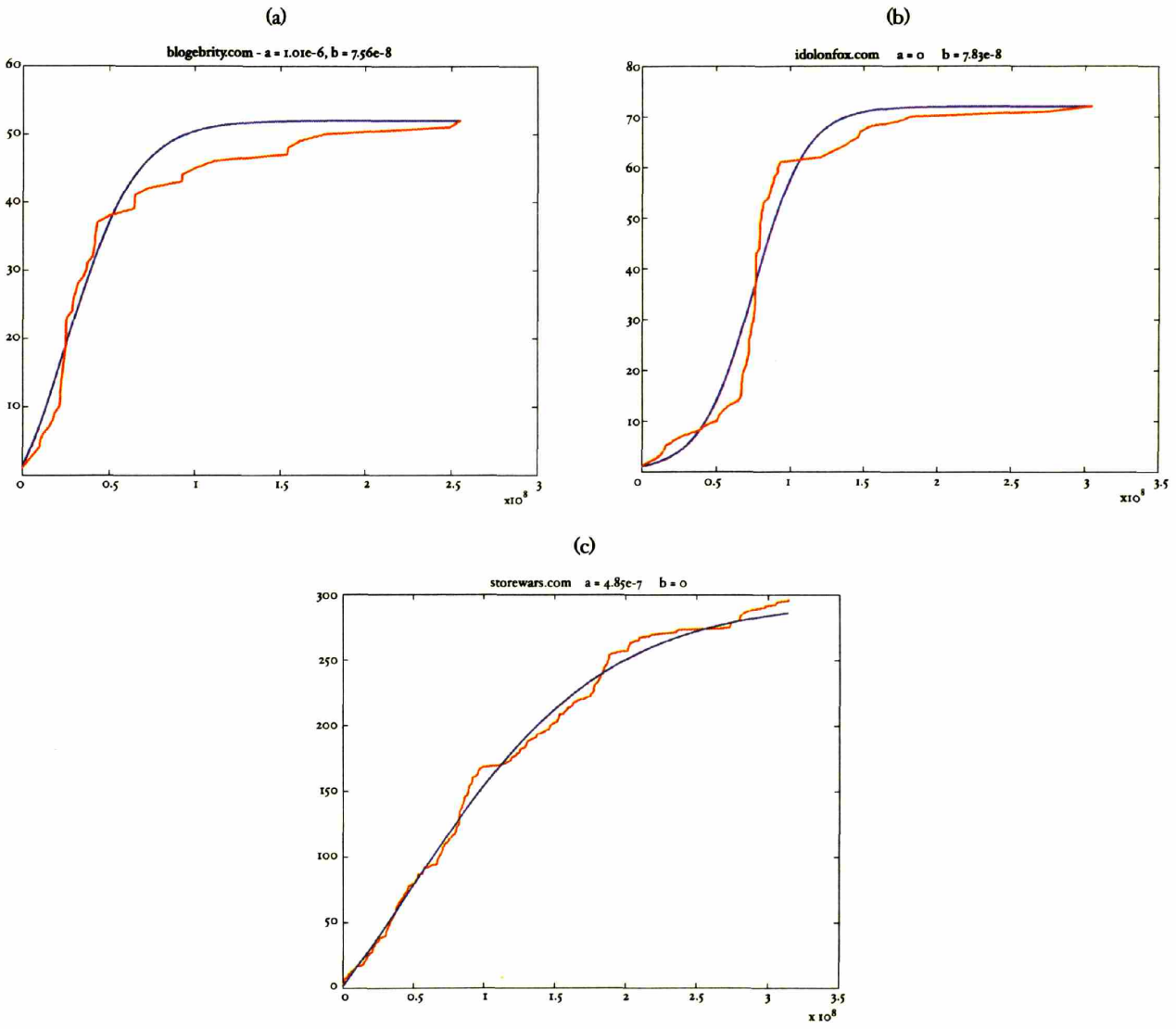


FIGURE 4.9. Actual and predicted curves for three diffusion types: (a) high a value (blogebrity.com); (b) high b value (idolonfox.com); (c) high $a + b$ (storewars.com)

"Join the organic rebellion," and fight with "Cuke Skywalker" against the forces of evil farms.

In the third case is a website presenting a humorous interpretation of the popular movie *Star Wars*, with the main characters replaced with organic food products† . This is a humorous website that was built to spread virally among sites just like weblogs, and the plot shows that it has received quite a bit of attention. My expectation would be that it would follow the familiar form of the S-curve, but over the course of the study it appears to have only grown approximately linearly. This could be the result of some external effects towards the beginning, but without better knowledge of its diffusion, it is hard to postulate.

While the regressed values of internal and external values showed promising results for separating the effects of internal and external diffusion, a qualitative look at the results shows that events that appear external can actually be internal, events that appear internal can be purely external, and events I would predict to be internal are actually a mixture of the two. Of course the internal/external dichotomy of growth are assumptions about the effect of structure on the growth of a media event; in order to validate this assertion, I need a better measure of structural diffusion.

STRUCTURAL MEASURES

While most diffusion studies have either no access to structural information (Deutschmann and Danielson, 1960; Funkhouser and McCombs, 1971; Greenberg, 1964) or information about the personal network of some respondents Burt (1987); Coleman et al. (1966); Valente (1995), very few have a whole network the size of that collected by the weblog aggregator. The scale of these data is both its largest windfall and its biggest hindrance; with more data comes the necessity to change various methodologies. I would prefer to validate measures such as structural equivalence, but the operating time of those measures precludes my using them.

Given the network of authorship and the set of diffusion events, the first question I will address is how many individual adoption events can be explained by structural contagion. Each diffusion event is comprised of a set of incidents I where we observe a link l on weblog w at time t. Every time we see a new incident, we can look at the adjacency list to see whether or neighbor of the new weblog has previously posted the link. At the end of the diffusion, there will be some number of components connected by structural diffusion and separated by nonstructural diffusion. This measure, the number of connected components, normalized by the size of the event, will be my first benchmark of the effect of structure on diffusion. This is defined by:

$$I_{CC} = 1 - \sum_{i=1}^{n} \frac{cp(i)}{n}, \tag{4.2}$$

where n is the number of incidents and $cp(i)$ is 1 if the node is connected to a component in the diffusion, and 0 otherwise. For instance, if none of the nodes of a diffusion event are completely disconnected, the number of connected components will be equal to the number of incidents, and the value of I_{CC} will be 0. If every new node in a diffusion event is connected however, the value of I_{CC} will approach 1 asymptotically as the event size increases. An important property of I_{CC} is that it is monotonically increasing over the course of an event, and as such it is comparable to other cumulative measures, such as adoption.

A large percentage (about 85%) of the events have zero contagion, and the rest include some component of inferred perceived structural diffusion. The mean value of I_{CC} is 0.29 implying that for any given link, roughly 30% of the edges could possibly explained by structure. However, this mean is deceiving given that such a large number have *no* perceived structural diffusion.

As an experiment, I considered a slightly extended version of I_{CC} where not only did first-degree contacts imply structural diffusion, but *second-degree* neighbors also did. For instance, if Blog A linked to something, followed by Blog B, even if A and B were not connected, I would consider the diffusion structural if there was some Blog C that both A and B had connections to. This measure I will call I_{CC2}. The surprising result is that while I_{CC2} greatly increases the space of possible contagion, the mean value is only 0.7, and only 35% of the events have a value greater than zero.

This measure is equivalent to adding all transitive links in the graph, e.g. if A links to B and B links to C, then we add the edge from A to C if it does not already exist. This increases the number of edges in the static graph from 1 million to about 9 million, almost an order of magnitude, and still only 35% have any contagion at all.

These values are far below what I expected; a large amount of weblog media exists within the weblog world. Take `blogebrity.com` for example. There is obviously no external diffusion in the sense described by Valente (1993), as one can be sure that the mass media is not involved, especially on the time scale of its diffusion. There are three plausible explanations for this growth. First, my sample of weblogs could be much smaller than the true census. If this were true, I could be missing many of the nodes that connect these weblogs to each other, and define the structural path. Unfortunately, this is difficult to

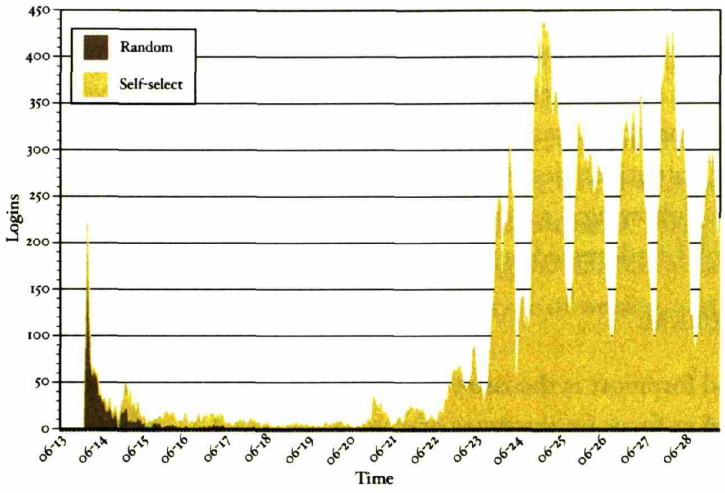
measure, and given other justifications of this sample (Herring et al., 2004), it seems unlikely that my sample would be this sparse. However, it could be the case that the links observed are not a good predictor of readership, and that much more inferential methods must be employed to acquire an accurate network (Adar et al., 2004).

Second, weblogs might not be the *primary* mode of diffusion for these events. Consider that some percentage of weblog authors are not only tied by their readership of each other, but also by email, instant messenger, telephone, or even an adjacent office. When something is read by someone, there could be a substantial amount of diffusion happening in other media. Akin to the justification made by Greenberg (1964), maybe this process happens much more as a push mechanism than pull: instead of posting something to my weblog, perhaps when I read something closely related to someone I know, I will send it to them through another medium. If this is true, then an event like `blogebrity.com` appear much more logistic if all of the other communication was included.

The third and most probable explanation for this discrepancy is simply that there is some force that inhibits weblog authors from linking to the things that their immediate neighbors do. If we think of neighbors as a sort of strong ties, then perhaps most bloggers think they share most of their audience with them. In this scenario, linking to something one degree away would be redundant; two degrees less so, and so on.

This issue needs to be addressed with further analysis. Given the data from this one month sample, a model has not emerged from my attempts to find any pattern to the contagious examples. Two methods showed promising data for predicting adoption patterns. First, PageRank was used to determine the probability that an individual would walk from any of the previous adopters to future adopters; this method proved to be a much better predictor than I_{CC} for contagion that was not over direct neighbors. Second, the characteristic path length over the *entire graph* for the set of adopters was equally good at predicting the amount of contagion that was involved. However, both of these measures were computationally intractable even for the limited number of examples in my data set. While I_{CC} is inaccurate, especially in the cases where hubs adopted early on, it is the only means I can find to circumvent the fact that bloggers might simply not be linking to the things their neighbors do.

FIGURE 4.10. New Subjects over time



4.2 SURVEY

The general social survey was released in two phases, first as an email to the random-sample subjects, and then publicized on both my personal weblog and from the top of the Blogdex service. My expectation was that the survey would diffuse naturally, given the self-reflective nature of bloggers. The number of new subjects over the course of the study is shown in Figure 4.10. Most notable in this diffusion is the takeoff after the introduction of some advertising buttons, after which the growth continued to be exponential for quite some time. I have provided a more in-depth analysis of this process as an example in Chapter **??**.

The official survey period was set to be 14 days, from Monday June 13th through Monday June 27th. At the end of the official survey period the growth of its diffusion was still at peak level, so in the interest of including as many people as possible, the survey was left online until the interest died off. However, for the purposes of completing this thesis, the data I will be presenting here will be only that gathered up through Tuesday June 28th (one day beyond the originally expected period).

I will break the analysis of these data into six sections: some caveats revealed in the execution of the survey; a synopsis of response rates and general demographics; and one for each of the other survey sections (links, weblog use, communication use, and social capital). Section **??** will provide a comparison of the survey results to the data collected by the weblog aggregator.

Caveats

Any survey that has a pilot of 25 individuals and eventually reaches 36,000 subjects is bound to have some gaps in its consideration of a topic. About 1.5% of the subjects responded via email after taking the survey, totaling about 450 personal emails to which I responded. These emails covered a range of potential holes and pitfalls which need to be recognized in my analysis. Following is a list of the major areas of discussion; question-specific issues will be addressed in their respective analyses.

LIVEJOURNAL

First and foremost is the issue of LiveJournal response. The survey was intended to focus on the types of weblogs observed by the aggregator (for the purposes of section **??**), and I did not expect the response that was eventually obtained by the LiveJournal community †. Out of the 36,000 respondents, almost 50% ended up specifying a LiveJournal site as their weblog, making them the largest subgroup that can be identified.

The breakdown of the survey sample is shown in the next section.

A number of problems arise for a LiveJournal subject of the survey. The structure of a LiveJournal site is quite different than any other; because the system is self-contained, the readership of a given weblog can be constrained to different security levels. For instance, an author can make certain posts for friends, another set for family, and yet another for the general public. Because the survey was not designed to handle such security issues, any content and links contained in private content will not be obtained in the Links section. A large number of responses I received claimed that the entirety of a subject's LiveJournal was in private form. Also along these lines, the LiveJournal system represents social links in a non-standard fashion rendering them invisible to my parser.

Reading LiveJournals one quickly becomes aware of the fact that the structure is qualitatively different than other weblogs. Because the survey was not piloted to *any* LiveJournal authors, many of the weblog-specific questions may be interpreted differently by a LiveJournal user. These issues will be addressed in the respective analyses.

WEBLOG CLASSIFICATION

The second-largest response received came from various individuals who felt that Question 23 (motivations for keeping their weblog) marginalized the type of writing that they were doing. Among these individuals the most emergent

categories were: fan-based (mostly arising from LiveJournal users), creative writing, and illustration-based.

Many other subjects also responded to Questions 29-31 (percentages of posts about personal matters, current events, and professional matters) as if these questions were meant to be mutually exclusive classifications of their content. Some were angered by the inclusion of professional matters at all, claiming that their use of weblogs should be studied in a different form. Many of these confusions could have been minimized by informing the subjects ahead of time of the difficulties of including every possible type of weblog, and offering a free-form input for adding their own motivation.

MULTIPLE AUTHORS

While the survey addressed the issue of multiple authors, it was not complete throughout the survey. For those individuals who worked collaboratively on a weblog with others, the Links Section proved difficult in the case that another author's links were selected. Questions about the specific attributes of the weblog were also ambiguous, soliciting either the author's individual contribution or the entire weblog (for which s/he would be less aware of).

LINKS

The Links Section was the least tested portion of the survey, and as a result was one of the most confusing parts for many subjects. In the places where it worked, it seemed to work well, but when it failed, it failed badly in many cases: the weblog could not be parsed (as with LiveJournal), the links it extracted contained automatically generated content, and often the authors were dissatisfied with the "randomness" of the selection† .

In most cases, subjects felt that too many links were taken from the static part of their site, not their posts, not realizing that a majority of their links were static.

Response and Demographics

Because the random and self-selected samples individually provide interesting results, I will interpret their representativity in this section. Characterizing the random sample is a straightforward task, but due to the open nature of the self-selected sample, dissecting the bias will be much more complicated. I will first deal with the issue of incomplete surveys (attrition), followed descriptions of both of the individual samples.

Because of a language mismatch between education systems around the world, the choices provided for the education question need to be normalized. The three choices of master's degree, doctoral degree and professional degree will

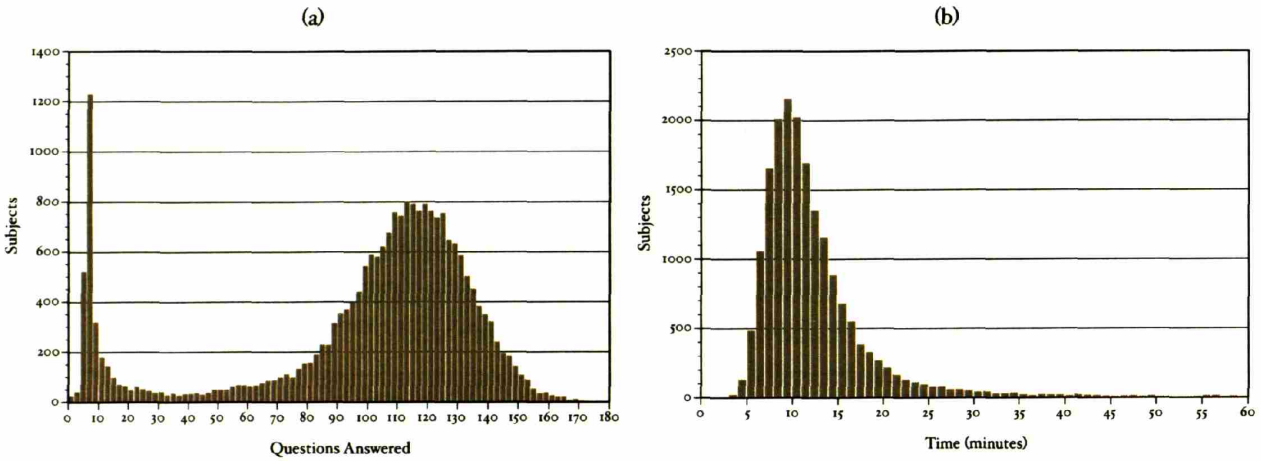


FIGURE 4.11. (a) shows the distribution of questions answered per subject and (b) shows the average time spent taking the survey

be combined into one "post-graduate" category, and the 2-year degree group will be folded into the college degree category.

As mentioned in the caveats, a large percentage of the self-selected sample (52.7%) identified themselves as LiveJournal users by entering a LiveJournal URL in the Links Section of the survey. For the reasons mentioned in the caveats, it is important that we separate these subjects from the rest of the respondents for various sections. This partitioning gives three total sample populations: *random*, those emailed directly to participate, and two self-selected groups that found the survey through other means, *LiveJournal*, those identified as being from LiveJournal, and *self-selected*, the remaining subjects.

ATTRITION

Not every subject who begins a survey finishes all of the sections; my first task is to define what constitutes a completed survey. The survey came in 6 parts, but every question answered by a subject was recorded individually. Because the survey had branches in many sections, a subject could have a fully completed survey with only 60 questions answered, ranging all the way up to 160+ questions if every possible question was answered. Figure 4.11(a) shows the distribution of questions answered per user. Statistically† and intuitively the break in incomplete and complete surveys seems to occur around 50 questions answered.

If we assume that the second half of the bimodal distribution is normal, two standard deviations would produce a cut point of 49.

The distribution in figure 4.11(b) shows the total amount of time spent by

subjects on only completed surveys (as defined above), measured from the first question sent to the server until the last question was registered. This response time is actually below my expectations based on the pilot survey, probably because some percentage of users answered considerably less questions than others.

RANDOM SAMPLE RESPONSE

5003 subjects were emailed as part of the random sample over the course of the suvery period. Of these initial emails, 201 were immediately rejected by email servers, 89 subjects opted-out for being misclassified using the link provided in the email, 219 individuals logged on but did not complete the survey, and 1,369 finished by the definition above. This left a remaining 3,125 who did not respond in any way.

Given these values, the overall response rate (for completed surveys) was 29%. This figure is very close to those obtained in other emailed random-sample surveys (Bosnjak and Tuten, 2003; Kypri and Gallagher, 2003). A number of factors differentiate this study from others, however. First, my expected rate of misclassification is much higher than would be expected of a targeted email survey; most surveys of this size deal with hand-collected emails whereas mine were collected automatically. The fact that 89 people actually went to the trouble of responding with this information suggests that there is a much larger population who simply discounted the email altogether.

Second, since the time of the last cited survey (2003), the email world has changed considerably. Spam detection has become a large part of the email landscape, and nearly every user today has a spam filter of some sort. Despite my pre-testing with various spam filters (Google, Yahoo, Hotmail, and SpamAssassin), it is impossible to know the actual configuration of each one, and my suspicion is that a number of these emails ended up classified as spam. With these caveats, a 29% response rate is quite respectable.

COUNTRY OF RESIDENCE

My initial intent was to translate the survey into a few popular, foreign languages, but time constraints forced me to release this version in a solely English language form. The subjects of the survey were not constrained by their country of origin or residence, but rather by their ability to read English. Table 4.6 shows the distribution of country of origin for all three samples.

While most of the samples have fairly similar representation, there are a few important distinctions. First, the random sample drew from a more diverse

TABLE 4.6. Observed Countries

	Country	Self-selected (%)	LiveJournal (%)	Random (%)
1.	United States	71.3	79.7	66.0
2.	Canada	6.5	6.3	6.3
3.	United Kingdom	4.9	5.4	6.2
4.	Australia	2.2	2.9	4.5
5.	Germany	1.6	0.8	0.4
6.	France	1.3	0.1	0.5
7.	Philippines	0.9	0.5	1.4
8.	Spain	1.3	0.1	0.1
9.	Netherlands	0.8	0.5	0.4
10.	Singapore	0.5	0.2	4.5
11.	India	0.8	0.1	1.4
12.	Finland	0.5	0.3	0.4
13.	New Zealand	0.3	0.4	0.7
14.	Ireland	0.4	0.3	0.5
15.	Malaysia	0.3	0.1	2.0
16.	Sweden	0.2	0.3	0.3
17.	Japan	0.3	0.2	0.4
18.	Belgium	0.4	0.1	0.1
19.	Denmark	0.4	0.1	0.2
20.	Iran	0.4		0.1

set of English-speaking countries, including Canada, the UK, Australia, and a large portion from Singapore, while the other samples had a significantly higher U.S. response. This variation is best explained by the prevalence of English in these countries along with the popularity of the blogging tool Blogspot†. Since Blogspot is the only major international hosting service my aggregator was tracking, many of these distinctions can be attributed to this bias.

Of the 51 random subjects from Singapore, 92% of them used a Blogspot weblog.

INCOMPLETE

Table 4.7 contains the general demographic information (age, sex and education) for all three samples and their respective incomplete groups labeled with an (I). The variables were coded as follows: age was measured as the current year (2005) minus the year they entered as their birth year, sex was coded as 0 for male and 1 for female, and education was coded as 0 being less than High School to 6 being a graduate degree. I will first address the possible bias in those individuals that did not finish the survey, and then explain some of the discrepancies in the three individual samples.

On average, those respondents that provided some demographic information but did not complete the survey were less educated, younger and more male

TABLE 4.7. Sample demographics

Sample		Age	Sex	Education
Self-selected	Mean	29.2	.55	2.6
	Std. Deviation	9.3	.50	1.1
	N	12,774	12,732	12,787
Self-selected (I)	Mean	27.4	.59	2.4
	Std. Deviation	9.6	.49	1.1
	N (24.4%)	3,638	3,667	3,648
LiveJournal	Mean	26.7	.71	2.4
	Std. Deviation	7.5	.45	.9
	N	15,776	15,736	15,817
LiveJournal (I)	Mean	25.1	.68	2.2
	Std. Deviation	7.1	.47	1.0
	N(8.8%)	1,537	1,527	1,542
Random	Mean	30.2	.31	2.6
	Std. Deviation	10.6	.46	1.1
	N	1,358	1,360	1,361
Random (I)	Mean	30.9	.28	2.5
	Std. Deviation	11.8	.45	1.1
	N (13.8%)	171	173	172
Total	Mean	27.8	.62	2.5
	Std. Deviation	8.7	.48	1.0
	N	35,254	35,195	35,327

than those who finished, except in the case of the self-selected sample which was more female. The most significant incomplete group occurred in the self-selected sample at almost 25%, with both the random and LiveJournal samples at nearly half that rate. On average the random and LiveJournal incomplete samples also answered more questions on average, albeit with a larger standard deviation. Table 4.2 summarizes these data.

TABLE 4.8. Survey completion rates

Sample	%	μ	σ
Self-Selected	85.6	112.1	20.4
Incomplete	24.4	8.8	9.5
LiveJournal	91.2	114.0	19.0
Incomplete	8.8	16.3	11.1
Random	86.2	111.2	23.7
Incomplete	13.8	11.2	12.4

Because all three samples have very similar biases in their incompleted populations, it is unlikely that something about the survey created this

disparity. The most likely explanation is that something about LiveJournal sites, or the way in which the survey was presented on these sites provided more motivation to finish than the other self-selecting sites.

LIVEJOURNAL REPRESENTIVITY

Since LiveJournal provides a list of their user demographics (LiveJournal, 2005), we can answer this question quantitatively.

TABLE 4.9. Demographics reported by survey and LiveJournal (LiveJournal, 2005)

Gender	Survey (%)	LiveJournal (%)
Female	71.02	67.3
Male	29.91	32.7
Country	**Survey (%)**	**LiveJournal (%)**
United States	79.38	79.88
Canada	6.24	5.61
United Kingdom	5.37	4.43
Australia	2.84	2.00
Germany	0.76	0.60
Age	**Survey**	**LiveJournal**
μ	26.68	20.50
σ	7.51	6.95

Table 4.2 shows the deviation between LiveJournal subjects and the statistics reported by Livejournal.com. While the survey sample seems surprisingly young and female, the service itself contains this same bias. Of the top 5 countries reported by respondents, LiveJournal reports a very similar breakdown for their users, although the survey is missing a large component of their population living in the Russian Federation.

The only striking difference between what I have observed and what LiveJournal has recorded is in the age of their community. With an average age of over 6 years younger, it is hard to explain away this bias without assuming that the survey was less appealing to the more youthful crowd. Since the statistics compiled on LiveJournal are for the lifetime of the service, there is a potential that the service has "grown up," or had an increasing average age since its inception.

The survey data was for just over two weeks—in the last month only 1.4 million LiveJournals were updated.

While the statistics reported cover all of the 7 million accounts created, under 1.4 million were updated over the time that the survey was live† . Without

knowing what the churn of this service is, and who exactly has stopped participating, it is hard to specify exactly what the error in my sample is.

GENDER

While the country of origin, education, and age distributions are roughly the same between the random sample and the self-selected, the gender profile is remarkably different. This is an important discrepancy, and there are a number of possible explanations. If some percentage of LiveJournal users decided not to identify themselves in the second section, the large majority of women users there could balance out the otherwise male-dominant population described by the random sample. If this were the case however, one would also expect to see a shift in the other demographics, which is not apparent.

The only potential LiveJournal subjects who could affect this remaining sample are those individuals who did not enter the address of their weblog. Removing these 4,785 subjects, the percentage of female respondents actually *increases* to 54.5%. In total, the male-female ratio of the entire sample is fairly similar to the statistics compiled by Perseus Development (2004). One of the biggest stipulations of their survey (and the subsequent survey the completed (Perseus Development, 2005)) was the fact that they do not include non-hosted weblogs. Since a large percentage of my random set comes from these weblogs, this bias could also be attributed to the fact that fewer women use non-hosted weblog tools, such as MovableType or WordPress.

Exactly 700 of the 1,369 random respondents came from Blogspot. Of these individuals, 440 (62.8%) were male while 260 (37.2%) were female. Of the remaining individuals who entered a URL, 412 (76.2%) were male and 129 (23.8%) were female. If we assume that these subjects came from non-hosted weblogs, we can assume that this environment is biased toward male users. However, if we assume that the Perseus data are still accurate, another more likely justification for this difference is that there is a bias in the group of individuals willing to put their email on the front page of their weblog. Given that our LiveJournal sample reached more women than expected, it is unlikely that the survey itself appealed more to men, or that in the weblog community men are more like to take surveys. Likewise, it could also be that women do not commonly respond to an unsolicited email.

Weblog Use

The third section of the survey aimed to characterize the different ways in which weblogs are being used. I will address it first since the analysis of other sections is in some way contingent to understanding this space. Weblog use was broken down into two primary sections, the first dealing with general weblog experience and use, and the second focusing on the subject's current weblog. Within these two groups, there are essentially four types of measures that can be extracted from this section of the survey: history (length of use), investment (time spent posting, commenting and reading other weblogs), self-reported audience size, and genres of authorship. I will attend to each of these variables independently, and address their relation to one another.

WEBLOG GENRES

One of the unknown factors that prompted this survey was the way in which weblogs were being used. Based on the pilot data, my hypothesis is that there are essentially three genres of weblogs: *journals*, with content mainly about personal experience, *editorial*, focusing on responding to news and online media, and *professional*, a sort of notebook for one's professional life. I assume there is quite a bit of overlap between these categories, but I expect that these will be the emerging dimensions.

In order to test this hypothesis, I took as many motivations as I could cull and created one question that asked subjects to check all of the primary reasons they kept their weblog from this list. The list of possible motivations is shown in Table 4.2. Each was chosen from a specific type of weblog that was given as an example by pilot subjects. Most pilot subjects classified themselves in numerous categories, with the mean around 2.5.

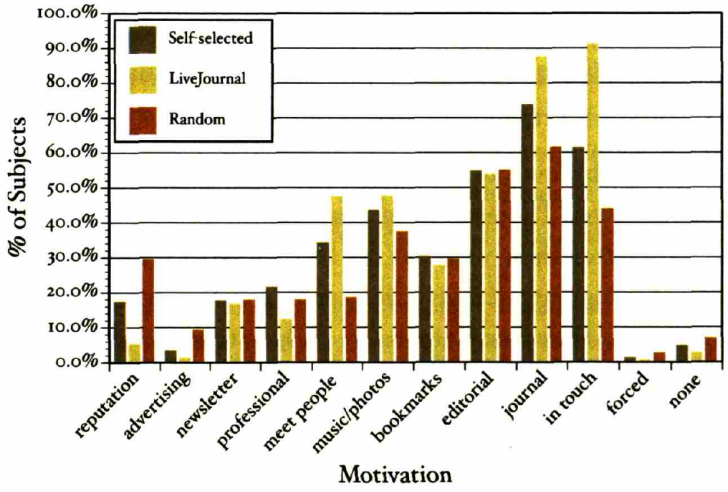
As mentioned in the caveats, I received a good number of emails from users feeling marginalized by this list, despite my efforts to be as inclusive as possible. In each of the samples, about 5% of the subjects chose "none of the above", but in half of these cases they also chose another motivation suggesting that they though some other category was left off the list. In the case of the self-selected sample, this value was slightly higher (about 2/3 of the "none" listed no other choices), and slightly lower in the case of LiveJournal.

The average number of motivations was between 3.6 in the random sample and 3.9 for LiveJournal users, all with a standard deviation of about 1.7. These results suggest that few people think of their weblogs as a tool for a single purpose. Despite the fact that the question stated "primary motivation(s)" as the indicator, most people chose many reasons, and this fact did not deviate

TABLE 4.10. Weblog Motivations

Motivation	Variable
Increase your professional reputation	reputation
Make money through advertising	advertising
To post news about an organization or project	newsletter
Keep notes for your professional interests	professional
Meet new people	meet people
Post photos you have taken or music you have made	music/photos
Keep a list of links to things you have read	bookmarks
Comment about things you read in the news	editorial
Keep notes or record what's going on in your life	journal
Keep in touch with friends	in touch
My work/school forces me to	forced
None of the above	none

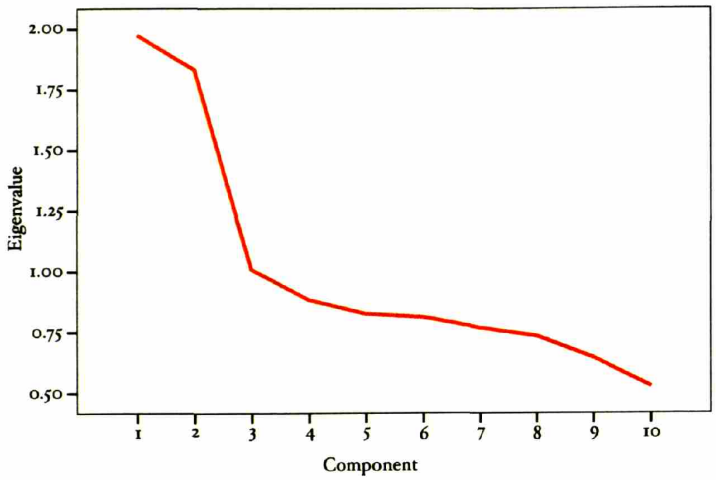
FIGURE 4.12. Weblog Motivations by Sample



across samples. The various choices are shown in Figure 4.12 normalized to the samples. A value of 100% in this case means that all subjects responded affirmatively to the motivation listed.

The three most popular categories are editorial, journal and in-touch, followed in descending order by music/photos, bookmarks, meet people, professional, newsletter, reputation, advertising, none and forced. Across the sample populations, there is divergence between LiveJournal and both the random and self-selected, where LiveJournal subjects tended to choose social motivations much more than the others, and professional matters much less. The most consistent category across all three was editorial writing, or

FIGURE 4.13. Scree Plot of PCA for Weblog Motivation



responding to things in the news. Over 50% in every sample chose this option; my initial expectation that LiveJournal was exclusively about social interaction and not current events turned out to be false.

This is still a very shallow understanding of the interactions between the different motivations. There are a number of different methods by which the relationships between these variables can be analyzed. One method would be to take the individual lists of motivations and cluster them, looking for obvious groupings. However, since we are looking to take a set of features and reduce them into more salient features, probably the most effective means is principle component analysis (PCA).

Correlations were used as dimensions, and Quartermax rotation was applied to make the vectors easier to interpret. The rotation converged in 3 iterations.

All ten dimensions were reduced using PCA† ; the scree plot of these components is shown in figure 4.13. This plot essentially shows the amount of information described by each component. As the figure clearly shows, two components stand out much higher than the rest. These two extracted vectors are shown in Table ??.

Looking at the various contributions to these two components, it becomes clear that Component 1 contains all of the variables associated with professional activity while Component 2 includes all of the purely personal motivations. Both components have some element of bookmarking and editorial writing, but in all other dimensions they are fairly divergent. I should note that adding the third component essentially splits Component 1 in half, removing the editorial and linking nature from the professional component, but this division is not as salient as the one I've just shown.

TABLE 4.11. Results of Motivation PCA

	Component	
	1	2
professional	.714	-.029
reputation	.709	-.225
newsletter	.540	.120
bookmarks	.434	.349
editorial	.408	.370
advertising	.419	-.064
journal	-.171	.653
in touch	-.257	.644
photos/music	.141	.608
meet people	.122	.550

While these components are not completely orthogonal with respect to the various motivations, they do provide an interesting measure of why the given subject has decided to write the weblog. My hypothesis that there were three main subjects that drove weblog—authorship, news, personal matters, and professional matters—was partially correct. It appears that news and linking are actually part of both communities described by professional and personal writing. These measures will be useful in future analyses, and I will refer to the components as motivational scores M_{prof} and M_{pers} respectively.

TABLE 4.12. Post-type frequency correlations

	M_{pers}	M_{prof}	Personal Freq.	News Freq.	Work Freq.
M_{pers}	1	.000	.305	-.158	-.194
M_{prof}	.000	1	-.418	.344	.355
Personal Freq.	.305	-.418	1	-.597	-.228
News Freq.	-.158	.344	-.597	1	.119
Work Freq.	-.194	.355	-.228	.119	1

All correlations significant at the 0.001 level (2-tailed)

At the end of the section, subjects were asked to report the percentage of their posts that were of personal nature, news-related, and related to their profession. To validate these measures, Table 4.2 shows the correlations between each component and the frequencies of these three post types. As expected, the M_{prof} is negatively associated with personal posts, while strongly correlated with both news- and work-related posts. Likewise, the converse is true for the M_{pers}, to a slightly lesser degree. This can be explained by the distributions of M_{prof} and M_{pers} across the subject population in Figure 4.2. While the mean and standard deviations of these

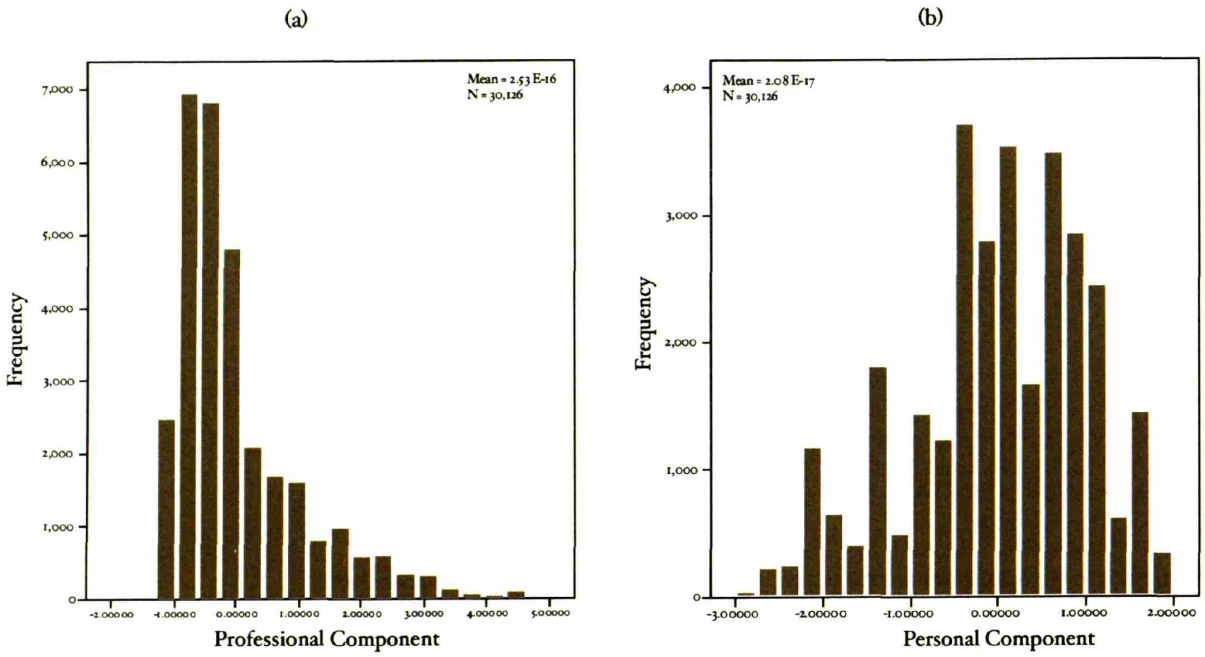


FIGURE 4.14. Personal and Professional Component Distributions

distributions are set by the PCA to be at 0 and 1 respectively, their shape is markedly different, with M_{pers} being much more evenly distributed over the population than the M_{prof}.

INVESTMENT

Given the diversity of update times observed by the aggregator, I expected to find a range of weblog use from amateur or busy authors who only updated their weblog once in a while to those that poured their hearts into the craft. To measure the level of commitment that a given subject had towards their site, I employed a number of questions about the time invested into various related activities. Weblog acts were divided into three different pursuits: reading other people's sites, posting to your own, and commenting on others. I also added a general question about the total time invested during an average week. Against these variables I wished to measure the effect that this input had on the popularity of their site, as quantified by their self-reported audience size, comments received in an average week, and the amount of time they expected to continue their weblog.

Each of the time-investment measures is was represented as an ordinal variable; in the case of reading, this took the form of the number of weblogs

read on a given day. For comments I worried that the frequency of commenting would be too low among some types of bloggers to be observable on a weekly timeframe, so the question was stated as the "total number of weblogs commented on in the past year." Posts and comments on the subject's own weblog were measured in terms of gradually diminished frequencies from many times per day to once every month, or never. Despite the discrepancy in scale of these variables, there was a very strong association between these various forms of activity.

TABLE 4.13. Investment into weblogging: *Read* is the number of weblogs an author reads weekly; *Time* is total time invested weekly, *Comt. Out* is the range of comments the author made, *Post* is their post frequency, *Comt. In* is the frequency of comments received, *Audience* is self-reported audience size and *Futures* is the amount of time they expect their weblog will remain

Self-selected							
	Read	Time	Comt. Out	Post	Comt. In	Audience	Futures
Read	1.000	.452	.411	.206	.220	.352	.175
Time	.452	1.000	.387	.310	.271	.308	.181
Comt. Out	.411	.387	1.000	.248	.475	.331	.122
Post	.206	.310	.248	1.000	.501	.322	.173
Comt In	.220	.271	.475	.501	1.000	.443	.086
Audience	.352	.308	.331	.322	.443	1.000	.149
Futures	.175	.181	.122	.173	.086	.149	1.000
LiveJournal							
	Read	Time	Comt. Out	Post	Comt. In	Audience	Futures
Read	1.000	.492	.519	.271	.342	.542	.224
Time	.492	1.000	.405	.350	.331	.327	.211
Comt. Out	.519	.405	1.000	.296	.436	.492	.183
Post	.271	.350	.296	1.000	.666	.280	.209
Comt In	.342	.331	.436	.666	1.000	.467	.174
Audience	.542	.327	.492	.280	.467	1.000	.187
Futures	.224	.211	.183	.209	.174	.187	1.000
Random							
	Read	Time	Comt. Out	Post	Comt. In	Audience	Futures
Read	1.000	.495	.392	.262	.260	.296	.138
Time	.495	1.000	.396	.422	.319	.313	.179
Comt. Out	.392	.396	1.000	.249	.459	.266	.121
Post	.262	.422	.249	1.000	.424	.324	.157
Comt. In	.260	.319	.459	.424	1.000	.480	.067
Audience	.296	.313	.266	.324	.480	1.000	.137
Futures	.138	.179	.121	.157	.067	.137	1.000

$p < 0.001$ for all values

Table 4.2 shows the correlations between each of these activities individually for every sample. Nearly every investment is positively correlated with all of the others, suggesting that as the amount of time spent increased, so did each of these various activities. Most notable is the relationship between commenting, posting, and receiving comments; regardless of the sample, the number of comments received is related both to the amount of time invested into the various activities, including reading other weblogs. Likewise, the relationship between audience size, albeit self-reported, varies according to these investment measures, as does the authors expectation of how long they will continue the activity.

Unfortunately due to the nature of this survey, I cannot definitively determine the direction of the causality; it might be the case that the more popular weblogs inspire their authors to invest more time, or the invested time could be rewarded with larger audiences and more frequent comments. The relationship between comments posted and comments received though, regardless of its origin, suggests that commenting is not an activity that can be maintained without some investment back into the community. Surprisingly this is true in every sample, not just the social environment engendered by LiveJournal.

Communication Use

The section on communication use had two purposes: first, I hoped to see the range of tools used by weblog authors and the extent to which they utilized them. My hypothesis here is simply that communication tools will be heavily dependent on age and marginally dependent on gender. Secondly, and more importantly, I hope to use the variables in this section to control for the amount of correspondence a subject engages in, in case some weblog variables are heavily correlated with communication use in general.

The correlations between demographic variables and various communication media shown in Table 4.14 exhibit the fact that the modality of communication we use is heavily dependent on who we are. Keeping in mind the fact that this is a sample of weblog authors who are already selected for being adopters of new communication media, some trends should be expected to emerge. The correlation between older technologies (phone and email) are expected as a younger generation transitions to message-based forms of

TABLE 4.14. Demographics and communication frequencies

	Age	Education	Sex
IM	-.294*	-.222*	-.048*
Phone	.163*	.113*	-.074*
SMS	-.125*	-.038*	-.009
Email	.341*	.319*	-.155*

* $p < 0.001$

communication (Grinter and Palen, 2002; Schiano et al., 2002), as exhibited by the negative relation between age and both IM and SMS. While the relationship between education and media disappears when controlling for age, the potential gender bias away from email should also be noted.

TABLE 4.15. Communication frequency correlations

	Phone	Email	IM	SMS
Phone	1.000	.316	.126	.149
Email	.316	1.000	.170	.120
IM	.126	.170	1.000	.197
SMS	.149	.120	.197	1.000

Control variables: Age, Gender

$p < 0.001$ for all measures

When we remove the variability in both age and gender, we see the generally positive relationships between frequency of media use shown in Table 4.15. Even without the tendency lent with age, there is a strong relationship between increased email and phone frequency. My expectation that there would be a negative relationship between short-messaging media and both email and phone is not confirmed, as increased use in one media generally reflects increased use in all others.

While communication use in these various media could reflect different orders of magnitude in communication use, the ordinal nature of the variables with such a small range makes them less significant. Because it appears that, for the most part, there aren't any negative relationships between these modalities, I will express the overall communication frequency as the sum of the individual measures, Comm^T. Since these variables are measured in terms of "total number of individuals on an average weekday," I cannot assume that these individuals are unique. In this fashion, this variable reflects at least an increased modality for individuals with frequent interaction in multiple media.

FIGURE 4.15. Distribution of Total Communication Frequency ($Comm^T$)

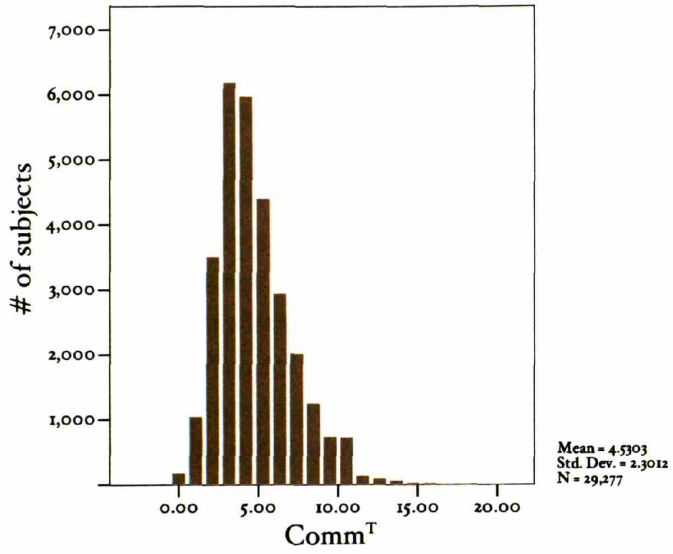
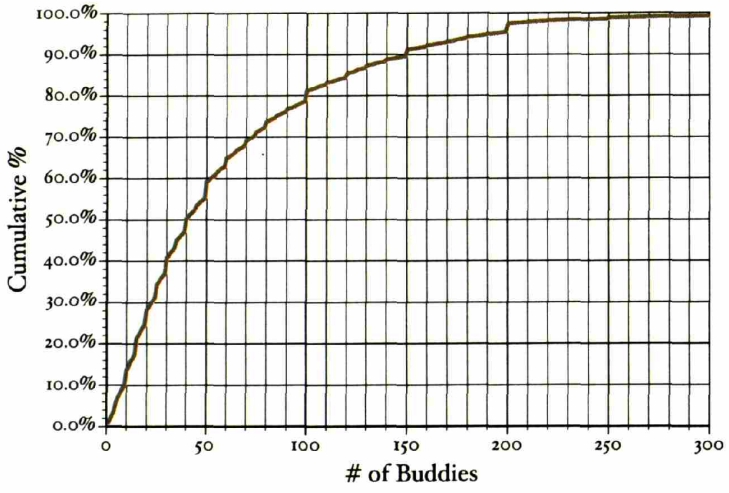


FIGURE 4.16. Buddy list size (cumulative)



The distribution of $Comm^T$ is shown in Figure ?? The distribution of $Comm^T$ across all survey populations shows a very pronounced mean very close to the median, and a falloff towards the extreme communicators. I will return to this measure in the next section of the survey.

The last part of the communication-use section probed the subjects' use of instant messaging as a method of exploring basic categories of interaction that

exist therein. Since I was not afforded the time to ask specific questions about these individuals, I chose to look at the breakdown between stronger ties (family and friends), professional ties, and everyone else. I first asked the subject to give the total size of their buddy list, $\sum_{IM}$, and then provide the percentage of this list that made up family, friends and professional ties, and the percentage of these that they met offline with once a month. The correlations between these percentages is displayed in Table 4.16.

TABLE 4.16. IM frequencies

	$\sum_{IM}$	Family	Friends	Work	Offline
$\sum_{IM}$	1.000	-.101*	-.137*	.060	-.039
Family	-.101*	1.000	.000	-.060	.124*
Friends	-.137*	.000	1.000	-.169*	.266*
Work	.060*	-.060*	-.169*	1.000	.139*
Offline	-.039*	.124*	.266*	.139*	1.000

Control variables: Age, Gender

* $p < 0.001$

The relationship between friends, family, and ties we meet offline once a month is expected; this reflects the amount to which our ties on IM are concurrent with our physical location. As the percentage of weak or specialized ties increases, the likelihood that they will be physically proximate goes down. However, this relationship is not as evident in the data as is the converse. I would expect to see an equal and opposite correlation with the size of one's buddy list and the percentage of them that they meet offline.

This could mean that instant messaging has multiple uses, and that a high percentage of strong ties elicits one type of use (supporting offline ties), and a lower percentage means it is more for keeping in touch with professional contacts or more specialized relationships. If this were true, the type of ties one communicates with would be a larger signifier of other modalities of communication than the sheer size of the list. Of course this is purely speculation, and without more accurate measures of these relationships it is a fairly shallow interpretation.

Links

In the links section of the survey, subjects were asked to answer questions about links that were extracted from their weblog. As noted in the caveats, this section needs to be carefully answered because of the issues presented by

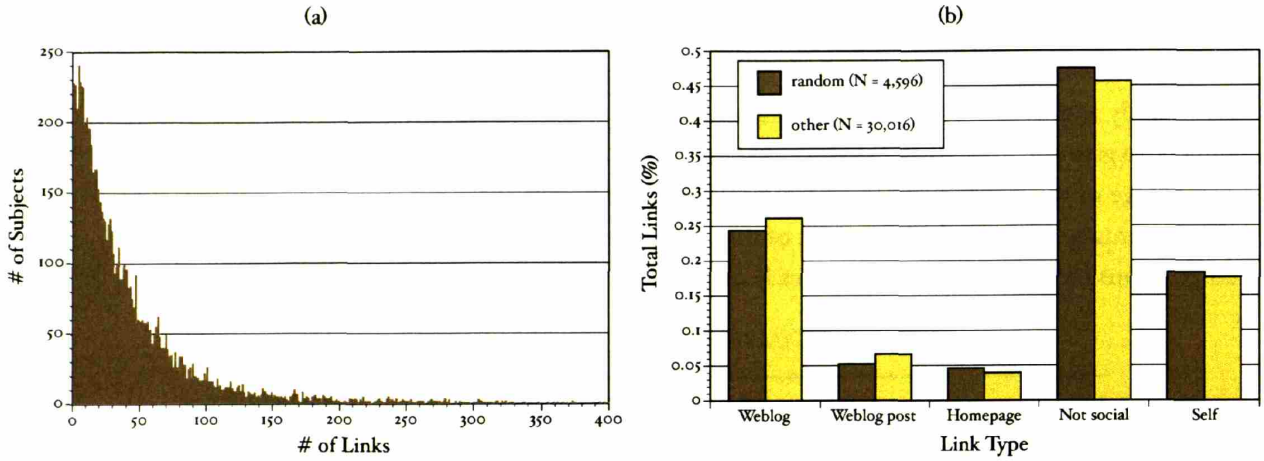


FIGURE 4.17. Observed Links (a) Distribution of links observed per weblog (b) Normalized link types for each sample population

Typically a LiveJournal user would be confronted with the response "no links found on your weblog," despite knowing that links were there.

the survey apparatus. Because of the structural differences between LiveJournal and other weblogs, very few LiveJournal weblogs succeeded in answering these questions† , I will only be considering those subjects in the random and "other" samples for this section.

The other issue with the links section arose from the fact that the categories were not all-inclusive, and a "none of the above" answer was not provided. This will introduce a large amount of noise into the results, and I should note that the analysis provided in this section be interpreted in proviso of these issues.

Figure **??** shows the distribution of links for the random and self-select samples. While the mean number of links per blog is 44.4, the standard deviation is 60 and the median is only 26 links. The second figure shows the classification of these links across the types provided to subjects. The first three categories (weblog, weblog post, personal homepage) are all instances of social links, carrying with them a set of questions about the relationship between the author and the person being linked to. The fourth category (not social) contains any links that the author noted as being in the class "other," while the fifth category are those links that referenced the author's own weblog.

I did not expect to have as many self-referential links as the data shows, almost 20% for both sample populations. Any link made to the same domain as the weblog are not considered, so these represent links that to other web sites that the author has still considered "part of their weblog." Feedback from

users who had difficulties classifying links suggests many utility type links, such as references to the weblog tool, Creative Commons licenses, or personal home pages were put into this category for lack of better groups. I tried to filter as many of these links out during the parsing process, but apparently a number of services fell through. I will return to this issue in the non-social links section.

NON-SOCIAL LINKS

When presented with a link from their own weblog, authors were asked to classify them into either one of three categories of social links, or the more generic case of "other." I will address the more general case of these non-social links first. Of the 62,660 classified as "other," 44,680 were excluded on the basis that they came from either the LiveJournal sample or an incomplete survey. Of the remaining links, 2,489 did not contain answers to one of the two contingent questions; after discarding these incomplete answers, 15,491 links remain in the data set. Finally, after comparing these links with the list of weblogs from the aggregator data, an additional 2,251 links that should have been classified as "weblog" were removed, arriving at a tally of 13,420 links.

For each of these links, two further classification questions were solicited: the source that the subject acquired the link from and a general motivation for putting the link on their site. Unfortunately, I could not find a good typology of internet links, but instead I hoped to find some way of recreating the environment of some of the early news diffusion experiments (Deutschmann and Danielson, 1960; Greenberg, 1964). If I could find those links that were related to diffusing information, I could identify whether or not personal interaction was playing a role in general news diffusion.

For the information source, two options, "can't remember" and "something I wrote," were both considered as excluded by the subject. The remaining sources were chosen based on feedback from pilot subjects and included personal communications, weblogs, bulletin board systems (BBS), news sites, search engines, and a more general category "stumbled upon it" meant to catch the happenstance nature of web surfing. Weblogs, BBSs and personal communications were meant to capture the role of either primary or secondary personal diffusion.

Motivations for posting the link included personal, newsworthy, important, funny and informative. The motivation behind these motivations was to separate timely or potentially diffusing links from those that were merely providing context or more static in nature (such as links to services and

affiliations). The two categories of interest are newsworthy and important, both of which suggest a sense of urgency, while personal and informative suggest a more contextual nature. "Funny" was added after feedback from the pilot study as the most delinquent category based on regular motivations.

TABLE 4.17. Distribution of non-social links

	Motivation for posting				
Source	Personal (%)	News (%)	Important (%)	Funny (%)	Informative (%)
Personal	30.8	10.4	21.0	17.4	17.0
Weblog	19.5	17.2	24.7	31.8	17.1
BBS	1.6	1.7	3.0	3.3	2.8
News site	5.8	44.2	20.3	12.9	11.5
Search engine	29.6	16.6	18.8	17.9	38.6
Surfing	12.6	9.9	12.3	16.8	13.0
N	2111	1241	1174	2178	6154

Table 4.17 shows the distribution of sources grouped according to the associated motivations. "Informative" was by far the largest category, containing almost half of the overall links while "funny" and "personal" followed with about a 15% of the links a piece. The two categories of most importance to this thesis turned out to be the least common. This occurred because the links were sampled from the entirety of external links on the given weblog, which included all of the utility links mentioned earlier. Table 4.18 shows the top ten most frequently listed links, all of which are services or utilities that might be considered functional. Since no such category existed, nearly all of these links were classified as "useful."

TABLE 4.18. Top non-social links

Rank	$\sum$	URL
1.	62	http://validator.w3.org/check/referer
2.	47	http://news.google.com
3.	40	http://www.hello.com
4.	39	http://gmpg.org/xfn
5.	37	http://www.google.com/
6.	35	http://www.flickr.com
7.	24	http://quizilla.com
8.	20	http://wholinkstome.com
9.	19	http://www.theonion.com
10.	17	http://jigsaw.w3.org/css-validator/check/referer

The first non-utility link is the Make Poverty History project ranked at

number 17†. At number 32 is the first news story on a Supreme Court ruling against property owners that was a large story during the survey. Without painstakingly selecting each of these diffusion examples by hand, I will have to rely on the "important" and "newsworthy" categories to determine the diffusing links.

This site is arguably in the utility class of links because it comes from a static badge authors leave on their site.

For links posted for their newsworthiness, it is not surprising that the most common source was news sites. The second most common sources, however, were weblogs, followed by search engines and personal sources. Because sites like Google News and Yahoo News obfuscate the division between search engine, portal, and news, it is hard to tell whether these users actually found the link via searching or through one of these associated services.

The sources for "important" information are almost evenly distributed across personal sources, weblogs and search engines. Given the noisy quality of the data, it is hard to make much of an inference, but I can speculate that people would label a link "important" over "news" and "personal" if the link had strong personal significance. Given this assumption, it is interesting to note that people use a range of sources to find these important pieces of information, and that smaller, more important information might come from personal sources more than mass media. This is in line with Greenberg's observation that the most common types of news to spread through personal ties are both big news stories and small, specialized ones (Greenberg, 1964).

Despite the instructions, the categories in this section were quite ambiguous. While the focus was really on discerning the nature of social links, I hoped to find some salient behavior from a very small set of questions (2 to be exact). Given some analysis of the noise and more specific categories, this apparatus could lend itself to a better understanding of the nature of personal information.

SOCIAL LINKS

Of the 26,075 links listed as social links, exactly 9,700 came from a disregarded sample (LiveJournal or incomplete), and an addition 1,351 were missing data for all of the subsequent questions. After excluding these links, the data includes 15,024 samples of links listed as weblog, weblog post, or personal homepage, with 10,275 (68.3%), 2,632 (17.5%), and 1,637 (10.9%) links respectively.

Removing the personal homepages, the ratio of static to dynamic social links is about 4:1, which is smaller than the rate observed by the aggregator, 2.6:1. I believe this discrepancy stems from two issues: first, the number of dynamic

links observed on all weblogs over the course of a month will aggregate to a larger number than would be found at any given day on the front pages of the same sites. Second, in cases where the subject does not know the author of a given weblog post, they might not see it as such. For instance, if I found a web page through a search engine that answered a question I was posing on my weblog, I might post the link without ever even making a mental note that the content was posted on a weblog.

TABLE 4.19. Social link type and relationship

	Link Type		
Relationship	Post (%)	Weblog (%)	Homepage (%)
No Relation	69.1	55.1	43.4
Acquaintance	12.7	18.2	17.0
Friend	13.8	23.6	30.1
Family	4.4	3.1	9.5

Table 4.19 displays the association between the various link types and the reported relationship between the subject and the other author. As would be expected by the notion that a dynamic link does not necessarily imply any sort of personal interaction, static links are associated with higher levels of acquaintance than dynamic. However, the number of ties identified as having no social basis is remarkably high; over 50% of the links that authors are making are made to weblogs written by individuals with whom the subject does not even consider an acquaintance.

TABLE 4.20. Readership and relationship

	Alter's relation to the author			
Last read	None	Acq.	Friend	Family
Never	4.5	.9	.7	.8
Over a year ago	.8	.9	.7	.2
6 Months–1 Year	2.0	1.8	1.2	1.7
1 month–6 months	8.8	7.4	6.1	3.1
1 week–1 month	21.5	19.4	14.0	15.2
This Week	32.7	33.8	31.6	23.6
Today	29.7	35.8	45.8	55.4

The next question to be addressed is how many of these links are "live," or denote weblogs that the subject reads regularly and how many are "dead," pointing to readership that no longer exists. Because of the higher variability of dynamic links, I will look specifically at static links for this measure, assuming that dynamic ones exhibit a diminished form of the same readership.

Table 4.20 shows the distribution of readership as described by the last time the author read the given weblog for each type of relation to the author.

While I expected to find high readership for friends' weblogs, I was surprised to see that for *all* levels of acquaintanceship over 80% of the identified weblogs had been visited in the last month, and over 60% in the last week. As would be expected, the stronger the social tie described by this link, the more likely the subject is to read them regularly. Over 50% of the familial weblogs were read the day the survey was taken, and nearly 50% for those denoted as friends.

These data raise two important issues. First, given the distribution of update times observed by the aggregator, one cannot expect that all of these weblogs were updated within the period that the subjects specified they had "read" them. With the use of weblog aggregating tools such as RSS readers or services like Bloglines (Bloglines, 2003), one would probably describe "reading" as "being aware of," where the frequency of reading is roughly the same as the frequency of update. Given this effect, the relative measure of live vs. dead is probably just as accurate.

The second issue is that there will probably be some level of bias associated with affirmation of the subjects self-image, namely that they would rather remember having read these weblogs more recently than they may actually have, especially for those that would be considered dead. The division of time periods was explicitly chosen to minimize the generalized bias shown by the pilot subjects, but it has the downside of including large ranges of time. For this reason, the best measure of live and dead links should be described by three stages: active, or within the last week, inactive, within the last month, and dead, longer than one month ago.

Another important part of describing these social links is to determine to what extent they denote other types of social interaction. For instance, I might say that someone I met through blogging is my friend because of the companionship we have shared in our writing online. However I may never have met this person face-to-face or even spoken with them in another medium. The correlations between the various social communication questions are shown in table 4.21. All of the measures are in terms of increasing frequency, except for friendship which is in increasing acquaintanceship.

The strongest correlations are between the level of acquaintance and various forms of communication, which should be expected. These data suggest that the stronger the tie, the more likely it is that a weblog author will use other

TABLE 4.21. Social links and communication

	Relat.	Read	F2F	Commented	Spoken
Relationship	1	.172	.764	.358	.770
Read	.172	1	.145	.438	.228
Face-to-face	.764	.145	1	.270	.697
Commented	.358	.438	.270	1	.419
Spoken with	.770	.228	.697	.419	1

$p < 0.001$ for all measures

forms of communication to interact with their weblog ties. The relationship between tie strength and the frequency of face-to-face interaction implies that in this case weblogs are part of a larger set of communication tools used to support offline ties. This is in accordance with Haythornthwaite and Wellman (1998), suggesting that the stronger a tie is, the higher the modality of interaction.

Along with the observations made in the weblog use section, these data reinforce the idea that there are a number of types of interaction being expressed in the form described as a weblog. While some of the more specialized, professional forms of weblogs can be non-social, the weblog is at its core a social tool, capable of reinforcing local, face-to-face communications, as well as allowing for new connections. The following section will address the extent to which weblogs engendering new social ties.

Social Capital

One of my fears about having a long survey instrument such as the position generator as the last section was that the drop-out rate in this section would be quite high. Furthermore, even though the instructions specified to check either yes or no for each position, I was also concerned that subjects would only check the "yes" answers and leave the "no" answers blank. My apprehension proved incorrect however, as shown in figure 4.18(a). For those subjects that completed the survey, a large majority submitted a value for all 30 occupations.

Figure 4.18(b) illustrates the responses over the entire subject population for each job, in the order that the job was presented in the survey. The gradual decline in answers suggests that as users got bored with the instrument, some percentage dropped out for every question. In analyzing the data, I will only use those subjects who completed at least 28 of the 30 questions; for those

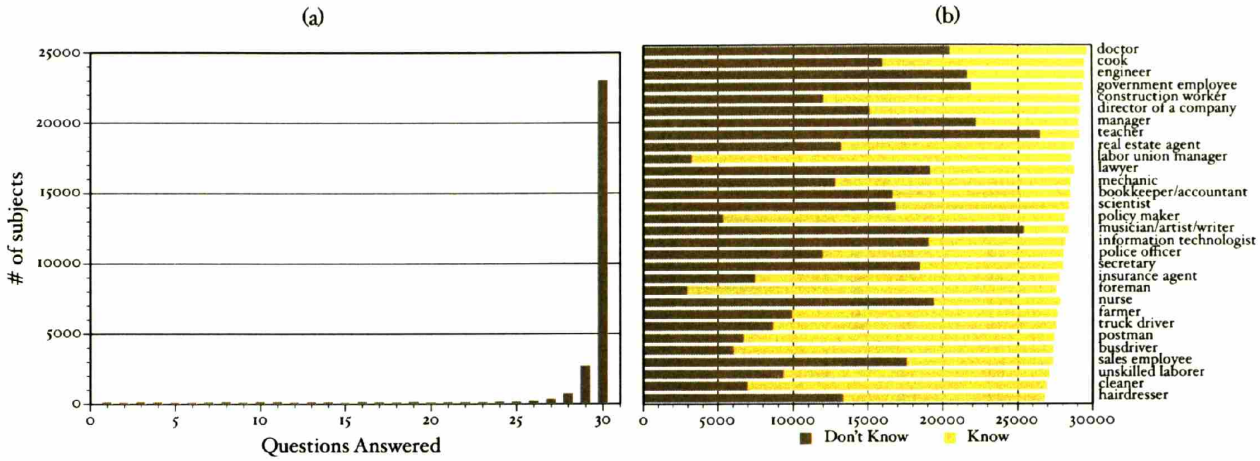


FIGURE 4.18. (a) The number of position generator questions answered per subject. (b) The number of responses per job.

subjects who did not answer one or two questions, it is quite likely that these came as the last two questions. Since these two occupations are among the least known among the entire set, it is safe to assume that an omission of the last one or two questions is equivalent to a "no" answer. Because my measure of total social capital will be the sum of the occupational prestige scores, these left out answers should not affect the data.

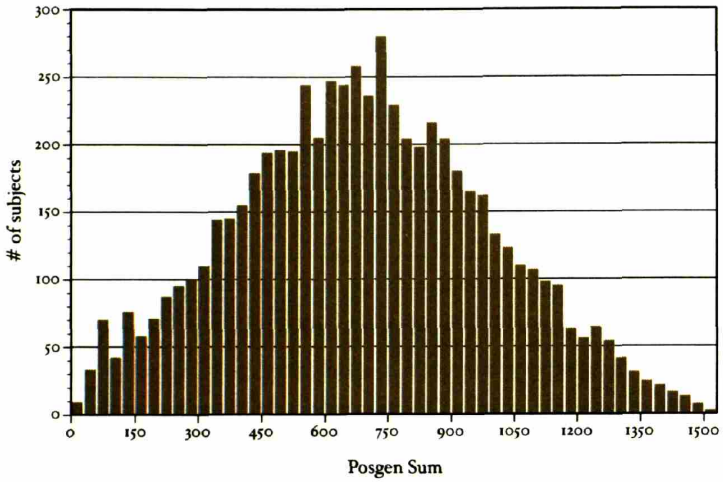
I have chosen to use the sum of occupational prestige scores as my measure of social capital based on the fact that it provides the widest degree of all of possible measures. The distribution of these scores can be seen in figure 4.19. The distribution shows a normal form† with some notable spikes towards the bottom, center, and top of the distribution. The spike at the highest value is probably the result of subjects answering "yes" to all of the questions either to finish the survey or maximize their "score."

The Anderson-Darling score for normality is 37.89 with $p < 0.001$.

The aggregate results for the position generator are presented in table 4.22. If the most common occupations, two stand out far and above the rest of the distribution; while teacher is in the top half of occupational prestige, any number of professions could satisfy this term, from pre-school to university professor. This ambiguous category makes it one of the most regularly known, and diminishes its informational value.

Likewise, the commonly known occupation of musician/artist/writer is probably caused by a misinterpretation of the question. While I was explicitly asking for people who's *profession* is the stated occupation, individuals probably interpreted this as anyone who fills that title. Since many people are

FIGURE 4.19. The distribution of position generator scores calculated as the sum of all occupational prestige scores.



Arguably, almost all weblog authors are amateur writers.

amateur musicians, artists, and writers† , most everyone knows someone who fits the description. The high rates of these two professions can explain the spikes in the lower values of the distribution; knowing only one or both of these professions places a subject in the first and second spikes respectively.

Demographics

As with weak social ties in general, one assumption is that social capital will accrue as a person ages; as we get older, we meet more people, and our overall network of acquaintances grows, along with access to associated resources (Lin, 1999b). Before comparing this measure of social capital to other measures in the survey, I will first look at the relationship to the demographic variables of age, gender and education.

As expected, there is a positive correlation with age and eduction, and a slight negative relationship with sex. Controlling for each of these variables independently does not remove these biases, so it will be necessary to control for all three in the following observations.

ONLINE/OFFLINE

One of the distinguishing factors of this instance of the position generator instrument was the addition of an online vs. offline distinction for each occupation. Knowing that the typically-public nature of weblogging allows

TABLE 4.22. Results from the Position Generator

		prestige		*% yes*		*% if yes*			
Item #	Job	U & S	ISEI	Know	Acq.	Friend	Family	Online	Offline
11	lawyer	86	83	66	42	38	20	21	79
1	doctor	84	87	68	51	28	21	11	89
15	policy maker	82	70	19	55	34	11	21	79
3	engineer	76	68	72	25	45	29	17	83
17	information technologist	68	70	67	25	59	16	31	69
7	manager	67	69	76	37	46	17	16	84
6	director of a company	67	69	51	46	33	20	17	83
10	labor union manager	66	65	11	59	24	17	19	81
14	scientist	65	71	59	34	51	16	26	74
4	government employee	64	61	73	29	40	31	20	80
9	real estate agent	64	61	45	58	26	16	14	86
12	mechanic	63	59	44	47	28	24	11	89
8	teacher	62	66	90	27	48	25	18	82
18	police officer	54	50	42	54	26	20	14	86
19	secretary	52	51	65	43	42	14	20	80
19	insurance agent	52	53	27	57	25	18	16	84
13	bookkeeper/accountant	52	54	58	40	36	24	17	83
16	musician/artist/writer	45	64	89	20	68	12	35	65
22	nurse	44	38	69	34	33	32	16	84
26	bus driver	44	26	22	61	23	16	11	89
30	hairdresser	39	30	49	58	31	11	11	89
2	cook	39	30	53	41	41	18	17	83
23	farmer	36	43	36	38	24	38	11	89
21	foreman	27	25	11	48	24	28	16	84
25	postman	26	39	24	60	21	19	10	90
24	truck driver	26	34	31	44	23	33	16	84
27	sales employee	22	43	64	33	54	12	23	77
29	cleaner	20	29	25	58	28	14	13	87
38	unskilled laborer	15	26	34	42	4	16	20	80
5	construction worker	15	26	41	39	33	28	12	88

TABLE 4.23. Position Generator and demographics

	Age	Education	Sex	PG_{Sum}
Age	1	.484	-.157	.313
Education	.484	1	-.109	.275
Sex	-.157	-.109	1	-.069
PG_{Sum}	.313	.275	-.069	1

N = 29,835 - p < 0.001 for all variables

for happenstance interactions, I was interested to what extent these new acquaintances could potentially be expanding an author's access to occupational resources. If this is true, namely that the process of weblogging increases one's social capital, then I would expect the length of authorship to be correlated with an increased amount of online occupational access. Before I address this measure, there are a few caveats to the online/offline distinction.

First, a few subjects appropriately asked the question, "how am I supposed to know the professions of my online acquaintances?" This inquiry is related to an even bigger question, which is whether or not one can extract resources from online social ties without knowing whether or not they exist. The survey was meant to elicit the potential of an individual to extract resources, and understanding how large groups of anonymous individuals can pool social capital should be the subject of another survey unto itself.

Also, the definition of "know" for online ties was flawed. The survey defined knowing as, "if you saw this person on the street somewhere you could remember their name and start a conversation with them." Even if two people had met online, considered each other friends, and were aware of each other's respective occupations, there is a potential that they would not be able to recognize each other offline. The fact that "knowing" is defined by an offline interaction biases the subject's memory towards offline ties.

Finally, the question was posed as a dichotomous answer. To disambiguate the case that the subject knew multiple individuals in the same occupation, the survey specified that they should choose the individual they "communicated the most with," as a measure of tie strength. I chose this language because I knew that my other definition of tie strength ("someone you feel especially close to") would bias the results towards offline interactions; communication would select for the individual they were most acquainted with currently.

TABLE 4.24. Online and offline Position Generator scores

	Investment	M_{prof}	M_{pers}	$Comm^T$	PG_{on}
PG_{on}	.226	.084	.144	.169	1.000
PG_{off}	-.001	.057	-.007	.197	1.000

N = 26,360. Control variables: Age, Gender, Education and PG_{on}/PG_{off} when observing the opposite variable. $p < 0.001$ for all correlations.

Given all of these caveats, the relationship between the online position generator sum PG_{on} and the offline sum PG_{off} are shown in Table 4.24. In early analyses, I looked at the total sum PG_{sum}, but this value seemed to only

be correlated with increases in either the online or offline scores, so I have chosen to present them instead. For each measure, I have controlled for age and gender, and to account for the tradeoff between the two, I have controlled for PG_{off} when looking at PG_{on} and vice-versa. These variables are shown with respect to the aggregate measures derived in the weblog and communication sections.

In the case of both online and offline position generator scores, the total communication frequency $Comm^T$ is correlated with increased access to occupational resources, and even more so for offline than online ties. The largest discrepancy comes from the amount of total investment that the subject puts into his or her weblog; those individuals that invest a lot of time and energy into the practice are associated with a higher number of online access to occupational resources. The same is ***not*** true for offline ties, as an increase in weblog investment has no effect on offline relationships whatsoever.

The other peculiarity of these two measures is the distinction between M_{prof} and M_{pers}; while the difference is not overwhelming, there is a discrepancy between the various social capital measures and these two types of motivation. Those who are blogging for professional reasons tend to have a slightly higher offline social capital while online social capital is correlated with both, and more so for those motivated by personal reasons.

Unfortunately, there was no strong relationship between the amount of time that a subject had been blogging and PG_{on}, nor was there any with the age of their current weblog. There were correlations with audience size, but this interaction was defined by the overall investment shown above.

Despite all of the caveats, I was not expecting to find these interdependencies at all. Without a longitudinal survey it is impossible to definitively say anything about this relationship; any number of other variables could be the source of both weblogging behavior and increased online social capital. However, the size of this correlation reinforces the need to study online social capital formation in a more rigorous form, and with respect to any number of communication technologies. It also remains to be seen whether or not these individuals can actually extract the resources implied within the position generator.

Chapter 5

Conclusions

This thesis has covered covered a broad range of topics with an equally varied number of methodological approaches. This chapter will serve as a summary of my empirical results and the associated theoretical implications.

5.1 SUMMARY

In this thesis I have observed the group of web authors known as *webloggers*, a community that engenders both online, social interaction and also the diffusion of information. Two different study tools were used to further my understanding of their social practices: first, a weblog aggregator collected data on the weblogs updated during the study period, extracted information about their relationships, both implicit and explicit, and tracked information as it spread across this structure. Second, a general social survey of this community was performed to both validate the data from the aggregator and better understand the makeup of the community.

Aggregator

The data collected by the aggregator provided a rich network of relationships that could be used to understand the possible social structure within the community. The extracted network consisted of over 300,000 nodes and 1.7 million edges; the distribution of these edges as represented by in-degree followed a power law, suggesting that a large percentage of the attention within the community was governed by a few select weblogs.

Unlike the small-world network that I expected to find, the weblog readership network has a very strange composure. The characteristic path length, or

degrees of separation, was well over 8 degrees, which reflects either narrow bridges across language barriers, or an extremely distributed network.

The relational ties between weblogs were broken into two categories: first, those links explicitly and directly to another weblog, and second those made implicitly in the course of interaction. These two measures were shown to produce different distributions of authority, despite the fact that the degree distributions appeared to be the same. The fact that different authors are popular in an explicit sense, and others in an implicit one suggests that these link types refer to different social processes: one is the expression of affiliation while the other interest or attention.

Because the assumed model of network growth within the community (preferential attachment) is not able to explain the dynamic nature of the implicit social structure within the community, I have presented an alternative model, *Dynamic Affinity*. Instead of relying on the addition of new nodes to the graph to create the necessary scaling, dynamic affinity relies on an affinity model distributed as a power law. In this way the network can constantly be rewired with the distribution of degree remaining intact, while new players can enter the network and gain prestige based on their place in the affinity distribution. Within the weblog community, the affinity model is based on the frequency of update, evident from its strong relationship with the distribution of in-degree.

Also, the diffusion of information across this network was addressed. Over 3,000 large-scale media events were observed by the aggregator during the study period. These events were shown to have adoption patterns very similar to those shown by previous studies of innovations, including some appearing to be comprised of external diffusion, some internal, and some a mixture of both. The mixed-model of diffusion was fit to these events using nonlinear regression, and the values of both internal and external growth extracted for each example.

As expected, the distribution of internal and external influence on diffusion contained three distinct clusters: internal-only, external-only and a mixture of the two. However, upon examining the resulting classifications, it became apparent that many of those classified as internal were controlled by external events, and those seen as external could actually be entirely structural. These perplexing results suggest that for the class of diffusion being studied, namely the diffusion of information, the controlling factor in growth may be more related to the type of information and less to the structural effects.

Looking at measures of perceived contagion, it was apparent that very little of

the weblog diffusion could be described by first-degree interactions. Only about 30% of individual weblogs' links could be shown as coming from a readership tie; the resulting analysis suggests that authors do not typically link to the things their neighbors link to.

Survey

In addition to the data collected by the aggregator, I have employed a social survey to investigate some of the social features of this community, as well as to validate the inferences I have made in the aggregator's observations. The survey was divided into 5 sections, covering demographics, weblog use, communication use, the meaning of links both social and non-social, and finally social capital. The survey sample consisted of both a randomly selected group identified by the aggregator, and a self-selected sample of weblog authors at large.

The response of the survey was good, with about a 30% response rate for the random sample, and over 30,000 respondents self-selected. Because such a large percentage of the authors came from the service LiveJournal, the sample was split into three parts: random, self-selected, and LiveJournal exclusively. The demographics of all samples reflected a well-educated group with about 55% women and a mean age of about 27 years. In all samples the authors came from a number of different English-speaking countries.

The section on weblog use produced a good measure of the motivations for producing a weblog: most authors were divided between professional and personal goals. While professional authors tended to use weblogs to increase their professional reputation, create newsletters, and make money off of advertising, personal authors typically wrote more about their own lives, posted more media, and used their blogs to keep in touch with friends. Both groups used their blogs to comment about current events and also to keep a record of things they had read. While most weblogs contained some component of personal use, a much smaller percentage were used to professional ends.

Probably the most important contribution to understanding this community was the observation of a strong relationship between investment in the weblog and payoff in terms of audience size and feedback. Five measures of time invested were consolidated into one variable which was shown to correlate very strongly with both the self-reported audience size, in-degree as observed by the aggregator, and the number of comments received on their

personal site. This also supports the model of dynamic affinity described above, as the distribution of investment across the community is certainly not uniform; these data show that the weblog community rewards the author who puts time into their work, and that the length of one's blogging history does not solely determine their future audience. In this case, it is the hard-workers who get richer, not the previously-rich.

After controlling for the demographic variables of age, sex and education, the communication patterns of weblog authors suggested that an increase in the frequency of communication implies an increase across all modalities. While phone and email were correlated with an older audience, and both instant messaging and text messaging with a younger one, when these age effects were removed, all modes of communication were correlated. For those subjects who specified the breakdown of their instant messaging buddy list, the data demonstrated that two uses of this medium were present: one, to support local, offline relationships, and the other to keep in touch with more distant acquaintances.

Another important section of the survey dealt with characterizing the the types of links that weblog authors make. These were divided into two categories: social, or those made to other weblogs or personal homepages, and non-social, everything that remains. The questions asked about non-social links were quite limited, given the range of possible links that could exist. A large majority of the links observed were characterized as "Informative," usually relating to services or tools that the author used. The second largest categories of links were "Funny" and "Personal," both of which had a large percentage of sources that suggest interpersonal contagion.

In looking at the social links, I hoped to ascertain the social basis and the extent to which they implied interaction. Over 50% of the "social" links described were actually not social at all, as the relationship between the subject and the linked author was specified as "no relationship." For weblog posts, this percentage was even larger, implying that links to posts are about readership, not dialog. The frequency of recent readership implied by these links was much higher than I anticipated, with over 60% of the subjects having read the other weblog within the last week. This suggests that the links do in fact imply readership, but readership is not indicative of further social relations.

The final section of the survey dealt with the social capital of subjects as measured by their access to occupational resources. While social capital was related to the expected demographic variables (age and education), the survey

revealed a relationship between those ties formed online and the author's investment into the medium of weblogging. While overall communication frequency was correlated with both online and offline social capital, investment in the weblog was quite strongly related to online access to resources through online ties. While I cannot make the definitive claim that time spent weblogging leads to social capital formation, the relationship clearly exists and warrants further investigation with a longitudinal study.

5.2 CONTRIBUTIONS

The major contributions of this thesis come in a number of different areas including methodology, theory, and the empirical results described above.

Theory

Two models were presented in this thesis, one to describe the structure of dynamic readership among authors, and one to describe the model of contagion contained therein. The first, ***dynamic affinity*** is the first model to describe a dynamic network topology that repeatedly produces a distribution that scales. While it is simplistic in its definition, it confirms that power laws need not arise in graphs from structural properties alone.

The mixed-model of diffusion was shown to be invalid in the context of media diffusion, and logistic growth was not associated with "internal" or structural propagation. In the context of frequent communication within large networks, media events tend to have purely exponential growth, regardless of how much structure is involved.

Methodology

The data collected by the weblog aggregator present an entirely new range of methodological issues in studying the structural components of diffusion. A data set with near millions of actors and millions of events requires rethinking many of the standard measures used to analyze social network data. Many of the measures employed today by researchers are computationally intractable for data sets of the magnitude presented here. In this thesis I have shown methods for working with large data sets, including the use of measures from related domains.

Measures of centrality are crucial to the study of any type of whole-network data. Even the best algorithms for betweenness or closeness centrality do not scale to nearly the scale of network shown here Brandes (2001). While they have not been employed in previous social network research, flow-based graph algorithms (Brin and Page, 1998; Kleinberg, 1998) are a novel approach to measuring centrality, and scale to fit data sets much larger than even I have addressed.

In the domain of social capital measurement, I have introduced a new measure of occupational resource access that employs a contextual feature measuring the modality of these ties. This Online/Offline Position Generator allows researchers to effectively gauge the extent to which an individual's resources are embedded in online relationships, and the extent to which they are offline. This is a preliminary study, and the first time this instrument has been used, and thus it comes with a number of caveats; this said, the Position Generator can be seen as an important tool for understanding online social networks.

A final contribution to the methodology of online research was presented in the validation of aggregated behavior with survey data. Many studies of web sites and social activity therein make assumptions about words such as "social," "community," and "communication." The two-armed analysis provided by automatically collected data and user surveys should be an important part of any online social analysis.

5.3 FUTURE WORK

The largest limitation to the data presented in this thesis is the limited time span for which it was collected. Without sub-sampling the aggregator data, it would have been difficult to scale the analysis to two or three times the number of events analyzed. Another approach would be to only considering those events that reached an increasingly large population. Since the distribution of these events follows a power law, it is simple enough to increase this bar as the sample becomes larger.

Many individual pieces of the survey warrant a follow-up exploration. The survey itself was primarily meant to reinforce the findings of the aggregated data; the fact that so many of the sections provided rich and interesting results only makes the medium of weblogging more interesting, and more worthy of study. The results of the social capital section need to be revalidated and tested in a longitudinal study to be confirmed.

Sometimes when you scratch flaking paint, you expose a wall you didn't know was there; sometimes the entire wall falls down on top of you. My experience with weblogs certainly falls in the latter camp. What I thought was an interesting side project has been transformed into four years of scholarly research. If every academic could be so lucky as to pick up on a massive sociological phenomenon, there would be no end of things to study.

Appendix A

Weblog aggregator

When this thesis work started, keeping up with the entire universe of weblogs was not such a difficult task; with only about 30,000 weblogs known at the time, a simple program could fetch and store them all comfortably within a day's time. Since then, the number of weblogs has grown exponentially, and current estimates put the total number of weblogs at around 10 million. Suffice to say, keeping up with weblogs is not such an easy task anymore.

The current version of the weblog aggregator performs a number of these tasks in parallel on a few different machines in order to keep up with the 2-3 weblogs that must be crawled and indexed per second. The basic architecture of this system is shown in Figure A.1. To describe the system in detail, I will assume a weblog has just been updated and follow it through the system as it gets indexed.

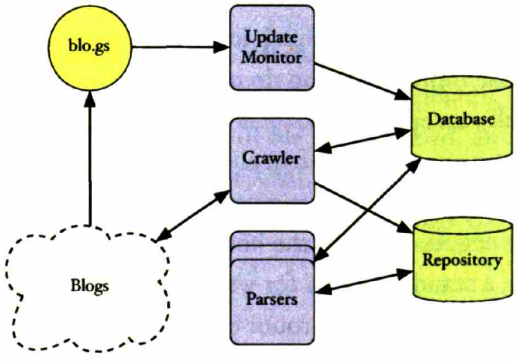


FIGURE A.1. Weblog aggregator system architecture

Update Monitor

Blo.gs provides a real-time push-service for acquiring ping data, sending an xml stream of pings as they are received by the system. For each weblog update, blo.gs provides the URL, a description string (usually a title), the URL of an XML representation if one exists, and a timestamp of unknown source.

UPDATE MONITOR

My *Update Monitor* remains connected to this stream, recording each weblog ping to my database, along with the time at which I observed the ping. Since the timestamp information tended to vary quite a bit from blo.gs (including times in the future and times far in the past), I used my observed time instead.

In addition to using blo.gs for my source of updates, it also doubles as my source of weblogs; every time I encounter a new weblog it is added to my database of weblogs. This means my sample will consist only of those weblogs that have pinged blo.gs in the time period of the study, but it normalizes my data and allows us to assume that my entire data set is accurate to within the error inherent in blo.gs. After these updates are stored in the database, the next phase of aggregation occurs when my crawler fetches the document.

Crawler

WEB CRAWLER

A *web crawler*, or web robot, is simply a computer system that takes a set of URLs, fetches them, and then performs some operation on them (typically storing them somewhere). Web crawlers are typically used to recursively store sets of documents based on the hypertext links contained in the documents they fetch (Burner, 1997; Heydon and Najork, 1999; McBryan, 1994). For my given task, the set of URLs to fetch has been specified (by the update monitor), and the task of the crawler is simply to fetch them as quickly and efficiently as possible.

There are two major constraints in crawling web sites: first, one must respect web servers and not impose too much load at any given time, and second, one must respect those individuals who do not want to be crawled. On the first issue, the generally accepted protocol is to avoid fetching pages from any given web host at a rate of more than about one per second. This standard has been adopted by all major search engines (Google, Inc., 2005), and any faster access should be negotiated with the host provider on an individual basis. On the second issue, a standard exists for web server administrators and web page authors to restrict which pages a robot can crawl. This is facilitated through either a file named `robots.txt` or through a command within the contents of the web page itself (Koster, 2005).

The system was written from the ground up using basic sockets in the Python programming language, allowing for extremely low-overhead interactions with web servers. Using the low-level Linux `select` library, it is able to maintain many simultaneous connections to remote servers without being dependent on the speed of any particular connection. In independent tests the crawler was able to fetch and store upwards of a million pages per day, well over my needs for the weblog aggregator. It also observes the HTTP 1.1 protocol (Fielding, Gettys, Mogul, Frystyk, Masinter, Leach, and Berners-Lee, 1999), including support for compression, which saves both parties on the amount of bandwidth consumed.

For a given weblog, there are a number of different analogous data sources that can be used to obtain the same content: the HTML front page, the archives, an RSS XML file, and possibly an Atom XML file. Because of the overlap in these different data types, I choose to crawl the one that provides the greatest information at the lowest cost. There are three different scenarios I use in deciding which pages to crawl:

XML

When XML files are available, the HTML is stored on the first crawl, but the XML is fetched on future updates, as they are both smaller and easier to process than the HTML counterpart.

INCOMPLETE XML

If XML files are available, but do not contain the full content of posts, the HTML front page is stored initially, then XML files are crawled on future updates. From the XML I obtain links to the individual archived posts which are fetched and stored.

HTML ONLY

If a weblog does not provide an XML alternative, I simply crawl the front page each time it is updated. Each time a weblog is updated, the system makes a determination of which type of file to download. In the event that a weblog adds support for a new type of XML, or if such an XML file disappears, it will adopt a new crawling pattern to reflect the change.

Repository

Once a weblog document has been downloaded from a remote server, the crawler must store the document somewhere. Because my systems need access to weblog content over the network, I have engineered a networked file repository, allowing for the storage and retrieval of any type of data. The repository has a server which runs on a machine with my primary storage device (one 80-gigabyte drive) and accepts connections from remote programs.

SITE-ID (SID)

Data is stored in the filesystem according to a unique identifier called the *site-id (SID)* associated with every weblog. Each weblog has a directory within the repository within which various files can be stored. Certain file types are also allowed to have multiple versions automatically archived, so that comparisons can be made between various instances. When a client connects to the repository, it must specify the weblog's SID, the type of file, and the archive number if one exists. The types and number of files are not constrained, so clients can store their state for a given weblog, or information that may be useful to other clients.

All of the files are stored in one directory tree in a ReiserFS file system, which can easily handle the hundreds of thousands of directories and millions of files necessary for the system to operate. Every file which is stored on the system is compressed before it is stored, then decompressed before it is retrieved. Since most HTML compresses quite well, one month of weblog content for over half-a-million sites adds up to a mere 10 gigabytes of data (including all of the associated files).

Parser

Once the source of updated weblogs have acquired and stored locally, they are analyzed by another program to extract relevant information from the source.

LINKS

This process happens in three individual stages: first hypertext *links* are extracted, sorted into *internal* and *external* links, and stored in respective database tables.

META-INFORMATION

Next, relevant *meta-information* is identified and stored; this includes references to other versions of the weblog (e.g. XML), the title, author's name, geographic coordinates, and email addresses. The availability of these data vary widely across weblogs, and are stored in the case that they are found.

LANGUAGE DETECTOR

In the final stage of parsing, I run a stochastic, trigram-based *language detector* on the natural language of the weblog. These data will be necessary when

gathering subjects for the survey component of my analysis. The language detector is based on the Languid system developed by Maciej Ceglowski (cite maciej). It has been ported to Python for better interoperability with the rest of the system. The training files provided support the detection of 26 different languages. When the accuracy of the language detection is above a certain threshold, the language of the weblog is stored in the database.

Database

For my database I have chosen to use a standard relational database provided by MySQL. This choice was made with the knowledge that the software would support my needs for the extent of my study period. For more details, the data schema is provided with the source code.

Availability

All of the aggregator code is released under an MIT License, and is thus available to anyone who wishes to use it. The license is as follows:

> Copyright (c) 2004-5 Massachusetts Institute of Technology
>
> Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:
>
> The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.
>
> THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

To obtain a copy, please contact me at `cameron@media.mit.edu`.

Appendix B

Email

Following is the email that was sent to the weblogs randomly selected to participate in the survey.

```
Hi there,

The Massachusetts Institute of Technology is
conducting an important academic study regarding
weblogs. We are investigating the role of weblogs in
the lives of their authors. Does your weblog make you
more connected to the rest of society? Does it
increase your chances of getting a job or finding
information that you are looking for? To answer these
questions, which are very important to our research,
we ask for your help.

Your weblog has been randomly selected as part of a
small group among millions to represent the entire
community of weblog authors. We obtained this email
address from what we assume is your weblog:

<URL>

If this is NOT your weblog, we would appreciate it if
you could tell us by clicking on this link:

http://blogsurvey.media.mit.edu/optout?e=<email>

This is not a commercial marketing survey, but
scholarly research to be used in academic publications
such as journals, conferences and books. By
participating in this study you will be assisting
scientific research and contributing to a better
understanding of how weblogs are influencing people's
lives. We would like you to please fill out our short
survey (about 15 minutes or less, on average).

To take the survey, either follow this link:

http://blogsurvey.media.mit.edu/?e=<email>\&k=<key>
```

or use the following information to log in:

http://blogsurvey.media.mit.edu
email: <email>
key: <key>

Your help is greatly appreciated. If you have any questions or concerns, feel free to contact me, or consult the link at the end of this email.

Yours Sincerely,
Cameron Marlow
MIT Media Laboratory

http://blogsurvey.media.mit.edu/help

Appendix C

MIT Weblog survey

The following is a textual representation of the survey used in this thesis. The actual visual representation looked very different; to obtain a copy, please send email to cameron@media.mit.edu.

Demographics

Instruction: This survey has 6 total sections which should take less than 15 minutes to complete all together. Thanks for your help in this important survey!1. What is your gender?

- Male
- Female

2. In what year were you born? Example: Four digits, e.g. "1976"
3. What is the highest level of education you have completed?
   - Less than High School
   - High School/GED
   - Some College
   - 2-Year College Degree (Associates)
   - 4-Year College Degree (BA, BS)
   - Master's Degree
   - Doctoral Degree
   - Professional Degree(MD, JD)
4. What is your country of residence?
5. What is your current marital status?
   - Single/Never Married
   - Married

- Separated
- Divorced

6. What is your race/ethnicity?

- Black/African American
- White/Caucasian
- Asian/Pacific Islander
- Spanish/Hispanic/Latino
- Native American/American Indian
- Mixed/Multi-racial
- Other

7. What is your current 5-digit zip code? Example: e.g. 02139

Links on your weblog

Instruction: For the following section we need to access your weblog and collect some information about the links you have made recently. After you click "submit," 5 random links will be selected from *your weblog*, and you will be asked to answer a few simple questions about each. The links will be provided in the context in which they were found on the page.

For the *link type* we ask that you use the following definitions: *weblog* is the front page of another person's weblog, *weblog entry* is a link to an individual entry on a weblog, *news story* is a story on a professional news website, *personal home page* is another person's non-weblog homepage, *web service* is a tool such as Google, and *other* is anything else.

If the weblog in question has multiple authors, please answer the questions about the one you know the best

Please enter your weblog URL below:

L1. How would you classify the link above?

- Part of my weblog
- Weblog
- Weblog entry/post
- Personal Homepage
- News story
- Web service
- Other

L-S1. When did you last read this weblog

- Today
- This week
- This month
- Past 6 months
- Past year
- Over a year ago
- Never

L-S2. When did you last post a comment on this weblog?

- The site doesn't have comments
- Today
- This week
- This month
- Past 6 months
- Past year
- Over a year ago
- Never

L-S3. When did you last meet the author in person?

- Today
- This week
- This month
- Past 6 months
- Past year
- Over a year ago
- Never

L-S4. When did you last speak with them online (IM, email, etc.)?

- Today
- This week
- This month
- Past 6 months
- Past year
- Over a year ago
- Never

L-N1. How did you first hear about this story/website?

- Someone told you about it

- Saw it on another weblog
- Saw it on a bulletin board
- Saw it on a news site
- Found it through a search engine
- Stumbled upon it
- Can't remember

L-N2. What would you say was the main motivation for posting it?

- Personal/Related to you
- Newsworthy/Urgent
- Important/Influential
- Funny/Amusing
- Informative/Useful
- No reason

Communication Use

Instruction: This section deals with your daily communication with friends, family and at work.16. How many distinct people do you write emails to on an average weekday?

- None
- 1 - 4
- 5 - 9
- 10 - 24
- 25 - 49
- 50 or more

17. How many distinct people do you have instant messenger (IM) conversations with on an average weekday?

- None
- 1 - 4
- 5 - 9
- 10 - 24
- 25 - 49
- 50 or more

18. How many distinct people do you send text messages (SMS) to on an average day?

- None

- 1 - 4
- 5 - 9
- 10 - 24
- 25 - 49
- 50 or more

19. How many distinct people do you have phone conversations with on an average weekday?

- None
- 1 - 2
- 3 - 5
- 6 - 10
- 11 - 20
- 21 or more

Instruction: For the following six questions we will ask you some details about your instant messenger client. Feel free to look at your list of buddies if you need to. If you do not use instant messaging, you can skip this section.20. How many members do you have in your buddy list?

Example: If you have twenty buddies, please enter "20"21. Approximately what percentage of your buddy list are family members?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- 100%

22. Approximately what percentage of your buddy list are close friends, i.e. someone you would feel comfortable borrowing money from?

- Less than 10%
- 10%
- 20%
- 30%

- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- More than 90%

23. Approximately what percentage of your buddy list are business-related, i.e. someone with whom you only talk about professional matters?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- More than 90%

24. Approximately what percentage of your buddy list are people you meet in-person at least once a month?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- More than 90%

Weblog Use

25. When did you first start weblogging? Example: (e.g. for April 2005, please use 04-2005)

26. How many weblogs have you been an author on in the past year?

- None
- 1
- 2
- 3
- 4
- 5 or more

27. How many weblogs would you estimate you have posted comments on in the past year?

- None
- 1-5
- 5-10
- 10-25
- 25-100
- 100 or more

28. How many hours per week would you estimate you typically spend using weblogs (reading, writing, or commenting on)?

- Less than one hour
- 1-2 hours
- 2-5 hours
- 5-10 hours
- 10-20 hours
- 20 or more hours

29. How many weblogs would you estimate you read on a given day?

- None
- 1-5
- 5-10
- 10-25
- 25-100
- 100 or more

Instruction: For the following questions, please answer them about your primary weblog, i.e. the weblog you post the most to, or most consider your

own. If the weblog is not currently active, please answer about the last time you updated it.

30. When did you create the weblog? Example: (e.g. for April 2005, please use 04-2005)

31. What would you say is the primary reasons for writing to this weblog? Check all that apply.

- Keep a list of links to things you have read
- Keep in touch with friends
- keep notes for myself or record what's going on in my life
- Comment about things I read in the news
- Meet new people
- Keep notes for my professional interests
- None of the above

32. How many times has it changed locations (i.e. changed your blog service or host, giving it a new URL)?

- Never
- 1 time
- 2 times
- 3 times
- 4 times
- 5 or more times

33. How often do you typically post to the weblog?

- Many times per day
- A few times per day
- Once a day
- A few times per week
- Once a week
- A few times per month
- Once a month
- Less than once a month

34. How often does the weblog typically receive comments (excluding spam or comments you have written)?

- Many times per day
- A few times per day
- Once a day

- A few times per week
- Once a week
- A few times per month
- Once a month
- Less than once a month

35. How many authors does the weblog have?

- One (you)
- 2
- 3 - 5
- 5 - 10
- 10 - 25
- More than 25

36. What percentage of your weblog posts would you say are about personal matters?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- More than 90%

37. What percentage of your weblog posts would you say are about the news, current events, or things you think are newsworthy?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%

- 90%
- More than 90%

38. What percentage of your weblog posts would you say are about topics relating to your profession?

- Less than 10%
- 10%
- 20%
- 30%
- 40%
- 50%
- 60%
- 70%
- 80%
- 90%
- More than 90%

Social Networks

Instruction: Last section! Here we will ask you about a number of different occupations. For each one we are interested in whether or not you know someone who holds that job. By ***knowing*** we mean that if you saw this person on the street somewhere you could remember their name and start a conversation with them.

If you do know such a person, we would also like you to specify what your relationship to this person is. By ***friend*** we mean someone whose house you could stay at if you needed to, and ***family*** being someone in your extended family. ***Acquaintance*** should be anyone else. We also would like you to specify if you were introduced to this person ***online*** (over email, instant messaging, bulletin board, etc.) or ***offline*** (in person).

If you know multiple people in one profession, please answer the questions about the individual you communicate with the most.

For each profession, the subject was asked:

Do you know someone who is a/an (No, Yes) Relationship? (Acquaintance, Friend, Family) Introduced? (Online,Offline)

Professions: doctor, cook, engineer, higher civil servant, construction worker, director of a company, manager, teacher, real estate agent, trade union

manager, lawyer, mechanic, bookkeeper/accountant, scientist, policy maker, musician/artist/writer, information technologist, police officer, secretary, insurance agent, foreman, nurse, farmer, truck driver, postman, bus driver, sales employee, unskilled laborer, cleaner, hairdresser,

Bibliography

ADAMIC, LADA, AND EYTAN ADAR. 2003. Friends and neighbors on the web. *Social Networks* 25(3):211–230.

ADAR, EYTAN, LADA ADAMIC, LI ZHANG, AND RAJAN M. LUKOSE. 2004. Implicit structure and the dynamics of blogspace. In *Workshop on the weblogging ecosystem at the 13th international world wide web conference.* New York.

BARABÁSI, ALBERT-LÁSZLÓ. 2002. *Linked: the new science of networks.* Cambridge, Mass.: Perseus Pub.

BARABÁSI, ALBERT-LÁSZLÓ, AND RÉKA ALBERT. 1999. Emergence of scaling in random networks. *Science* 286(5439):509–512.

BARLOW, JOHN PERRY. 1996. A declaration of the independence of cyberspace. http://www.eff.org/~barlow/Declaration-Final.html.

BARRETT, CAMERON. 1998. Anatomy of a weblog. http://www.camworld.com/journal/rants/99/01/26.html.

BASS, FRANK. 1969. A new product growth for model consumer durables. *Management Sciences* 15(1):215–227.

BLOGHOP. 2000. Bloghop. http://www.bloghop.com/.

BLOGLINES. 2003. Bloglines.com. http://www.bloglines.com.

BLOOD, REBECCA. 2000. Weblogs: A history and perspective. http://www.rebeccablood.net/essays/weblog{_}history.html.

BOLLOBÁS, B. 1985. *Random graphs.* London: Academic Press.

BOORMAN, S. A., AND H. C. WHITE. 1976. Social structure from multiple networks ii. role structures. *American Journal of Sociology* 81(6):1384–1446.

BOSNJAK, M., AND T.L. TUTEN. 2003. Prepaid and promised incentives in web surveys: An experiment. *Social Science Computer Review* 21(2):208–217.

BOURDIEU, PIERRE. 1985. *Handbook of theory and research for the sociology of education*, chap. The forms of capital, 241–258. New York: Greenwood.

BOYD, DANAH. 2002. Faceted id/entity: Managing representation in a digital world. Department of Media Arts and Sciences, Massachusetts Institute of Technology.

———. 2004. Friendster and publicly articulated social networks. In *Proceedings of the ACM Conference on Human Factors and Computing Systems (CHI 2004).*

BRANDES, ULRIK. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25(163).

BRIN, SERGEY, AND LARRY PAGE. 1998. The anatomy of a large-scale hypertextual search engine. In *Proceedings of the ACM Conference on the World Wide Web.*

BURNER, MIKE. 1997. Crawling towards eternity: Building an archive of the world wide web. *Web Techniques Magazine* 2(5).

BURT, RONALD S. 1987. Social contagion and innovation: Cohesion versus structural equivalence. *The American Journal of Sociology* 92(6):1287–1335.

———. 1992. *Structural holes: The social structure of competition.* Cambridge, MA: Harvard University Press.

CEGLOWSKI, MACIEJ. 2002. Blog census.
`http://blogcensus.net/.`

———. 2005. Languid: A language identification system.
`http://languid.cantbedone.org.`

COLEMAN, JAMES S. 1988. Social capital in the creation of human capital. *American Journal of Sociology* 94(Issue Suppliment: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure):S95–S120.

COLEMAN, JAMES S., ELIHU KATZ, AND HERBERT MENZEL. 1966. *Medical innovation: A diffusion study.* New York: Bobbs-Merrill.

CORMEN, LEISERSON CHARLES E. RIVEST RONALD R. STEIN CLIFFORD. 2001. *Introduction to algorithms.* 2nd ed. MIT Press.

COUPER, MICK. 2000. Web surveys: A review of issues and approaches. *Public Opinion Quarterly* 64:464–494.

CYWORLD. 2005. Cyworld.
`http://www.cyworld.com.`

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 (Statistical, Nonlinear, and Soft Matter Physics)* 65(5):055103.

DILLMAN, DON A. 1978. *Mail and telephone surveys: The total design method.* New York: Wiley.

————. 2000. *Mail and internet surveys: The tailored design method.* New York: Wiley.

DIMAGGIO, PAUL, ESZTER HARGITTAI, W. RUSSELL NEUMAN, AND JOHN P. ROBINSON. 2001. Social implications of the internet. *Annual Review of Sociology* 27(307-336).

DONATH, JUDITH. 1997. *Communities in cyberspace: Perspectives on new forms of social organization*, chap. Identity and Deception in the Virtual Community. Berkeley: University of California Press.

EAGLE, NATHAN. 2005. Machine perception and learning of complex social systems. Department of media arts and sciences, Massachusetts Institute of Technology.

EATON, BRIGITTE. 1999. Eatonweb portal.
`http://portal.eatonweb.com`.

EROI INC. 2004. Monday [is the] best day to send email.
`http://www.eroi.com/news/email-marketing-report-monday-101204.htm`.

EXACTTARGET INC. 2004. Myth debunked: No such thing as a bad day.
`http://email.exacttarget.com/pdf/Best-Day.pdf`.

FELD, SCOTT. 1982. Social structure determinants of similarity among associates. *American Sociological Review* 47:797–801.

FIELDING, R., J. GETTYS, J. MOGUL, H. FRYSTYK, L. MASINTER, P. LEACH, AND T. BERNERS-LEE. 1999. Hypertext transfer protocol - http/1.1. Tech. Rep., The World Wide Web Consortium.

FISCHER, CLAUDE S. 1982. *To dwell among friends: Personal networks in town and city.* Chicago, IL: University of Chicago Press.

FISCHER, CLAUDE S., AND LYNNE MCCALLISTER. 1978. A procedure for surveying personal networks. *Sociological Methods and Research* 7:131–148.

FLAKE, G., S. LAWRENCE, AND C. LEE GILES. 2000. Efficient identification of web communitites. In *Proceedings of the 6th ACM Conference on Knowledge Discovery and Data Mining*, 150–160. Boston, MA.

FREEMAN, LINTON C. 1979. Centrality in social networks: Conceptual clarification. *Social Networks* 1:215–239.

FUNKHOUSER, G. RAY, AND MAXWELL E. MCCOMBS. 1971. The rise and fall of news diffusion. *Public Opinion Quarterly* 35(1):107–113.

VAN DER GAAG, MARTIN, TOM A.B. SNIJDERS, AND HENK D. FLAP. 2005. *Measurement of individual social capital*, chap. Position generator and their relationship to other social capital measures.

GALLANT, A.R. 1987. *Nonlinear statistical models.* New York: Wiley.

GANZENBOOM, H. B. G., AND D. J. TREIMAN. 2003. *Advances in cross-national-comparison. a european working book for demographic and socio-economic variables*, chap. Three internationall standardized measures for comparative research on occupational status, 159–193. Kluwer Academic Press.

GARRETT, JESSE JAMES. 2005. Ajax: A new approach to web applications. http://www.adaptivepath.com/publications/essays/archives/000385.php.

GARTON, LAURA, CAROLINE HAYTHORNTHWAITE, AND BARRY WELLMAN. 1999. *Doing internet research*, chap. Studying on-line social networks. Thousand Oaks, CA: Sage.

GIBSON, D., J. M. KLEINBERG, AND P. RAGHAVAN. 1998. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia.*

GLOBAL REACH. 2005. Global internet statistics (by language). Tech. Rep., Global Reach. http://www.glreach.com/globstats/.

GLOBE OF BLOGS. 2001. Globe of blogs. http://www.globeofblogs.com/.

GOOGLE, INC. 2005. Google information for webmasters. http://www.google.com/webmasters/bot.html.

GRANOVETTER, MARK. 1973. The strength of weak ties. *The American Journal of Sociology* 78(6):1360–1380.

———. 1978. Threshold models of collective behavior. *American Journal of Sociology* 83(6):1420–1443.

———. 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.

GRINTER, R.E., AND MARGERY ELDRIDGE. 2003. Wan2tlk?: Everyday text messaging. In *Proceedings of the ACM Conference on Computer-Human Interaction.* Ft. Lauderdale, FL.

GRINTER, R.E., AND L. PALEN. 2002. Instant messaging in teen life. In *Proceedings of the ACM Conference on Computer-Supported Collaborative Work.*

GRUHL, D., DAVID LIBEN-NOWELL, R. GUHA, AND A. TOMKINS. 2004. Information diffusion through blogspace. In *Proceedings of the ACM Conference on the World Wide Web.* New York, NY.

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.

HARTIGAN, J. A., AND M. A. WONG. 1978. Algorithm as 136: A k-means clustering algorithm. *Applied Statistics* 28:100–108.

HAYTHORNTHWAITE, CAROLINE. 2000. Online personal networks: Size, composition and media use among distance learners. *New Media & Society* 2(2):195–226.

————. 2002. Strong, weak and latent ties and the impact of new media. *The Information Society* 18(5):385–401.

HAYTHORNTHWAITE, CAROLINE, AND BARRY WELLMAN. 1998. Work, friendship and media use for information exchange ina networked organization. *Journal of the American Society for Information Science* 49: 1101–1114.

HERRING, S. C., I. KOUPER, J. C. PAOLILLO, L. A. SCHEIDT, M. TYWORTH, AND P. WELSCH. 2005. Conversations in the blogosphere: An analysis "from the bottom-up". In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS'05)*. Los Almitos: IEEE Press.

HERRING, S. C., L. A. SCHEIDT, S. BONUS, AND E. WRIGHT. 2004. Bridging the gap: A genre analysis of weblogs. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS'04)*. Los Alamitos: IEEE Press.

HEYDON, ALLAN, AND MARC NAJORK. 1999. Mercator: A scalable, extensible web crawler. *World Wide Web* 219–229.

ISAACS, ELLEN, STEVE WHITTAKER WALENDOWSKI, DIANE J. SCHIANO, AND CANDACE KAMM. 2002. The charachter, functions and styles of instant messaging in the workplace. In *Proceedings of the ACM Conference on Computer-Supported Collaborative Work*. New Orleans, LA.

KATZ, ELIHU, AND PAUL F. LAZARSFELD. 1955. *Personal influence*. Glencoe, IL: Free Press. By Elihu Katz and Paul F. Lazarsfeld. With a foreword by Elmo Roper. "A report of the Bureau of Applied Social Research, Columbia University." Bibliography: p. 381-393.

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.

KLEINBERG, JON M. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*.

KOSTER, MARTIJN. 2005. The web robots faq.
`http://www.robotstxt.org/wc/faq.html`.

KOTTKE, JASON. 2004. Repeat after me: inbound links do not indicate either readership or influence.
http://www.kottke.org/remainder/04/08/6257.html.

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.

KUMAR, RAVI, PRABHAKAR RAGHAVAN, JASMINE NOVAK, AND ANDREW TOMKINS. 2003. On the bursty evolution of blogspace. In *Proceedings of the ACM Conference on the World Wide Web.*

KYPRI, KYPROS, AND STEPHEN J. GALLAGHER. 2003. Incentives to increase participation in an internet survey of alcohol use: A controlled experiment. *Alcohol and Alcoholism* 38(5):437–441.

LEE, SU HYUN. 2004. Souped-up blog takes south korea by storm.
http://www.iht.com/articles/2004/12/30/business/ptkorblog.html.

LILJEROS, FREDRICK, CRISTOFER EDLING, LUÍS AMARAL, EUGENE STANLEY, AND YVONNE ÅBERG. 2001. The web of human sexual contacts. *Nature* 411: 907–908.

LIN, ENSEL W. M. VAUGHN J. C. 1981. Social resources and strength of ties: Structural factors in occupational status attainment. *American Sociological Review* 46:393–405.

LIN, NAN. 1999a. Building a network theory of social capital. *Connections* 22(1):28–51.

———. 1999b. Social networks and status attainment. *Annual Review of Sociology* 25:467–487.

———. 2001. *Social capital: A theory of social structure and action.* Structural analysis in the social sciences 19, Cambridge, UK: Cambridge University Press.

LIN, NAN, AND M. DUMIN. 1986. Access to occupations through social ties. *Social Networks* 8:365–385.

LIVEJOURNAL. 2001. Livejournal.
http://www.livejournal.com.

LIVEJOURNAL. 2005. Livejournal.com statistics.
http://www.livejournal.com/stats.bml.

MARLOW, CAMERON. 2002. Getting the scoop: Social networks for news dissemination. In *Sunbelt International Social Networks Conference XXII.* New Orleans, LA.

———. 2003. Modeling emergent communitites through diffusion. In *Sunbelt International Social Networks Conference XXIII.* Cancun, Mexico.

———. 2004. Audience, structure and authority in the weblog community. In *54th Annual Conference of the International Communications Association.* New Orleans, LA.

MARSDEN, PETER. 1984. Measuring tie strength. *Social Forces* 63:482–501.

———. 1987. Core discussion networks of americans. *American Sociological Review* 52(1):122–131.

MCBRYAN, OLIVER A. 1994. Genvl and wwww: Tools for taming the web. In *Proceedings of the First International World Wide Web Conference*, 79–90.

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.

MEYERSOHN, ROLF, AND ELIHU KATZ. 1957. Notes on a natural history of fads. *The American Journal of Sociology* 62(6):594–601.

MICROSOFT. 2004. Msn spaces. `http://spaces.msn.com/`.

MILGRAM, STANLEY. 1967. The small world problem. *Psychology Today* 2:60–67.

NEUMAN, W. RUSSELL. 1991. *The future of the mass audience.* New York: Cambridge University Press.

NIE, N., AND L. ERBRING. 2000. Internet and society: A preliminary report. Tech. Rep., Stanford Institute for the Quantitative Study of Society: Stanford University.

PARETO, VILFREDO. 1896. *Cours d'economie politique.* Genève: Droz.

PARK, ROBERT. 1967. *On social control and collective behavior.* Chicago, IL: University of Chicago Press.

PARKS, MALCOLM R., AND KORY FLOYD. 1996. Making friends in cyberspace. *Journal of Communications* 46(1):80–97.

PASTOR-SATORRAS, ROMUALDO, 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. 2004. The blogging iceberg: Of 4.12 million hosted weblogs, most little seen and quickly abandoned. Tech. Rep., Perseus Development.

———. 2005. The blogging geyser: 31.6 million hosted blogs, growing to 53.4 million by year end. Tech. Rep., Perseus Development.

PEW INTERNET AND AMERICAN LIFE PROJECT. 2005. Demographics of internet users.
http://www.pewinternet.org/trends/User_Demo_05.18.05.htm.

PORTES, ALEJANDRO. 1998. Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology* 24:1–24.

PUTNAM, ROBERT D. 1993. The prosperous community: social capital and public life. *American Prospect* 13:35–42.

———. 2000. *Bowling alone : the collapse and revival of american community.* New York: Simon & Schuster.

QUAN-HAASE, ANABEL, AND BARRY WELLMAN. 2004. *It and social capital*, chap. How does the internet affect social capital? MIT Press.

RHEINGOLD, HOWARD. 1994. *The virtual community: Homesteading on the electronic frontier.* The MIT Press.

ROGERS, EVERETT M. 1962. *Diffusion of innovations.* 5th ed. New York: Free Press.

———. 2000. Diffusion of printing and the internet. In *Conference on the Printing Press and Internetted Computers.* Santa Monica, CA: Rand Corporation.

RYAN, B., AND N. GROSS. 1943. The diffusion of hybrid seed corn in two iowa communities. *Rural Sociology* 8(1):15–24.

SCHIANO, D.J., C.P. CHEN, GINSBERG J., AND U. GRETARSDOTTIR. 2002. Teen use of messaging media. In *Proceedings of the ACM Conference on Human Factors in Computing.*

SHIRKY, CLAY. 2003. Power laws, weblogs, and inequality.
http://shirky.com/writings/powerlaw_weblog.html.

SHRUM, W, N. H. CHEEK, AND S. M. HUNTER. 1988. Friendship in school: Gender and racial homophily. *Sociology of Education* 61:227–239.

SMITH, MARC. 1999. *Communities in cyberspace*, chap. Inivisible crowds in cyberspace: Mapping the social structure of the Usenet. Routledge.

SNIJDERS, T.A.B. 1999. Prologue to the measurement of social capital. *La Revue Tocqueville* XX:27–44.

SPROULL, L., AND S. KIESLER. 1986. Reducing social context cues: Electronic mail in organizational communication. *Management Science* 32:1492–1512.

———. 1991. *Connections: New ways of working in the networked organization.* Cambridge, MA: MIT Press.

TURKLE, SHERRY. 1995. *Life on the screen.* New York: Touchstone.

VALENTE, THOMAS W. 1993. Diffusion of innovations and policy decision-making. *Journal of Communications* 43(1):30–45.

———. 1995. *Network models of the diffusion of innovations.* Quantitative methods in communication, Cresskill, N.J.: Hampton Press. Thomas W. Valente. Includes bibliographical references (p. 153-163) and indexes.

VALENTE, THOMAS W., AND EVERETT M. ROGERS. 1995. The origins and development of the diffusion of innovations paradigm as an example of scientific growth. *Science Communications* 16(3):242–273.

VAN SONDERSON, E., J ORMEL, E. BRILMAN, AND C. VAN LINDEN VAN DEN HEUVELL. 1990. *Social network research: Methodological questions and substantive issues*, chap. A comparison of the exchange, affectivve, and role-relation approach, 101–120. Lisse: Swets & Zeitlinger.

WASSERMAN, S., AND K. FAUST. 1994. *Social network analysis.* Cambridge: Cambridge University Press.

WATTS, DUNCAN, AND S. H. STROGATZ. 1998. Collective dynamics of "small-world" networks. *Nature* 393:440–442.

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. 1997. Structural analysis: From method and metaphor to theory and substance. In *Social structures: A network approach*, ed. Barry Wellman and S. D. Berkowitz, chap. Structural analysis: From method and metaphor to theory and substance, 19–61. Greenwich, CT: JAI Press.

———. 2001. Computer networks as social networks. *Science* 293: 2031–2034.

WELLMAN, BARRY, AND MILENA GULIA. 1999a. *Networks in the global village*, chap. The network basis of social support, 83–118. Boulder, CO: Westview.

———. 1999b. *Networks in the global village*, chap. Net surfers don't ride alone: Virtual communities as communitites. Boulder, CO: Westview.

WELLMAN, BARRY, ANABEL QUAN-HAASE, JAMES WITTE, AND KEITH HAMPTON. 2001. Does the internet increase, decrease or supplement social capital? social networks, participation and community commitment. *American Behavioral Scientist* 45:437–456.

WIKIPEDIA. 2005a. Entry for "computer mediated communication". `http://en.wikipedia.org/wiki/Computer_mediated_communication`.

———. 2005b. Entry for "weblog". `http://en.wikipedia.org/wiki/Weblog`.

WINER, DAVE. 2000. Weblogs.com recently updated weblogs. `http://www.weblogs.com/`.

WINSTEAD, JIM. 2005. Blo.gs.
http://blo.gs/.

WOOLCOCK, MICHAEL. 1998. Socal capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society* 27: 151–208.

WORDPRESS. 2005. Wordpress weblog software.
http://www.wordpress.org.

WYNN, ELEANOR, AND JAMES E. KATZ. 1997. Hyperbole over cyberspace: Self-presentation and social boundaries in internet home pages and discourse. *The Information Society* 13(4):297–327.

XANGA.COM. 2003. Xanga.com - the weblog community.
http://www.xanga.com/.

ZIPF, GEORGE KINGSLEY. 1949. *Human behavior and the principle of least effort.* Addison-Wesley.