YOUR MEMBERS ARE ALSO YOUR CUSTOMERS: MARKETING FOR INTERNET SOCIAL NETWORKS

Michael Trusov, Anand V. Bodapati, Randolph E. Bucklin [1]

Working paper
September 29th, 2006

[1]*Michael Trusov (Michael.Trusov@anderson.ucla.edu) is a doctoral candidate, Anand V. Bodapati (abodapat@anderson.ucla.edu) is an Assistant Professor, and Randolph E. Bucklin (randy.bucklin@anderson.ucla.edu) is a Professor of Marketing, Anderson School at UCLA, 110 Westwood Plaza, Los Angeles, CA 90095.*

ABSTRACT

Perhaps the fastest growing arena in the World Wide Web is the space of social networking sites (e.g., Friendster, Facebook, MySpace). The success of these sites directly depends on the number and activity level of their users. What attracts users to the site is a continually changing digital content (e.g., messages, pictures, photos, music, videos, blogs) generated by other users. In contrast, e-commerce sites (e.g., Amazon) or portals (e.g., Yahoo!) can support their well being by manipulating content (adding products, adjusting prices, updating news, *etc*.). We propose a model which allows social networking site managers to identify community members whose network influence makes them important for the business. We link individual user activity to the online behavior of other users in the network. Using Bayesian methods and a variable selection approach, we infer which users in the community are "influential" in affecting the site usage of other members. The developed method is specifically tailored to the variable selection task with a very large number of predictors.

We show that in a user-generated content environment, network effects are critical in determining a customer's value to the firm. We find significant heterogeneity among users on two dimensions: susceptibility to influence from other users and the extent of influence on others in the network. Our findings indicate that social influence in an online community is similar to what is experienced offline. The average user is influenced by very few other network members and, in turn, influences very few others. Finally, a small proportion of people (25% of community members) participates in a major share (75%) of all influential ties identified in the social network.

Keywords: Social Networks, Internet, Bayesian Methods, Stochastic Search Variable Selection

Introduction

The idea of online communities is as old as the Internet. After all, the Internet was created for people to be able to interact. Nevertheless, it has only been in the past one or two years that the phenomenon of online social networking (SN) suddenly exploded. A year ago, the popular SN site MySpace.com was acquired by News Corporation for $580 million. This summer, the much smaller San Francisco-based SN site Bebo.com turned down a $500 million acquisition offer from British Telecom. Well established Internet industry players such as Microsoft, AOL, Google and Yahoo! are also trying to spin off their own SN sites. Today, SN sites (see Table 1 for a list) attract more than 90% of all teenagers and young adults in the U.S. and have a market of about 80 million members. The front cover of *BusinessWeek* magazine (December 12, 2005) suggests that the next generation of Americans could be called the MySpace generation – a generation of people who are literally growing up online and spending a considerable amount of their time socializing on the Internet. As an indication of SN's importance, the Marketing Science Institute has set an overarching theme of "The Connected Customer" as its top research priority for the next two years (Marketing Science Institute, 2006).

Users visit SN sites to interact with other users, to search for new friends or business contacts or simply to "hang out" in virtual space. The core of a SN site is a collection of user profiles where registered members can place information that they want to share with others. Any two users can create a connection between their profiles by exchanging electronic invitations. For the most part, users are involved in two kinds of activities on the site: they either *create new content* by editing their profiles (adding

pictures, uploading music, writing blogs and messages, *etc.*), or they browse through profiles *consuming content* created by others (looking at pictures, downloading music, reading blogs and messages, *etc.*). Essentially, popular SN sites are simply massive collections of continuously changing profiles and emerging connections.

In terms of revenue generation, the most popular SN site business model is advertising-based. Technically, users' site usage translates into sequences of web pages served by the site to users' screens. Each page served to users may carry a few ads. SN firms earn revenue either from showing ads to site visitors (impressions) or being paid for each click/action taken by site visitors in response to an ad (e.g., a click that leads to a purchase). Some SN sites also generate revenue from their members through the pay-per-contact model or subscriptions.

Due to the user-generated content at its core, an SN site depends on users to a greater extent than an online business which has full control over the products it sells. For example, Amazon.com can choose which products to offer and what prices to set. On the other hand, an SN site cannot go much beyond basic content censoring. The bulk of digital content – the key "product" which attracts traffic to the site – is produced by its users. Users are not all created equal. There is a great deal of heterogeneity across community members in terms of frequency, volume and attractiveness of digital content generated. A recent article in *The Wall Street Journal* (Holmes, 2006) provides anecdotal evidence that site traffic highly correlates with content updates produced by some key content contributors. Holmes reports that when a popular blogger left a particular site for a two-week vacation, the site's visitor tally fell (Figure 1). Clearly, not all user-generated

content is going to be demanded by other users. In Holmes' example, content produced by three invited substitute bloggers could not prevent site traffic from falling.

From a managerial perspective, it is important to understand who the users are and who keeps the SN site running. Who are the key players? What drives site visitation and therefore advertising revenues? Who makes the site attractive to others? How should the SN site segment its customers? To manage the site, make it a better place for consumers, and obtain better information for advertisers, the site's management needs to know who the influential members are. The objective of this study is to develop a model that will help managers identify their key players – the influential people in the network.

In the social network analysis literature, there are many examples of how the importance of an individual member can be inferred from his or her location in the network (e.g., Iacobucci, 1998). The critical element for such analysis, however, is the known topology (structure) of the network. In the world of SN sites, however, connections between two profiles are not very informative. A connection indicates that two users were introduced to each other; however, it does not carry any notion of strength, direction or influence. Besides asking people directly about their network acquaintances, there are several ways in which the strength and direction of connections can be derived from secondary data. Possible measures might include the number of e-mail exchanges between profiles, the number of content uploads/downloads, or the number of profile views. Clearly, the notion of influence should be driven by research objectives. Site traffic is a primary revenue driver for most of the SN sites. Therefore the notion of influence in this study is linked to factors affecting site usage.

In our dataset, which comes from a major social networking site, we tracked daily login activities of community members (treating the frequency of logins as a reasonable proxy for usage). We can use the data to ascertain the effect of any change in a user's behavior (either increasing or decreasing usage) on the behavior of those connected to him or her. If a member increases his or her usage, and the people connected to him or her also increase their usage (as they are very interested in what this person is creating), we propose that this identifies this person as influential. On the other hand, if a member's login behavior goes up or down and nothing happens to the people connected to him or her, then we propose that this person is not an influencer. With the data and this notion of influence, we seek to identify users whose behavior on the site has the most significant impact on the behavior of other individuals in the network.

Substantively, we show that in a user-generated content environment, network effects are critical in determining a customer's value to the firm. We find significant heterogeneity among users on two dimensions: susceptibility to influence from other users and the extent of influence on others in the network. Our findings indicate that social influence in an online community is similar to what is experienced offline. The average user is being influenced by very few other network members and, in turn, influences very few people. Also, having many friends does not make users influential *per se*. Finally, a small proportion of people (25%) is responsible for a major share (75%) of all existing influential ties.

Many online communities have thousands, if not millions, of members. Therefore, to be useful for marketing practitioners, the proposed model must be able to handle large

networks. We use a Bayesian stochastic search method for best subset selection that is scalable to a large number of explanatory variables.

In the next section, we provide a brief overview of existing SN sites. Then, we briefly discuss four streams of the literature related to our research – social influence, online communities research, Internet studies on repeat site visitation, and the statistical literature on Bayesian variable selection methods. We continue with a description of the dataset and model formulation, followed by empirical estimation results. We conclude with managerial implications, limitations and directions for future research.

Background on Online Communities

According to an entry posted on Wikipedia (www.wikipedia.org), there are currently about 30 social networking websites which have more than one million registered users each and several dozen significant, though smaller, sites. In terms of web traffic, ComScore MediaMetrix reported that the largest online SN site (as of March 2006) was MySpace.com, with 42 million unique visitors per month, followed by FaceBook.com with 13 million and Xanga.com with 7.4 million.[2] ComScore MediaMetrix numbers suggest that every second Internet user in the U.S. visits one of the top 15 SN sites (Table 1).

A social networking site is typically initiated by a small group of founders who send out invitations to join the site to the members of their own personal networks. In turn, new members send invitations to their personal networks, and so on. Typical SN sites allow a user to build and maintain a network of friends for social or professional interaction. In the core of a SN site are personalized user profiles (Figure 2). Individual

[2] For a comprehensive overview of major online community sites, we refer interested readers to O'Murchu, Breslin and Decker (2004).

profiles are usually a combination of users' images (or avatars), lists of interests and preferences and links to affiliated profiles ("friends"). Different sites impose different levels of privacy in terms of what information is revealed through profile pages to non-affiliated visitors and how far "strangers" and "friends" can traverse through the network of a profile's friends. Profile holders acquire new "friends" by browsing and searching the site and sending requests to be added as a friend. The term "friend" in SN settings does not carry the same meaning of "friendship" in the traditional sense of the word. In social networking, the only reliable inference that a "friend" connection implies is that two users were introduced to each other online.

Classical social network analysis (e.g., Wasserman and Faust, 1994) defines a social network as a set of actors and a set of relations among them. In the case of online communities, actors are site users and relations are "friend" connections. From a firm's perspective, a network is a connected, undirected graph with binary edges (Figure 3). Due to the ambiguity of "friendship" on SN sites, classical methods using topological analysis of social networks should be regarded with caution. We illustrate this point with the following hypothetical example.

Let us focus on a specific individual in the network – Allison (Figure 4). Network analysis calls this the ego-centered view. Allison's ego-centered network is a network of her "friends." In Figure 4, "friends" (the people with whom Allison has exchanged invitations) are represented by solid circles. It is quite likely, however, that among "friends" there are just a few people who actually make the site attractive to Allison. From Allison's perspective, these are "important" friends. She comes back to the site

looking for new content produced by them, while tending to ignore updates in other connected profiles.

In our dataset, the average number of "friends" per user is 90. Sociologists tell us that, on average, we have about two real "close friends" (McPherson *et al*., 2006). If, for each user in the network, we could pinpoint his or her influencers, the topology (structure) of the network may appear to be quite different. In Figure 5, we show the hypothetical network after non-influential connections are removed. This network is reconstructed from the original network (Figure 3) by taking the ego-centered networks of each user and identifying the influential friends within his or her network. Compared to the original network, this updated network depicts the important connections among profiles more clearly. Using standard SN analysis tools, we could then derive a range of useful measures for the updated network. The most obvious one (*degree centrality* in SN terms) would be to count how many people a particular individual is influencing. We will discuss our approach to ego-centered analysis in the following sections.

Literature

Our study is related to four distinct research streams: social influence, online communities, repeat Internet site visitation, and Bayesian variable selection methods. We provide brief highlights from each of these four areas.

Social Influence

Social influence occurs when an individual adapts his or her behavior, attitudes or beliefs to the behavior, attitudes or beliefs of others in the social system (Leenders, 1997). Social influence has been the subject of more than 70 marketing studies since the 1960s. Overall, scholarly research on social and communication networks, opinion

leadership, source credibility, and diffusion of innovations has long demonstrated that consumers influence other consumers (Phelps *et al*., 2005). Influence does not necessarily require face-to-face interactions, but is based on information about other people (Robins *et al*., 2001). In an online community information is passed among individuals in the form of digital content. We contribute to an extremely rich literature on social influence by considering a particular type of social influence, which takes place in an online community when community members change their site usage in response to changes in the behavior of other members.

Online Communities

Although still a relatively new area in marketing research, online communities have attracted the attention of many scholars. Dholakia *et al*. (2004) studied two key group-level determinants of virtual community participation – group norms and social identity. The authors tested the proposed model using a survey-based study across several virtual communities. Kozinets (2002) developed a new approach to collecting and interpreting data obtained from consumers' discussions in online forums. Godes and Mayzlin (2004) looked at the effect word-of-mouth communications in online communities had over TV show ratings. They showed that a measure of the dispersion of conversations across communities has explanatory power in a dynamic model of TV ratings. By studying consumer reviews posted on Amazon.com and BarnesandNoble.com, Chevalier and Mayzlin (2006) found that an improvement in a book's reviews leads to an increase in sales at that site and that the impact of negative reviews is greater than the impact of positive ones. Dellarocas (2005) analyzed how the strategic manipulation of Internet opinion forums affects the payoffs of consumers and

firms in markets of vertically differentiated experience goods. The author showed that while under certain conditions forum manipulation can be beneficial to consumers, in more general cases (e.g., manipulation costs are convex and the number of "honest" consumers' post is high), such activities end up reducing the profits of all competing firms as well as the overall social welfare. Finally, Narayan and Yang (2006) studied a popular online provider of comparison shopping services – Epinions.com. The authors modeled the formation of relationships of "trust" that consumers form with other consumers, whose online product reviews they consistently find to be valuable.

Our research contributes to the existing literature on online communities in several ways. First, we believe that online social networking sites are a unique type of online community. Some aspects of socializing in the virtual worlds of MySpace or Facebook are similar to the online bulletin board-type of interactions found on movie or consumer product review sites – the most common type of online communities studied in the marketing literature. There is too much dissimilarity, however, from the number of people involved, the motives for and nature of interactions, to the revenue generating models, to treat them alike. Second, previous research has not looked into the problem of peer influence on individual level site usage, which is the focus of the present research. Finally, from a methodological perspective, none of the above empirical studies has simultaneously modeled the individual-level influence within a group of users. They have either focused on interaction within a dyad (e.g., Narayan and Yang, 2006) or considered aggregated group-level measures (e.g., Dholakia *et al*., 2004, Godes and Mayzlin, 2004).

Repeat Site Visitation

Repeat web site visitation is a focal construct of several studies on Internet marketing. In the present research we use the frequency of logins to the social networking site as a proxy for site usage. It is acknowledged that e-business needs to know the "visit frequency" as well as the "visit timing" of its customers (Lee *et al.*, 2001). For example, Moe and Fader (2004) developed a model of visit-to-purchase conversion that predicts a customer's probability of purchasing based on an observed history of visits and purchases. Park and Fader (2004) developed a stochastic model of cross-site visit-timing behavior to understand how the visitation pattern from one site might explain customer behavior at another. Lee, Zufryden and Drèze (2001) developed the NBD model to measure the effects of a consumer's Internet use and demographics as well as a firm's marketing activities on consumers' repeat visit behavior. None of the above studies, however, considered the peer-influence effect, which is the focus of our research.

Bayesian Variable Selection Methods

The problem of variable selection occurs in modeling the relationship between a variable of interest and a subset of potential explanatory variables when there is uncertainty about which subset to use (George, 2000). In our application, the potential explanatory variables are site usage measures for each "friend" of a particular individual. We want to search for the subset of "friends" whose site usage explains variation in login activities for the focal individual (also known as the *ego*). The main challenge here is the possibly very high dimensionality of the search space. For an average user with 90 "friends," we would need to compare 2^{90} possible subsets – computationally too expensive and time consuming. Classical methods for variable selection have employed

different algorithms, including backward elimination, forward selection, and stepwise regression. In the Bayesian literature, several approaches to variable selection have been developed. George and McCulloch (1993) proposed a Stochastic Search Variable Selection (SSVS) procedure for identifying "promising" subsets of predictors. SSVS would avoid estimating all 2^{90} possible models in our example and, instead, stochastically search among the models which have the highest posterior probability. SSVS is based on the Gibbs sampler for simulating draws from a posterior distribution. Because high probability models are more likely to appear quickly, the Gibbs sampler can identify such models with relatively short runs (George and McCulloch, 1997). In the spirit of the SSVS approach, Kuo and Mallick (1998) developed a simpler method of subset selection. They associate each predictor in the full model with an indicator variable (*i.e.*, 0 or 1). The posterior mean of each indicator is then interpreted as a probability that the corresponding predictor belongs to the "best" model. Gilbride *et al*. (2006) extended the SSVS by introducing individual-level heterogeneity and context dependence in a choice model setting.

Making use of the aforementioned studies, we have developed our own version of the SSVS algorithm, tailored to the variable selection task with a very large number of predictors. Our approach, described in detail in the following sections, accounts for heterogeneity in user-level susceptibility to the influence of "friends". We use the proposed method for weeding out "non-influential" friends.

Modeling Approach

Model Specification

For each user in an online community, we perform a search for influential friends within his or her ego-centered network (Figure 4). Next, we reconstruct a full network, using only influential links obtained from the ego-centered analysis (Figure 5). Finally, on a full network we use standard SN analysis measure – *degree centrality* – to evaluate how many people a particular individual is influencing.

To identify the influential friends within an ego-centered network, we model a user's login activity as a function of the user's characteristics, the user's past behavior on the site and, most importantly, the login activity of the user's friends. We suggest that the count of individual daily logins follows a Poisson distribution with rate parameter λ_{ut}, which may vary across users (u) and time (t). Accordingly, we model the number of daily logins y_{ut} as a Poisson regression (equation 1).

$$(1) \qquad y_{ut} \sim Poisson(\lambda_{ut})$$

The Poisson regression model is derived from the Poisson distribution by parameterizing the relation between the rate parameter λ_{ut} and predictors (Nelder and Wedderburn, 1972; Cameron and Trivedi, 1999), which we group into two sets: self-effects and friend-effects. Self-effects include such covariates as user-specific intercepts, day of the week and past logins. Friend-effects consist of friends' lagged login activity. It is customary to use the exponential rate parameterization. In words, the logarithm of the rate parameter is specified as

$$(2) \qquad \log(\lambda_{ut}) = Self\ \ Effects_{ut} \ \ + \ \ Friend\ \ Effects_{ut}$$

In symbols, we have

$$(3) \qquad \log(\lambda_{ut}) = \alpha_{u1}x_{u1t} + \alpha_{u2}x_{u2t} + \ldots + \alpha_{uK}x_{uKt} + \beta_{u1}z_{u1t} + \beta_{u2}z_{u2t} + \ldots + \beta_{uF_u}z_{uF_ut}$$

where

x_{ukt} = user-specific covariate k (e.g., intercept, day-of-the-week effect, logins at t-1),

z_{uft} = weighted average of lagged login activities of friend f of user u at time t,

F_u = number of friends of user u,

α_{uk} = coefficient of user-specific covariate k, and

β_{uf} = coefficient of friend f for user u.

We offer the following rationale for this model. Community members are attracted to the site by continually changing digital content generated by other users. Digital content constitutes pretty much anything that leaves some "digital" trace on the site and can therefore be either directly observed or in some other way experienced by network members. The most obvious examples of user-generated content are messages and testimonials, updated pictures, uploaded photos, music, and videos, and blog posts. There are also some other, less apparent, kinds of digital content, which, nevertheless, could be of high relevance to network members. For example, many SN sites allow profile-holders to check whom among his or her friends visited his or her page in the past. Clearly, for many users, the "popularity" of their profile is a very strong motivation to generate new content. The last login date and time are typically known for each user; from this information, network members could infer how active a specific individual has been on the site. Finally, when online, site users are able to see which of their friends are currently logged in. The fact that the users' friends are also "hanging out" there is likely to be appealing to users, and therefore would make the site visitation experience more enjoyable (Figure 6).

Let us assume that one of Allison's friends (e.g., Gordon) becomes more active on the SN site – he goes there more often, puts out more content, spends more time surfing through other profiles, etc. By means of digital traces (as discussed above), Gordon's increased activity level is likely to be noticed by Allison. For Allison, this "learning" process may take some time, or it could happen instantaneously. Now, if Gordon is among Allison's "important" friends, and, therefore, his presence on the site makes a difference to her, then the site's attractiveness to Allison will also increase. Accordingly, she may respond by increasing her level of activity on the site. From the firm's perspective, changes in site activity levels can be inferred from changes in login behavior – an individual who is more involved with the site is likely to log in more often.

The intuition behind the model defined in equation (3) is that a user's site usage at any given point in time depends, among other things, on the site usage of this individual's "important" friends. The proposed model captures this process through the friend-specific coefficients β_{uf}. Given that learning typically takes time, for each friend f of user u we construct a covariate z_{ufl} as a weighted average of friend f's login activities over the past D days, as specified in equation (4):

$$(4) \qquad z_{ufl} = \sum_{d=1}^{D} w_d \times y_{f(t-d)} \text{ , where } \sum_{d=1}^{D} w_d = 1$$

We use a grid search method to choose optimal weight coefficients w_d and number of lags D in equation (4).[3]

[3] We recognize that the proposed measure of a friend's past login activities can be improved. A more comprehensive model should control for possible heterogeneity in weights (w) and lags (D) across users. Another approach would be to use an exponential smoothing over lagged logins (e.g., Guadagni and Little, 1983). A smoothing constant could be made user-specific and be estimated in the sampler.

While the Poisson assumption may appear simplistic, there are several arguments that support our modeling choice. First, the fundamental limitation of Poisson regression, equidispersion, does not present a significant problem for our analysis. In Figure 7, we plot variance in daily login activities against the mean daily logins on an individual level. While we can spot overdispersion for part of our sample, the difference between mean and variance is not dramatic. Second, we benchmarked the performance of our model against the more flexible NBD specification and did not find any significant difference in estimation results. Finally, the Poisson formulation integrates well with our variable selection algorithm (described below), which results in a very parsimonious and scalable solution.

The model defined in equation (3) presents certain challenges. The major problem is that the number of friends varies from user to user and, for some users, this number is rather large. Some users have hundreds, if not thousands, of friends. This would require a very large number of coefficients to be estimated, but even for relatively long panels (like over 80 daily observations we have), this exceeds the available degrees of freedom ("the large p small n problem"). A related problem is that different users have almost entirely different sets of friends. In the marketing literature, for short panels, it is common to use Bayesian shrinkage to obtain individual level parameters. Typically, we can obtain a price response coefficient for a household by leveraging the population average response. We cannot use this idea here without modification; in scanner data the variables such as price are common across users. But the variable set in our case is not the same across all users. We illustrate this in Table 2. For predicting John's behavior, the predictor variables are *John's* friends' behavior. For predicting Emily's behavior, the predictor

variables are *Emily's* friends' behavior. To address these problems, we propose a different kind of shrinkage, a shrinkage used in the variable selection context.

As mentioned earlier, we believe that the average person probably tracks just a few other people. This means that most of the β_{uf} coefficients in equation (3) are likely to be zero. The challenge is to find which ones are not. We look at this task as a variable selection problem, *i.e.* as a problem of best subset selection. In our data, the average number of friends per user is 90. For a user with 90 friends, we would need to consider 2^{90} possible subsets. Even for a moderate number of explanatory variables *f*, estimation of all 2^f models can be prohibitively expensive. Thus some reduction of the model space is needed (George, 2000). We solve this problem using a stochastic search approach implemented by Markov Chain Monte Carlo (MCMC).

As a first step, we decompose each friend-specific coefficient into a user's susceptibility to a friend's influence, denoted by β_u, and a binary parameter γ_{uf}, which serves to sort out a user's non-influential friends:

$$\beta_{uf} = \beta_u \times \gamma_{uf} \tag{5}$$

This captures two phenomena: (1) either a friend is influential or is not, and (2) susceptibility to friends' influence can vary from user to user. Equation (6) shows how γ_{uf} is drawn in the Gibbs sampler:

$$\gamma_{uf}, \forall f \mid \bullet \propto Bin(\gamma_{uf} \mid \frac{c_{uf}}{c_{uf} + d_{uf}}) \tag{6}$$

where

$$c_{uf} = p_u \times L_u(\bullet, \gamma_{uf} = 1),$$

$$d_{uf} = (1 - p_u) \times L_u(\bullet, \gamma_{uf} = 0),$$

L_u = the Poisson likelihood function for user *u*, and

p_u = prior probability of friend *f* being influential for user *u* (estimated in the sampler).

The intuition behind equation (6) is as follows. In each iteration of the Gibbs sampler, a friend-specific γ_{uf} is drawn as Bernoulli random variable, with success probability based on the ratio of the likelihood with friend *f*'s effect included (*i.e.* with γ_{uf} =1) to model likelihood without friend *f* (*i.e.* with γ_{uf} =0). So, for example, after 10,000 iterations for each friend, we observe 10,000 draws of γ_{uf} (Figure 8). From the variable selection method's perspective, the posterior mean of γ_{uf} is a probability that the corresponding covariate z_{uf} should be included in the model. The behavioral interpretation is that friend *f* is influential for user *u*.

Let IF_u denote the sum of all the γ_{uf} terms for any particular user *u*. The IF_u can be interpreted as the number of influential friends. The p_u term is updated by drawing from the beta distribution **B**(1+ IF_u , 1 + F_u - IF_u), which has a mean that is approximately equal to the empirical fraction of influential friends.

The model defined in equation (6) does not control for a possible reciprocity of influence between user *u* and friend *f*. We recognize that treating influence in a dyad as independent may result in biased estimates. We plan to address this issue in the next version of our model. As a possible solution, we would consider the introduction of friend effect priors into equation (6) in the following manner:

$$c_{uf} = \tilde{p}_u \times L_u(\bullet, \gamma_{uf} = 1)$$

$$d_{uf} = (1 - \tilde{p}_u) \times L_u(\bullet, \gamma_{uf} = 0)$$

where

$$\tilde{p}_u = p_u \times p_f \times e^{\tau(\gamma_{uf}-0.5)}$$

In these settings, the reciprocity effect is captured by parameter τ. Also, adding p_f to the prior allows us to control for a friend's influence across multiple users.

For a real social networking site, we might need to estimate millions of ego-centered networks. The proposed decomposition results in a very scalable solution. Updating β coefficients collapses to a simple Poisson regression with a small number of parameters. Indeed, conditional on γ_u , sufficient statistics for β_{uf} become an inner product of vector $z_{ut} = [z_{u1t}, z_{u2t}, \ldots z_{uFt}]$ and vector $\gamma_u = [\gamma_{u1}, \gamma_{u2}, \ldots \gamma_{uF}]$. Accordingly, equation (3) can be rewritten as equation (7):

$$(7) \qquad \log(\lambda_{ut}) = \alpha_{u1}x_{u1t} + \alpha_{u2}x_{u2t} + \ldots + \alpha_{uK}x_{uKt} + \beta_u \sum_{f=1}^{F_u} \gamma_{uf} z_{uft}$$

Also, instead of the traditional random walk, we use an independence sampler, where the Poisson likelihood function is approximated by a normal distribution with Metropolis correction. We generate parameter values α_{uk} and β_u using a normal distribution centered at the maximum likelihood estimate for equation (6) with the variance equal to the asymptotic variance (approximated by inverse Hessian). For more details and in-depth discussion of our estimation approach, we refer the interested reader to a technical appendix (available upon request).

Empirical Analysis

Data Description

We applied our model to the data obtained from one of the major social networking sites, which wishes to remain anonymous. In the 12-week dataset, we track daily login activities for 330 users and their 29,478 friends. For each user/friend in the

dataset we observe full profile information (e.g., networking goals, number of friends, number of profile views) as well as self-reported demographics (e.g., age, education, income, zip code). Table 3 provides descriptive statistics on users' profiles: of the users included in the sample, 45% are male, 29% are married, 73% are reportedly interested in friendship, 24% would consider dating, and 15% are interested in serious relationships. The mean age of users is 19.7 years, and 62% of users are white, 24% are Hispanic, 7% are African-American, and 6% are Asian (Figure 9). Figure 10 shows users' geographical locations and "friendship" connections among profiles.

The average number of logins per day in our sample is 2.48. Figure 11 shows the distribution of the number of logins. Figure 12 provides examples of login time series for four users. Each bar on these graphs corresponds to the number of logins on a specific date for a specific individual. It is apparent that site usage varies considerably from one user to another. For example, on average, user #118 logs in about 4 times per day, user #97 logs in about 3 times, and user #12 logs in approximately 1.4 times per day. To a certain extent, the cross-sectional differences in average login numbers can be explained by profile data (Table 4). A basic comparison of average site usage across different segments of users shows that the most active user is a single white male, less than 20 years old, who is interested in a serious relationship.

Model Estimation

We estimate our model using a Bayesian approach, implemented with MCMC methods. To complete the model specification, we introduce priors over the parameters common to all users (see technical appendix). We monitor chains for convergence and,

after convergence, we allow long chains to run. We burn 50,000 draws and simulate an additional 50,000.

Estimation Results

Using the methodology described in the previous section, we present results from estimating a series of models, each predicting the number of daily logins for each individual in our sample. Model 1 is a benchmark model incorporating self-effects only. From a network effects perspective, Model 1 ignores friend *f*'s effects, setting all of the γ_{uf} to zero. Model 2 is another benchmark in which we incorporate the effect from all of the user's friends. In other words, all of the γ_{uf} are set to "one" in this model. Finally, Model 3 is our proposed model, using a variable selection approach and friends' effect decomposition. To assess model fit and to conduct model comparisons, we use the Deviance Information Criterion (DIC) (Spiegelhalter *et al*., 2002). As shown in Table 5, the proposed model provides a significantly better fit to the data compared to the two benchmark models.

On an individual level, we observe considerable heterogeneity among users in terms of the number of influential friends they have. In Figure 13, we show two users with a similar number of friends but quite distinct patterns of influential friends. Bars on these graphs correspond to the posterior probability that a particular friend is influential for the user. In Figure 13, we show two users with about 100 friends each, so we plot 100 γ_{uf}. For the first user, we observe a few influential friends, while for the second there are none. Figure 14 shows the resulting empirical distribution of the posterior mean γ_{uf} for the whole sample. The distribution is considerably skewed to the left, which means that most of the posterior means γ_{uf} are relatively small. In other words, friends corresponding to

these small γ_{uf} have a very low probability of having any influence on users' site usage. For further analysis, we count a friend as influential if his or her posterior mean γ_{uf} exceeds 0.5. Under this condition, our findings indicate that a user's influential friends are about 8% of the total number of friends. The average user has about three important friends, while 30% of users do not have any.

The probability of being influential in a dyad can be explained by static user measures readily available from users' online profiles. Table 6 reports results of posterior analysis.[4] We regress a logit-transformed posterior mean γ_{uf} on covariates extracted from profiles. For example, we find that gender is a good predictor of influence – overall, male users are more susceptible to social influence than female users. Also, in dyads, female users tend to have a stronger influence on males than vice versa. We find that the posterior mean γ_{uf} in a female friend/male user pair tends to be much higher than for any other possible gender combinations. By comparing the popularity of two users in terms of the number of profile views they receive on the network, we find that more popular friends tend to have more influence over less popular friends. Networking goals can also predict susceptibility to friends' influence. For example, users with dating objectives are less sensitive to other people's actions. Users' ethnicity also makes a difference. A friend of the same ethnicity as the user is more likely to affect the user's site usage than a friend of different ethnicity. Finally, married users are less sensitive to friends' influence. We speculate that married people are more likely to have other priorities than socializing on the Internet.

[4] Generally, the model defined in equations (1) through (4) can be easily modified to incorporate profile data in hierarchical fashion by adding priors on γ_{uf}. We reserve this extension for a subsequent version of this manuscript.

Findings from posterior analysis could help the firm's management to predict the probability of influence for any dyad in the network. We should note, however, that profile data help to explain the relatively small variation in influence – R^2 in this regression is quite low, at about 11%.

We can also use the estimated ego-centered networks to reconstruct the full network of influential ties (Figure 5). To evaluate a user's influence in the network, we count the number of friends for whom this user is important. For that we analyzed the ego-centered networks for each friend of 50 users selected from our sample.[5] Results appear in Figure 15. Our findings indicate that 40% of users influence no one. This is the left side of the distribution. Some, however, influence as many as eight or nine people. In Figure 16, we plot the total number of a user's friends versus the number of friends a user influences. We find that the extent of influence varies considerably, even after controlling for the number of friends. For example, in Figure 16, some users with 30 friends have an influence over eight or nine people, while other users with more than 50 friends influence only one or none.

Managerial Implications

In economics there exists the "80/20" principle, in which 80% of the work is done by 20% of the participants (e.g., Koch, 1999). In our analysis of social networking website, we find that the top 25% of influencers in the sample are responsible for 75% of all important ties identified in the network. For SN site management, this 25% might constitute the segment of members most suitable for further study or targeting actions.

A social networking site should clearly be concerned with retaining its most influential customers. In network settings, customer value to the firm is not solely a

[5] Due to computational considerations we use a smaller sample in this analysis.

function of the cash flows generated by a customer, but also of the effect that this customer has on other customers (Gupta *et al*., 2006). The negative impact of an influential user leaving the site is not limited to the lost revenues from, say, ad impressions not served to this particular individual. Rather, the site usage of all linked (dependent) users will be affected as well. Therefore, when determining how much a firm should be willing to spend to retain a particular customer, the user's network influence must be an important component of the valuation. Also, if there is some cost associated with site usage stimulation, a firm needs to know whom they want to target. For example, should Friendster.com allow some of its members to use blog space for free, hoping that their online activities will have a positive network effect on others? The valuation of a user's network influence is essential in addressing this question.

Conclusions and Future Research

The purpose of this study is to shed light on the marketing issues in the emerging phenomenon of online social networking. Specifically, we propose a model which allows social networking site managers to identify community members whose network influence makes them the most important for the business. To do this, we employ site usage activity data from a major social networking site. In our study we link individual user activity to the online behavior of other users in the online community. Using Bayesian methods and a variables selection approach, we infer which users in the community are "influential" in affecting the site usage of other members.

Our findings indicate that social influence on the Internet is similar to what is experienced offline. The average user has very few important friends and influences very few people. Also, having many friends does not make users influential *per se*. Finally, a

small proportion of people is responsible for a site's well-being. From a methodological perspective, we developed a scalable solution to an otherwise very difficult best subset selection problem. To the best of our knowledge, existing research on variable selection has not done this before in an application to massive right-hand-side expressions. As an outcome of our research, site management should now be able to identify who these influential people are.

Our research also has some limitations. The number of daily logins is a coarse proxy for site usage. A more accurate measure would incorporate actual changes to the content as well as the total time spent on the site during each session. Also, we do not observe if particular digital content was actually consumed by an individual user. More detailed data would help to address these issues. We believe, however, that the core of our modeling approach – the best subset selection task – will extend to handle additional data. Another limitation of our research is shared by most Internet-related studies. It is difficult, if not impossible, for a firm to comprehensively know when and how users interact with one another through digital media outside a specific SN site. From interviews with a number of users, we have learned that many of them maintain profiles on multiple SN sites, as well as using other means of digital communications from Instant Messaging (IM) to SMS and BlackBerry e-mails. Other researchers, such as Park and Fader (2004), have previously expressed concern about the inherent limitations of models built entirely upon behavioral data collected on a single site when users' activities may span several sites. Padmanabhan, Zheng and Kimbrough (2001) also warn of the possibility that erroneous conclusions may be drawn from models based only on such data. Of course, the offline interactions of users are also potentially relevant. In this paper

we focus on handling what can be tracked behaviorally on the Internet and leave the role of offline activities for future research.

References

Cameron, A. Colin and Pravin K. Trivedi (1999), "Essentials of Count Data Regression," In *Companion in Econometric Theory*, B. Baltagi (Ed.), New York: Basil Blackwell.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43(3).

comScore Media Metrix (2006), "Total Number Of Unique Visitors To Selected Social Networking Sites, as of March 2006", as appears at http://www.emergencemarketing.com/archives/2006/05/social_networking_sites_d.php, May 22, 2006.

Dellarocas, Chrysanthos N. (2005), "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms", working paper, revision March 2005, University of Maryland, College Park, MD.

Dholakia, Utpal M., Richard P. Bagozzi and Lisa Klein Pearo (2004), "A Social Influence Model of Consumer Participation in Network- and Small-group-based Virtual Communities", *International Journal of Research in Marketing*, 21, 241-263.

George, Edward I. (2000), "The Variable Selection Problem", The Wharton School, University of Pennsylvania, working paper.

George, Edward I., and Robert E. McCulloch (1997), "Approaches for Bayesian variable selection," *Statistica Sinica*, 7, 339–373.

Gilbride, Timothy J., Greg M. Allenby and Jeff Brazell (2006), "Models of Heterogeneous Variable Selection," *Journal of Marketing Research*, 43(3).

Godes, David, and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, 23 (4), 545-560.

Gupta, Sunil, Carl F. Mela, and Jose M. Vidal-Sanz (2006), "The Value of a "Free" Customer," *working paper,* August 25.

Holmes, Elizabeth (2006), "No Day at the Beach: Bloggers Struggle With What to Do About Vacation," *The Wall Street Journal*, August 31, Page B1.

Iacobucci, Dawn (1998), "Interactive Marketing and the Meganet: Network of Networks," *Journal of Interactive Marketing*, Winter, 5–16.

Koch, Richard (1999), *The 80/20 principle: The Secret of Achieving More with Less.* Garden City, NY: Doubleday and Company, Inc.

Kozinets, Robert V. (2002), "The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities," *Journal of Marketing Research*, 39(1), 61-72.

Kuo, Lynn and Bani Mallick (1998), "Variable Selection for Regression Models," Sankhya, *The Indian Journal of Statistics*, B 60, 65–81.

Lee, Sukekyu, Fred Zufryden, and Xavier Dreze (2001), "Modeling Consumer Visit Frequency on the Internet," In 34th Annual Hawaii International Conference on System Sciences (HICSS-34), Volume 7, 2001.

Leenders, Roger Th.A.J. (2002), "Modeling Social Influence through Network Autocorrelation: Constructing the Weight Matrix," *Social Networks*, 24, 21-48.

Marketing Science Institute (2006), "*2006-2008 Research Priorities,*" as appears at http://www.msi.org/msi/rp0608.cfm, July 30, 2006.

McPherson, Miller, Lynn Smith-Lovin, and Matthew E. Brashears (2006), "Social Isolation in America: Changes in Core Discussion Networks over Two Decades," *American Sociological Review*, 71(3), 353-375.

Moe, Wendy W. and Peter S. Fader (2004), "Dynamic Conversion Behavior at e-Commerce Sites," *Management Science*, 50 (3), 326-335.

Narayan, Vishal and Sha Yang (2006), "Trust between Consumers in Online Communities: Modeling the Formation of Dyadic Relationships," *Working paper*, NYU.

Nelder, John A. and R. W. M. Wedderburn (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Series A (General), Vol. 135, No. 3, 370-384.

O'Murchu, Ina, John G. Breslin, and Stefan Decker (2004), "Online Social and Business Networking Communities," DERI Technical Report 2004-08-11.

Padmanabhan, Balaji, Zhiqiang Zheng, and Steven O. Kimbrough (2001), "Personalization from Incomplete Data: What You Don't Know Can Hurt," In Proceedings of KDD 2001, 154-164, 2001.

Park, Young-Hoon and Peter S. Fader (2004), "Modeling Browsing Behavior at Multiple Websites," *Marketing Science*, 23 (3), Summer, 280-303.

Phelps, Joseph E., Lewis, Regina, Lynne Mobilio, David Perry, and Niranjan Raman (2004), "Viral Marketing or Electronic W-O-M Advertising: Examining Consumer Responses to Pass Along Email," *Journal of Advertising Research*, 44(4), 333-348.

Robins, Garry, Philippa Pattison and Peter Elliott (2001), "Network models for social influence processes," *Psychometrika*, 66, 161-190.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A., (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society*, Series B, 64, 583-639, 2002.

Wasserman, Stanley, and Katherine Faust (1994), *Social Network Analysis: Methods and Applications.* New York and Cambridge, ENG: Cambridge University Press.

Wikipedia (2006), "List of social networking websites," as appears at http://en.wikipedia.org/wiki/List_of_social_networking_websites, May 22, 2006.

Table 1. Social Networking Sites Ranking

Social Networking Sites	**Number of Visitors (in thousands)**
MYSPACE.COM	41,889
FACEBOOK.COM	12,917
XANGA.COM	7,448
LIVEJOURNAL.COM	4,047
Yahoo! 360^{o}	3,614
MYYEARBOOK.COM	3,613
HI5.COM	2,609
TAGWORLD.COM	2,275
TAGGED.COM	1,668
BEBO.COM	1,096
FRIENDSTER.COM	1,066
Tribe	871
43THINGS.COM	661
SCONEX.COM	372
Internet Total	**171,421**

Source: ComScore MediaMetrix, March 2006 Report

Table 2. Complications with Traditional Shrinkage

John's friends	Ana	Bret	Allison	Noel	Danny	Stan	Earl
	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Emily's friends	Cindy	Gordon	Felix	Rita			
	β_1	β_2	β_3	β_4			

Table 3. Descriptive Statistics

Profile Info	***Share***
Males	45%
Married	29%
Networking goals:	
- Friendship	73%
- Dating	24%
- Serious Relationships	15%

Table 4. Variations in Average Daily Logins by Segments

Profile Info	***Usage is…***
Males	3% higher
Married	9% lower
Under 20 years old	4% higher
Whites	5% higher
Looking for Serious Relationships	8% higher

Table 5. Model Fit

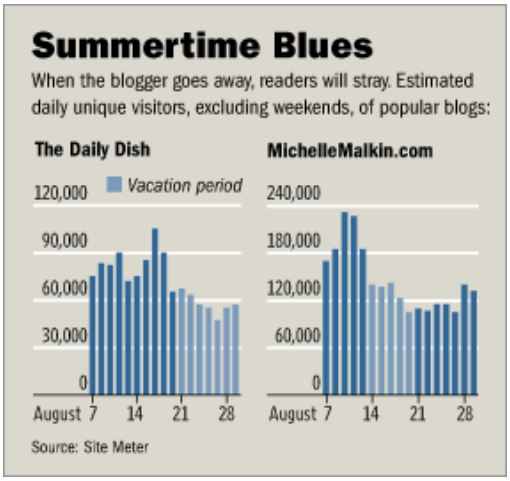
Model	***Description***	***Fit (DIC)***
Model 1	User self effects only (none of friends is influential)	77618.66
Model 2	Self plus all friends' effects (all friends are influential)	77411.55
Model 3	Self and friends effects with variable selection (some friends are influential)	74204.93

Table 6. Explaining Posterior Mean γ_{uf}

Covariate	***Coeff.****	***t-stat***
Gender (Male user)	0.76	13.22
Gender (Female friend/Male user)	0.74	14.25
Friend is More Popular (more profile views)	0.12	2.92
Friend is of the same ethnicity as user	0.46	11.36
User is looking for a date	-0.6	12.18
User is married	-0.1	2.27
R^2		0.11

*LHS: $\log\left(\hat{\gamma}_{uf} / \left(1-\hat{\gamma}_{uf}\right)\right)$

Figure 1. A Decline in Readership Due to Vacation of a Popular Blogger.



Source: The Wall Street Journal, August 31, Page B1.

Figure 2. Profile Example

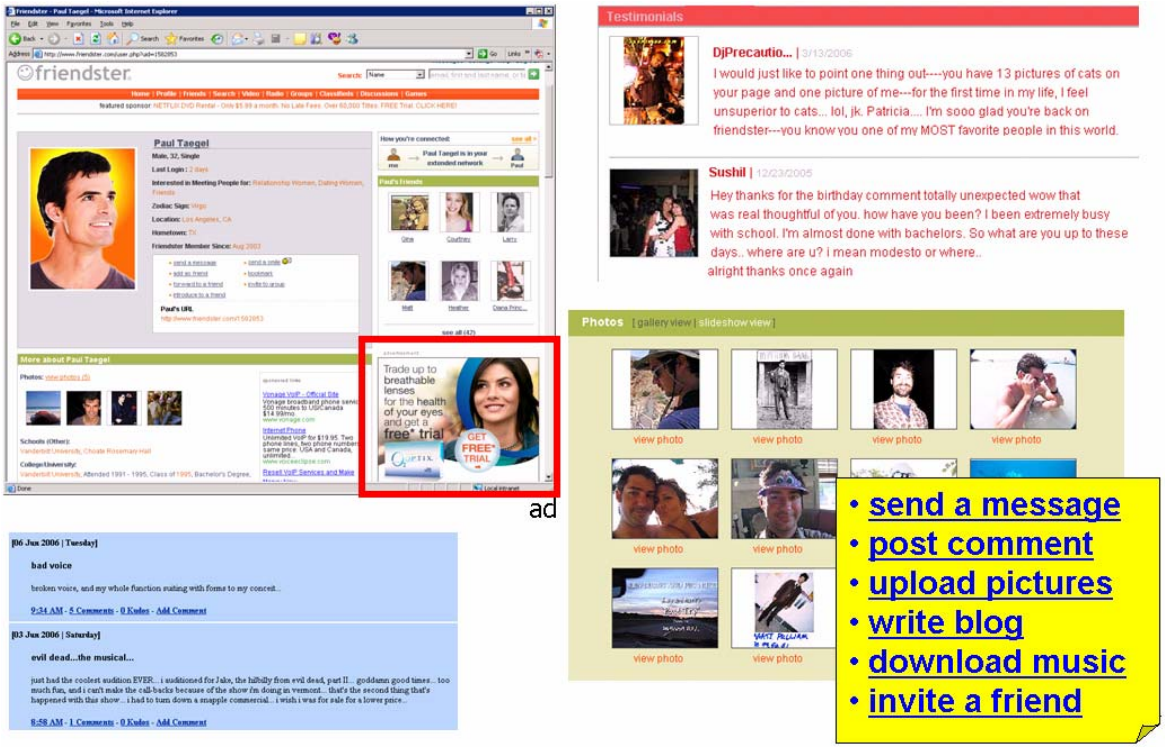


Figure 3. A Firm's View on the Network

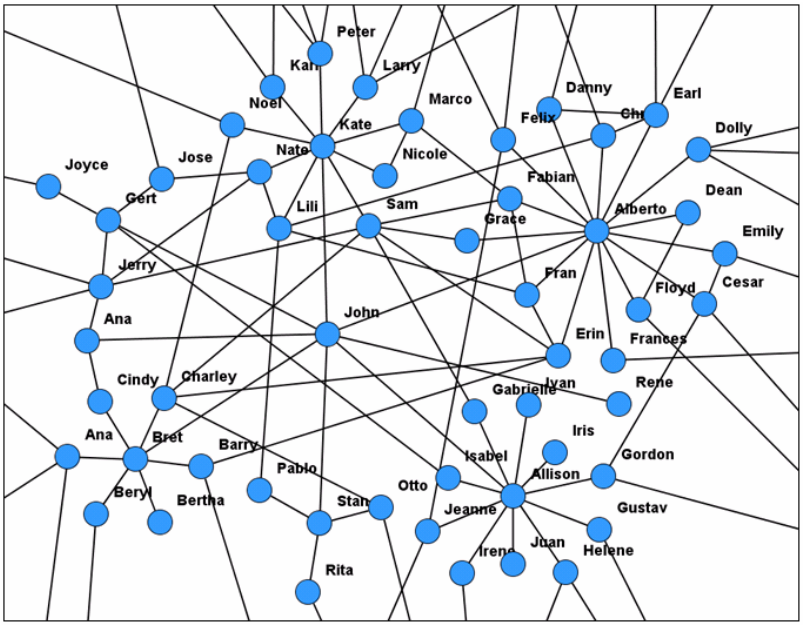


Figure 4. User (ego) Centered View.

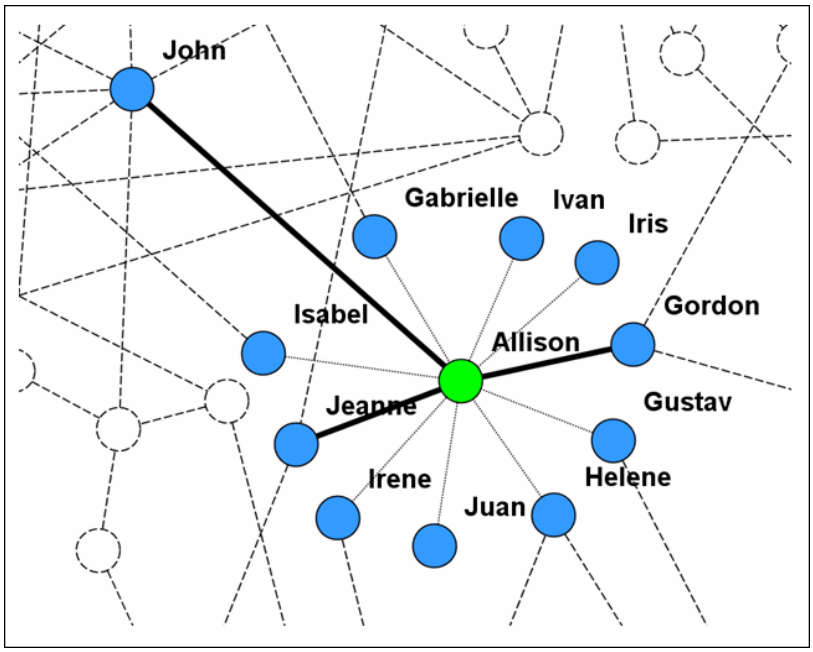


Figure 5. In Search for “Influencers”

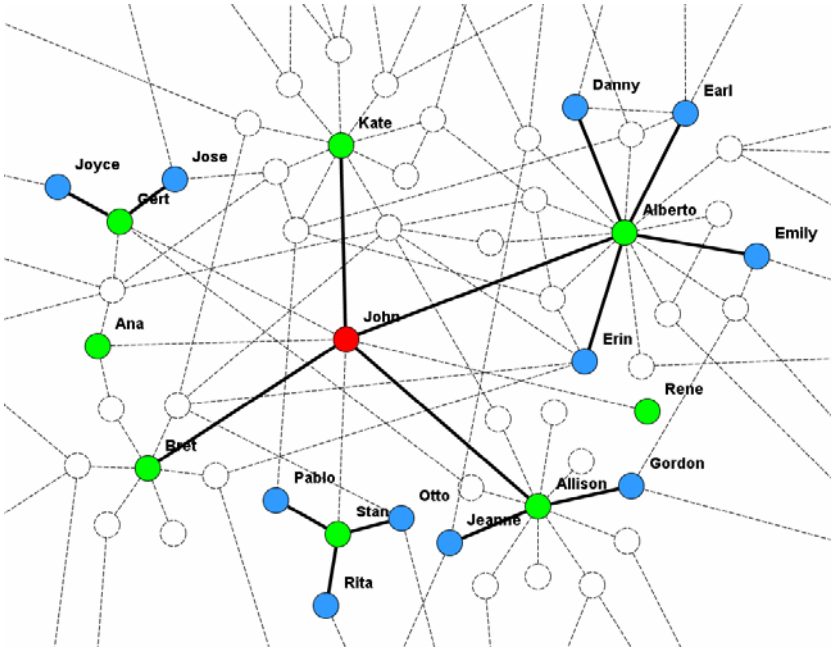


Figure 6. Knowing Who Else is Online.

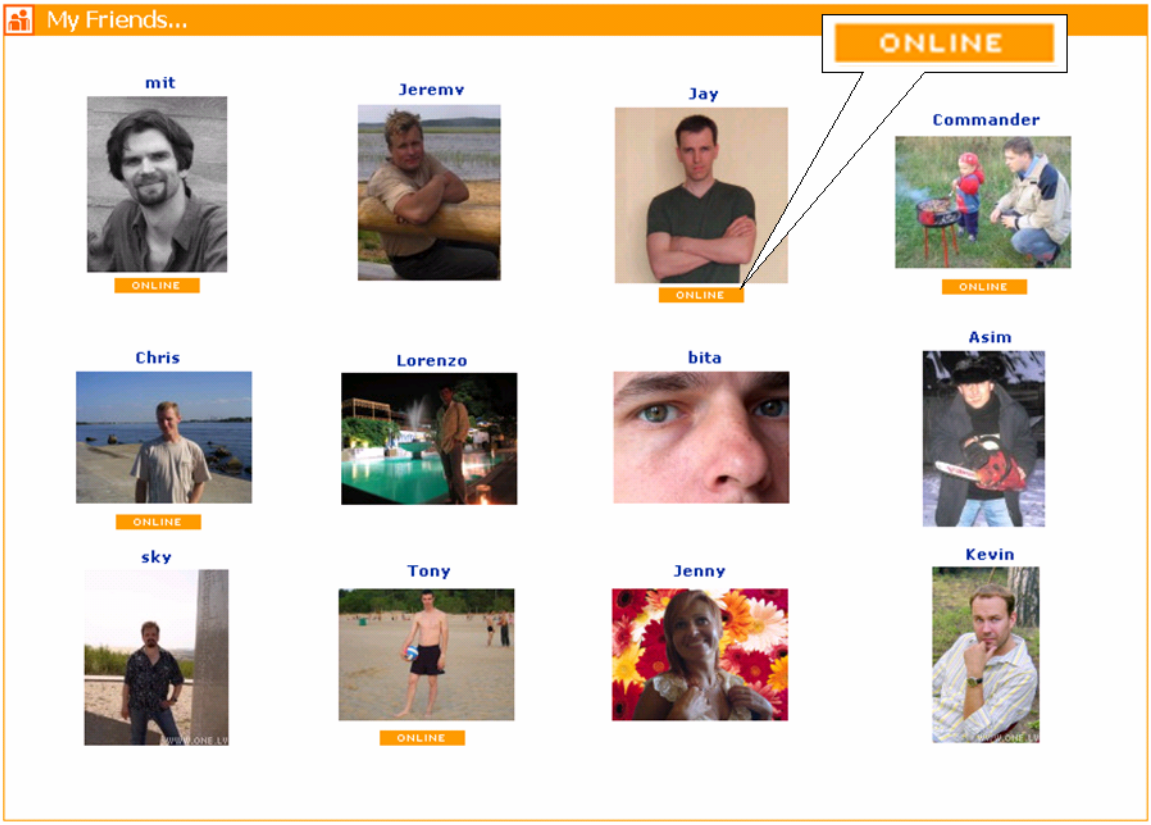


Figure 7. Mean vs. Variance in Daily Logins

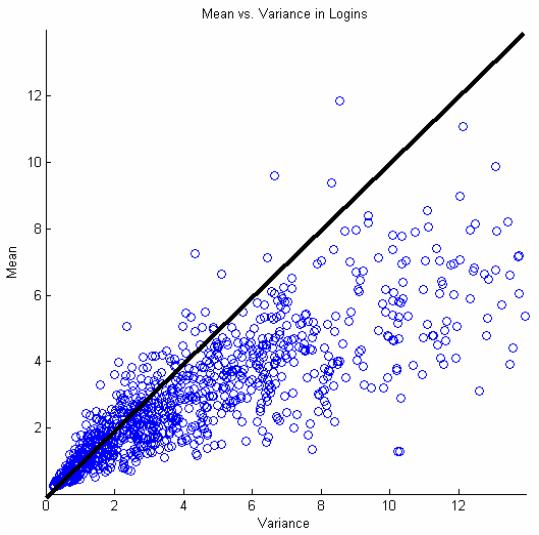


Figure 8. Draws from Posterior Distribution of Gamma for Two Friends of the Same User

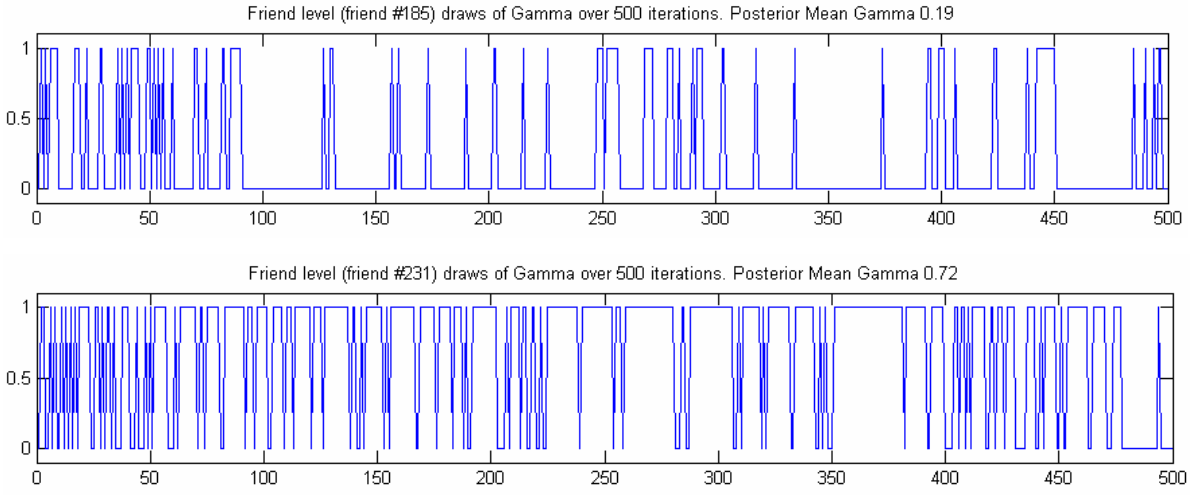


Figure 9. Users' Ethnicity

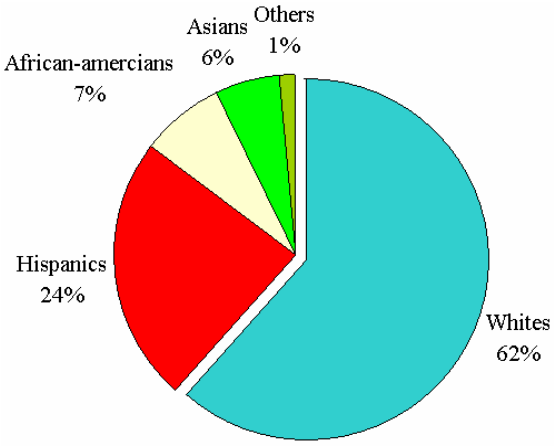


Figure 10. Users' Geographical Location and Friendship Links

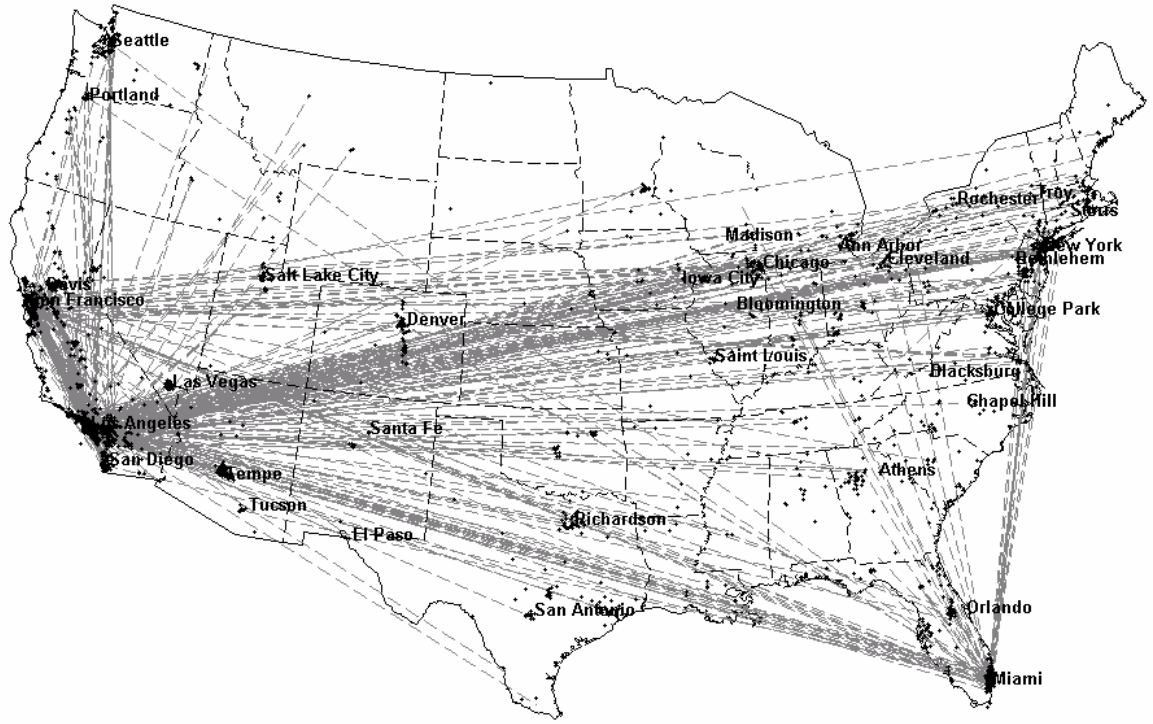


Figure 11. Distribution of Logins per Day

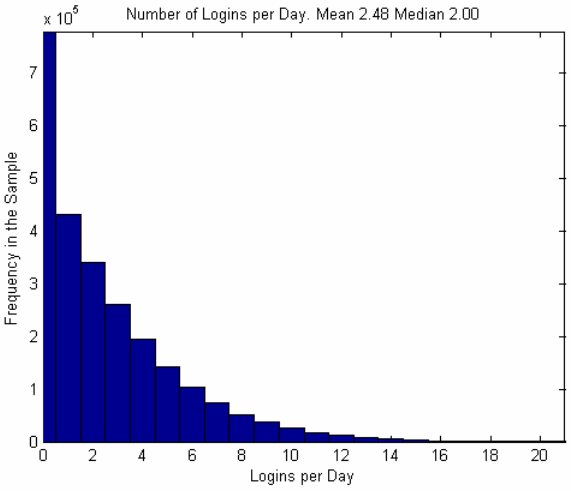


Figure 12. Login Time Series Examples for Four Users

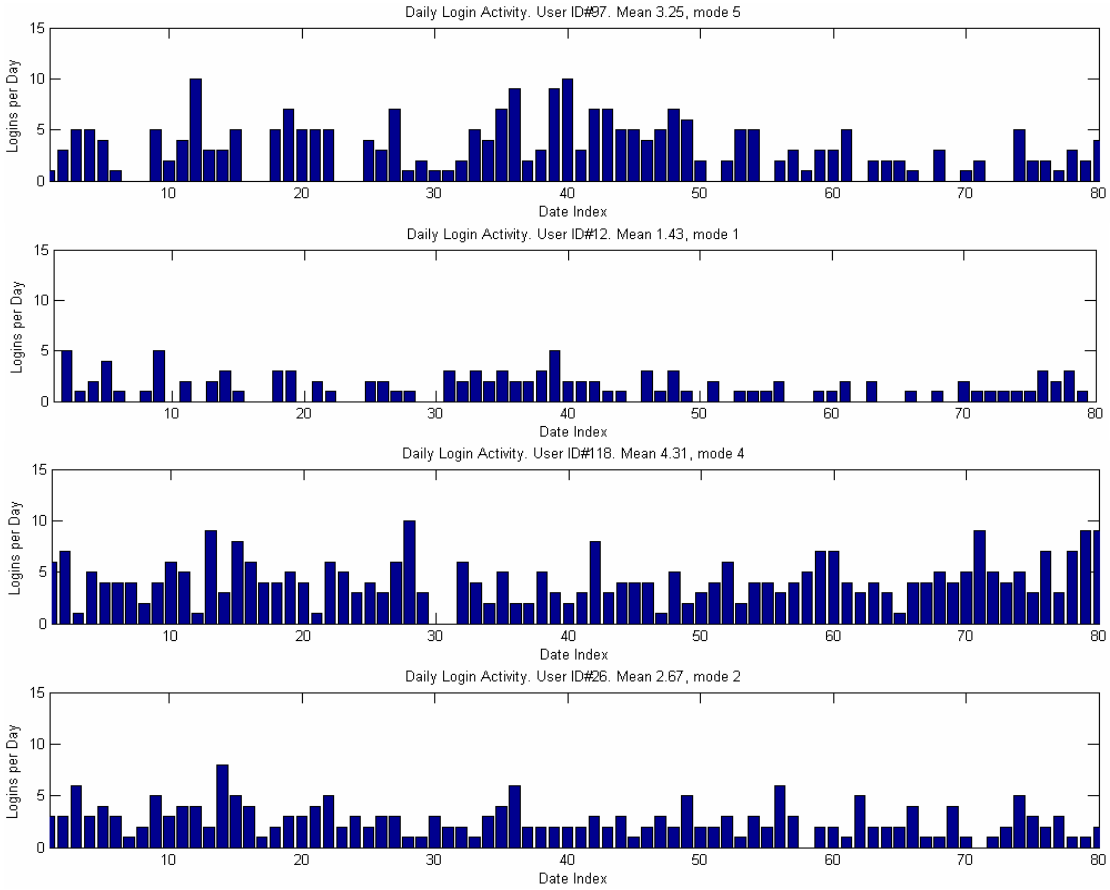


Figure 13. Estimation Results for Two Users

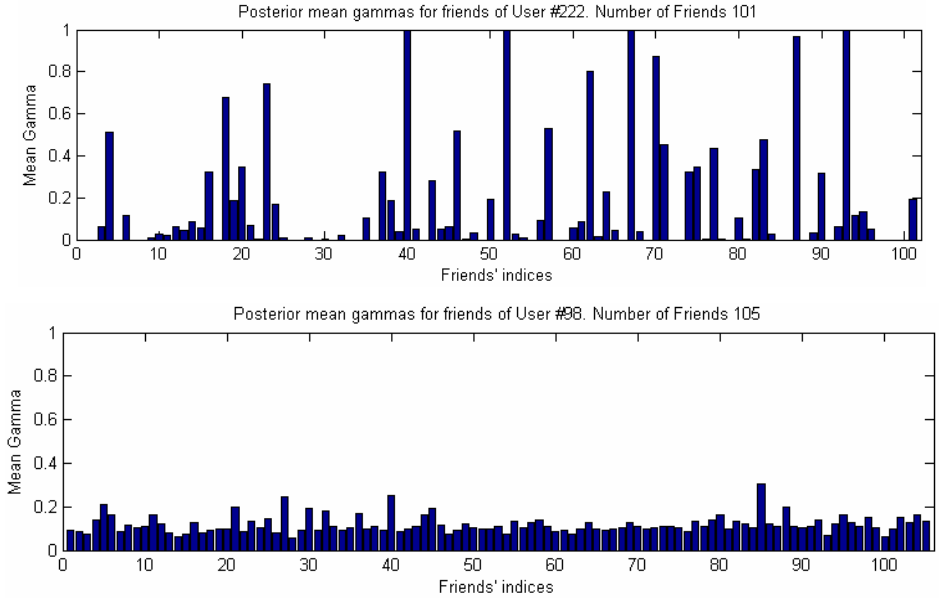


Figure 14. Distribution of Posterior Mean γ_{uf}

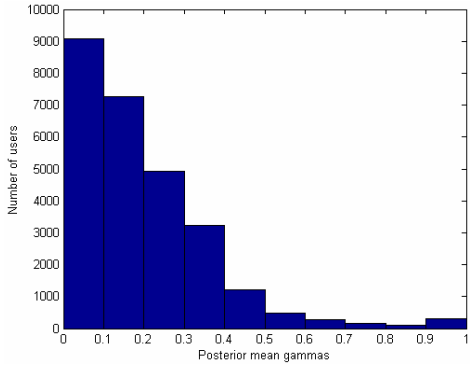


Figure 15. Distribution of Influential Users

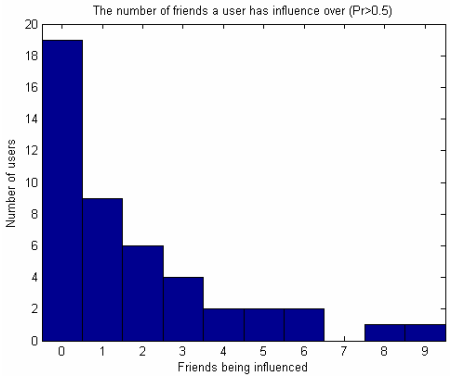


Figure 16. Number of Friends vs. Number of Friends User Influences