MICHAEL TRUSOV, ANAND V. BODAPATI, and RANDOLPH E. BUCKLIN*

The success of Internet social networking sites depends on the number and activity levels of their user members. Although users typically have numerous connections to other site members (i.e., "friends"), only a fraction of those so-called friends may actually influence a member's site usage. Because the influence of potentially hundreds of friends needs to be evaluated for each user, inferring precisely who is influential—and, therefore, of managerial interest for advertising targeting and retention efforts—is difficult. The authors develop an approach to determine which users have significant effects on the activities of others using the longitudinal records of members' log-in activity. They propose a nonstandard form of Bayesian shrinkage implemented in a Poisson regression. Instead of shrinking across panelists, strength is pooled across variables within the model for each user. The approach identifies the specific users who most influence others' activity and does so considerably better than simpler alternatives. For the social networking site data, the authors find that, on average, approximately one-fifth of a user's friends actually influence his or her activity level on the site.

Keywords: Internet, social networking, Bayesian methods

Determining Influential Users in Internet Social Networks

In 1995, when the first notable social networking (SN) Web site, Classmates.com, was launched, few might have guessed that 15 years later, SN sites would have tens of millions of users and would be valued at billions of dollars. Currently, SN sites attract more than 90% of all teenagers and young adults in the United States and have a market of approximately 80 million members. The cover of *BusinessWeek* magazine's issue dated December 12, 2005, suggests that the next generation of Americans could be called the MySpace generation. Emphasizing the importance of SN to marketing, the Marketing Science Institute (2006) designated "The Connected Customer" as its top research priority.

*Michael Trusov is Assistant Professor of Marketing, Robert H. Smith School of Business, University of Maryland (e-mail: mtrusov@rhsmith.umd.edu). Anand V. Bodapati is Associate Professor of Marketing (e-mail: anand.bodapati@anderson.ucla.edu), and Randolph E. Bucklin is Peter W. Mullin Professor (e-mail: rbucklin@anderson.ucla.edu), Anderson School of Management, University of California, Los Angeles. The authors are grateful to Christophe Van den Bulte and Dawn Iacobucci for their insightful and thoughtful comments on this work. John Hauser served as associate editor for this article.

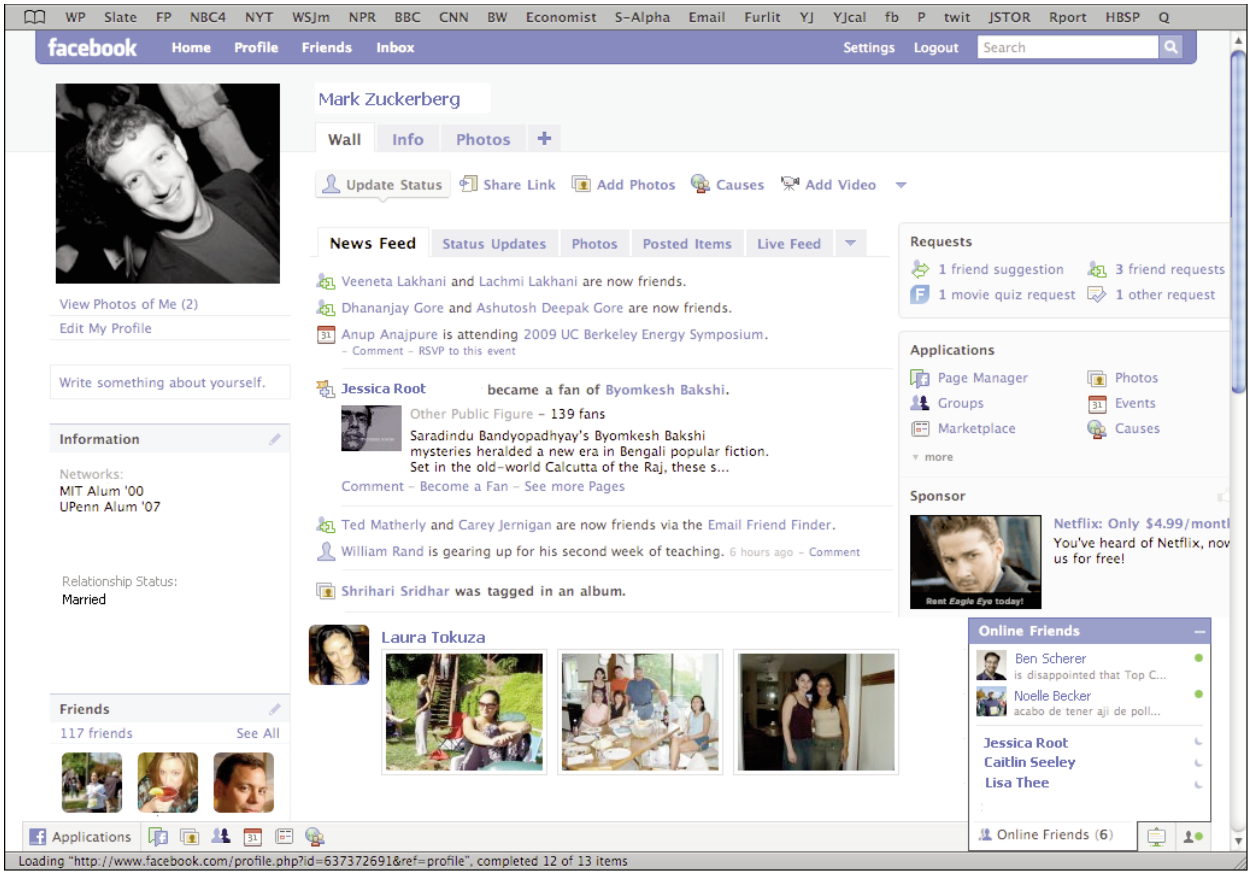
The core of an SN site is a collection of user profiles (Figure 1) where registered members can place information that they want to share with others. For the most part, users are involved in two kinds of activities on the site: Either they create new content by editing their profiles (e.g., adding pictures, uploading music, writing blogs and messages), or they consume content that others create (e.g., looking at pictures, downloading music, reading blogs and messages). On most SN sites, users can add other users to their networks of "friends." Usually, one user initiates the invitation, and the other user accepts or rejects it. When accepted, the two profiles become linked.

The most popular SN site business model is based on advertising. As users surf through a site, advertisements are displayed on the Web pages delivered to the users. Social networking firms earn revenue from either showing advertisements to site visitors (impressions) or being paid for each click/action taken by site visitors in response to an advertisement. Consequently, user involvement with a site (e.g., time spent on the site, number of pages viewed, amount of personal information revealed) directly translates into firm revenue. Social networking firms have commonly used members' profile information for ad targeting purposes.

© 2010, American Marketing Association
ISSN: 0022-2437 (print), 1547-7193 (electronic)

Journal of Marketing Research
Vol. XLVII (August 2010), 643–658

Figure 1
PROFILE EXAMPLE



THE ROLE OF USERS IN USER-GENERATED CONTENT ENVIRONMENTS

At SN sites, the content is almost entirely user generated. To attract traffic, an SN firm itself cannot do much beyond periodic updates of site features and design elements. The bulk of digital content—the driving force of the site's vitality and attractiveness—is produced by its users. However, users are not all created equal. Community members differ widely in terms of the frequency, volume, type, and quality of digital content generated and consumed. From a managerial perspective, understanding who keeps the SN site attractive—specifically, identifying users who influence the site activity of others—is vital. Such understanding permits more precise ad targeting as well as retention efforts aimed at sustaining and/or increasing the activity of influential existing users (and, therefore, future ad revenue).

The importance of identifying influential users on an SN site has been recently highlighted by Google's efforts to improve ad targeting at MySpace.com. Apparently disappointed with the returns from its recent deal to place advertising on MySpace.com, Google is developing algorithms to improve the identification of influential users (Green 2008). The idea is to target advertisements at site members who get the most attention, not simply those with certain characteristics in their profiles. Ad buyers have indicated that they would pay premium rates for this type of influence-based targeting. Google has filed a patent application, having apparently adapted its PageRank algorithm to the problem (Green 2008). The financial implications of improved ad targeting are significant because display ad pricing varies widely with audience attractiveness. For example, Walsh (2008) reports cost per thousand impressions varying from $.05 to $.80 or more.

Our objective is to develop and test a methodology to identify influential users in online social networks on the basis of a simple metric of their activity level. In this article, we consider a user "influential" in a social network if his or her activity level, as captured by site log-ins over time, has a significant effect on others' activity levels and, consequently, on the site's overall page view volume. An example of this type of influence is given by Holmes (2006), who reports that when a popular blogger left his blogging site for a two-week vacation, the site's visitor tally fell, and content produced by three invited substitute bloggers could not stem the decline.

EMPIRICAL INFERENCE OF SITE USAGE INFLUENCE

Most existing studies on social influence adopt a survey approach. Although questionnaires may work well for small groups, for an online community with millions of members, surveys are problematic. Fortunately, in computer-mediated environments, users' online activities can be tracked and recorded. The model we develop infers site usage influence from secondary data on member log-in activity. It can be easily extended to other online activity measures that are

potentially available to SN site managers, such as the amount of time spent on the Web site and the number of messages sent. In this study, we use the number of log-ins per day as an effective correlate for these other measures.

Our approach is based on the following logic: Users log in to the site to consume new digital content that other users produce. From the traces others leave (e.g., the last log-in date and time, time-stamped content updates), users can infer how active a specific person has been on the site and update their expectations for future activity. Logging in is more attractive when a user expects that there is likely to be new content to view. Accordingly, we propose that a member's site usage level at each point is driven by his or her expectation about the volume and update frequency of relevant new content created by others. User expectations are formed from recent experiences. For example, if, for the past few log-in occasions, the user observed an increase in volume and/or update frequency of new content, he or she might choose to wait less time before logging in again.

In our data set, which comes from a major SN site, we tracked daily log-in activities of anonymous community members, treating the frequency of log-ins as a proxy for usage. A higher number of log-ins per day is taken to be a sign of higher usage, while a lower number of log-ins implies lower usage. In addition, because both processes—content consumption and content creation—constitute site usage, we assume that during high-usage days, the user has more opportunities to produce more content than during low-usage days. 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 linked to him or her. If a member increases his or her usage and the people connected also increase their usage (possibly because of their interest in what this person is creating), we propose that this identifies this person as influential. Conversely, if a member's usage goes up or down and usage does not change among the people connected to him or her, we propose that this person is not influential.

With the data and this notion of influence, we attempt to identify users whose behavior on the site has the most significant impact on the behavior of others in the network. Note that this formalization of "influence" fits well with the firm's business objectives. Managers need to know who stimulates activity levels among other members of the network. As expected, our empirical results show significant heterogeneity among users on two dimensions: susceptibility to influence from other users and the extent of influence on others in the network. The findings indicate that social influence in an online community is similar to what, according to studies in sociology, is experienced offline. The average user is influenced by relatively few other network members and, in turn, influences few people. In addition, having many friends (i.e., linked profiles) does not make users influential per se.

The problem of identifying influential users with site activity data is difficult because the data are typically sparse relative to the number of effects that need to be evaluated. Thus, simple approaches to effects estimation do not do well (as we demonstrate subsequently). A common way to address the sparse number of individual-level observations in, for example, UPC (Universal Product Code) scanner panel data is to use Bayesian shrinkage, in which strength is pooled across panelists. This approach is grounded on the assumption that the effect of a certain variable (e.g., price) for a given panelist is related to the effect of the same variable on other panelists. Unfortunately, we cannot use this type of Bayesian shrinkage here because the variable set differs across users. At an SN site, each user has a unique set of friends, potentially hundreds, and it is the friends' activity levels that make up the variable set. To accommodate this, we propose to use a different type of Bayesian shrinkage in which strength is pooled across variables within an individual rather than across individuals.

We organize the rest of this article as follows: In the next section, we provide a brief overview of the study context and introduce key terminology. Then, we touch on two streams of related literature: social influence and online communities. We continue with a description of the proposed method, model formulation, and empirical estimation on field data from an SN site. Next, using simulated data, we assess the ability of our model to recover true levels of user influence. We then discuss managerial implications and illustrate the potential financial benefits of applying the approach versus some naive methods. We conclude with a summary, limitations, and directions for further research.

THE STUDY CONTEXT AND KEY TERMINOLOGY

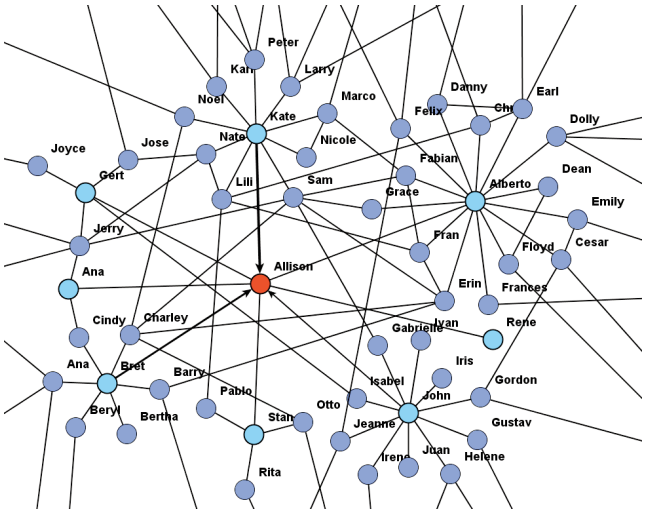
A profile holder at an SN site can acquire new friends by browsing and searching the site and sending requests to be added as a friend. The resulting "friendship" network can be represented by a connected, undirected graph with binary edges. As an example, we focus on a specific person in a hypothetical network, Allison (Figure 2). Allison's ego-centered network is a network of her friends. In Figure 2, Panel A, the friends (the people with whom Allison has exchanged invitations) are Kate, Alberto, Rene, John, Stan, Bret, Ana, and Gert. Among these friends, there are just a few 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. She also updates her profile and posts new content motivated by the expectation that her important friends will view these updates. In turn, from the perspective of some of her friends, Allison also might be important. It is possible that some of Allison's friends are regularly checking for content she produces or are motivated to contribute new content in the hope that Allison will view it. In this sense, Allison's online activity influences their behavior. The goal in this study is to develop a method to estimate the extent and the direction of the influence associated with each edge in the graph.

The study of relationships among interacting units is a cornerstone of social network analysis (SNA)—a set of theories and methods that enable the analysis of social structures; these methods are specifically geared toward an investigation of the relational aspects of these structures (Scott 1992).[1] Usually, the importance of an individual actor (in this case, a community member) can be inferred from his or her location in the network (e.g., Iacobucci 1990, 1998; Iacobucci and Hopkins 1992). On most SN sites, the network is based on friendship, or links established through exchange of electronic invitations. Because these links are easily observable by the firm, it might be tempting to apply SNA to infer a person's importance. This would likely

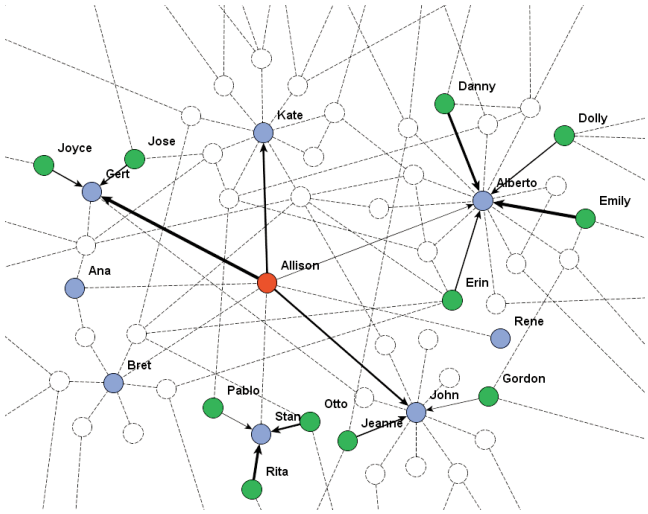
[1]For a broad overview of SNA applications in marketing, see Iacobucci (1996) and Van Den Bulte and Wuyts (2007).

Figure 2
INFERRING A USER'S INFLUENCE

A: Allison's Ego-Centered Network of Friends



B: Allison's Network of Influence



imply that a person who has more linked profiles is more important than someone with fewer links.[2]

However, analyzing the network of friendship links might not be the best alternative when a firm wants to know who is important in terms of influencing site usage. A link between two profiles on an SN site does not necessarily imply influence.[3] To study a user's influence, we need to use a network in which the relationships represent influence on site usage. Therefore, a better input for SNA techniques would be a network in which the link from Actor A to Actor B has a weight proportional to User A's influence on User B's site usage. Our approach is intended to infer such a network. We can then use SNA tools to evaluate a person's importance in this influence-based network, in addition to, or in place of, a person's importance based on, for example, network connections and nonconnections alone.

LITERATURE

Social influence occurs when a person adapts his or her behavior, attitudes, or beliefs to the behavior, attitudes, or beliefs of others in the social system (Leenders 2002). 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. 2004). Influence does not necessarily require face-to-face interaction but rather is based on information about other people (Robins, Pattison, and Elliott 2001). In an online community, information is passed among individual users in the form of digital content. Here, we consider a particular type of social influence that takes place in an online community—namely, when members change their site usage in response to changes in the behavior of other members.

Though a relatively new area in marketing research, online communities have attracted the attention of many scholars. Dholakia, Bagozzi, and Pearo (2004) study two key group-level determinants of virtual community participation—group norms and social identity—and test the proposed model using a survey-based study across several virtual communities. Kozinets (2002) develops a new approach to collecting and interpreting data obtained from consumers' discussions in online forums. Godes and Mayzlin (2004) and Chevalier and Mayzlin (2006) examine the effect of online word-of-mouth communications. Dellarocas (2005) analyzes how the strategic manipulation of Internet opinion forums affects the payoffs to consumers and firms in markets of vertically differentiated experience goods. Narayan and Yang (2006) study a popular online provider of comparison-shopping services, Epinions.com, and model the formation of relationships of "trust" that consumers develop with other consumers whose online product reviews they consistently find to be valuable. Finally, Stephen and Toubia (2010) examine a large online social commerce marketplace and study economic value implications of link formation among sellers.

The current research's contribution lies at the intersection of social influence and online communities. First, we believe that online SN sites are a unique type of online community. Some aspects of socializing in the virtual worlds of MySpace and 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). However, the dissimilarities (e.g., the number of people involved, the motives for and nature of interactions, the revenue-generating models) are too numerous to treat them the same way. Second, previous research has not examined peer influence on individual-level site usage, the focus of our study. Finally,

[2]Social network analysis can also use other structural measures to flag possibly "important" network players.

[3]Numerous anecdotal examples from press and personal interviews with SN site participants suggest that for many members, having their profiles linked to a large number of other users is a matter of prestige or competition rather than a sign of importance or popularity. In addition, people often accept invitations from others just to avoid seeming impolite. Thus, a network of friends might consist of a network of total strangers with whom almost no interaction takes place.

from a methodological perspective, none of the aforementioned empirical studies simultaneously model individual-level influence within a group of users. They either focus on interaction within a dyad (e.g., Narayan and Yang 2006) or consider aggregated group-level measures (e.g., Dholakia, Bagozzi, and Pearo 2004; Godes and Mayzlin 2004).

MODELING AND ESTIMATION

Our objective is to estimate the influence of each SN site member on the site usage of other members. For an online community with N members, we would need to evaluate $N \times (N - 1)/2$ possible pairs of users on two dimensions: direction and strength. For a real-world network with millions of members, this task is both infeasible and unnecessary. In a typical large online community, most members never interact and are not even aware of one another's existence. To take advantage of this sparseness, we used a prefiltering condition that significantly reduces the number of potential connections to be evaluated. We argue that a good candidate for a prefiltering condition is the existence of an explicit connection between profiles established through an invitation mechanism.[4] Accordingly, we estimate the direction and strength of an influence only for profiles linked by a friendship connection.

A person is part of a user's "first-level network" of friends if the person has a friendship connection with the user in the sense we described previously. The person is taken to be in the user's "second-level network" of friends if he or she is not part of the user's first-level network but is in the first-level network of someone else who is in the user's first-level network. Higher-level networks are defined in a similar way. In our modeling approach, we consider a user's activity independent of the activity of second-level friends conditional on the activity of first-level friends. This does not imply that a second-level friend has no effect on a user. Rather, it means that a second-level friend has an effect only through a first-level friend. Formally, this means that first-level friends' activities represent sufficient statistics for other users' activities. For example, in Figure 2, Panel A, Allison may have an effect on Emily, but Alberto's activities are sufficient for characterizing this effect. This is a natural assumption to make, given the nature of the network; a closely analogous assumption is typically made in spatial statistics models. (Note that this assumption would be flawed if the network structure was misspecified, as when Allison and Emily know each other directly outside the online social network and there are interactions other than those through Alberto.)

The assumption of conditional independence from second- and higher-level friends enables us to adopt an ego-centered approach to the analysis. Taking one user of an online community at a time, we search for influential friends within the user's ego-centered network. This answers the question, Who is influencing a given user? For example, in Figure 2, Panel A, we treat Allison as an ego and evaluate the impact of her friends—Kate, Alberto, Rene, John, Stan, Bret, Ana, and Gert—on her site usage. The outcome is that Bret, Kate, and John influence Allison to different degrees, while the others have no impact on her. Repeating this process for every person in the network, we treat each user as an ego once and as a potential "influencer" the number of times equal to the number of friends he or she has. In Figure 2, Panel B, a by-product of the ego-centered analysis for Kate, Alberto, Rene, John, Stan, Bret, Ana, and Gert is Allison's influence on them. Here, Allison has a strong impact on Gert, a moderate impact on Kate and John, and a weak impact on Alberto. This answers the second question, Who is influenced by the same given user? In network terms, the procedure reconstructs a full influence-based network by performing a series of ego-centered estimations.

Model Specification

To identify influential friends within an ego-centered network, we model a user's log-in activity as a function of the user's characteristics, the user's past behavior on the site, and the log-in activity of the user's friends. We suggest that the count of individual daily log-ins 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 log-ins y_{ut} as a Poisson regression:

$$y_{ut} \sim \text{Poisson}(\lambda_{ut}). \tag{1}$$

We derive the Poisson regression model from the Poisson distribution by specifying the relationship between the rate parameter λ_{ut} and predictors (e.g., Cameron and Trivedi 1998). We group the predictors into two sets: self effects and friend effects. Self effects include covariates such as user-specific intercepts, day of the week, and past log-ins. Friend effects consist of friends' lagged log-in activity. It is customary to use exponential rate parameterization, giving the logarithm of the rate parameter as follows:

$$\log(\lambda_{ut}) = \text{Self Effects}_{ut} + \text{Friend Effects}_{ut}. \tag{2}$$

More specifically, we have

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

where

x_{ukt} = user-specific covariate k (e.g., intercept, day-of-the-week effect, log-ins at t – 1),
z_{uft} = weighted average of lagged log-in 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.

Equation 3 specifies that a user's site usage at any given time depends, among other things, on the site usage of this person's friends. The proposed model captures this process through the friend-specific coefficients β_{uf}. If friend k is among the user's "important" friends (i.e., his or her activity level has an impact on the site attractiveness for the user), the corresponding coefficient β_{uk} will be significantly

[4]Some other alternatives, such as cross-profile visitations and message exchanges, might also be considered candidates for the prefiltering condition, but unfortunately, these are not available to us in the data set. However, this limitation of the data set does not present a serious problem in this research, because the bulk of the interaction on a typical SN site occurs among profiles that have been connected through invitations. Moreover, an absence of invitation-based connections between two profiles often imposes serious limitations on a level of interaction (e.g., ability to consume and/or exchange digital content).

different from zero. We note that learning about the activity levels among a user's friends is likely to take time. Thus, for each friend f of user u, we construct a covariate z_{uft} as a weighted average [$w_u(d)$] of friend f's log-in activities over the past D days. This is given in Equation 4:

$$(4)\qquad Z_{uft} = \sum_{d=1}^{D} w_u(d) \times y_{f(t-d)}, \text{ where} \sum_{d=1}^{D} w_u(d) = 1,$$

where $y_{f(t-d)}$ is the number of log-ins for friend f at time t – d.

We also adopt an exponential smoothing expression for $w_u(d)$, following the spirit of Guadagni and Little (1983). We define $w_u(d)$ (Equation 5) as a function of lag d and a smoothing parameter ρ_u:

$$(5)\qquad w_u(d) = \frac{\exp(-d \times \rho_u)}{\sum_{k=1}^{D} \exp(-k \times \rho_u)}.$$

We believe that the past actions of friend f are likely to have a diminishing-in-time impact on user u's activity level at time t. So, we expect $w_u(d)$ to be a decreasing function in d. We allow for full heterogeneity in the smoothing parameter across users. The flexibility of this approach allows the weight distribution over the past D days to vary from being equal for all lags when $\rho_u \to 0$ to the entire mass concentrating on the first lag for larger ρ_u (e.g., when ρ_u equals 5 the first lag receives 99.3% of the weight). Figure 3 plots a few examples of $w_u(d)$ for different values of ρ_u. We preset D to be seven days.[5]

Among several alternative specifications for count data models, we chose the Poisson regression because it integrates well with the variable selection algorithm (described subsequently) and results in a parsimonious and scalable solution, which is critical in real-world applications. We benchmarked the performance of the Poisson model against the more flexible negative binomial distribution specification (sometimes preferred to the Poisson because of the equidispersion assumption) and did not find any significant difference in the estimation results.

A limitation of the model is that explosion is a theoretical possibility. For example, consider a situation in which one user amplifies another user's future activity and, in turn, this other user amplifies the first user's activity. This situation can create a positive feedback loop that causes each user's activity level to increase indefinitely.[6] We believe that the characteristics of the data and the estimation results make the risks of this relatively small. First, we observe that user activity levels are roughly equivalent at the start and end of the 12-week observation period. This suggests that explosion, if any, is slow. Second, the estimated directions of influence between users are primarily unidirectional; that is, when user u influences user u′, user u′ does not influence user u. This also greatly diminishes the likelihood of explosion.

Estimation Challenges

In the data, the typical user has approximately 90 friends, and many have hundreds of friends. The panel has approximately 80 observations for each user. Thus, we encounter the "large p, small n" situation in which the number of parameters to be estimated is large relative to the number of observations. This means that these parameters cannot be reliably estimated in a fixed-effects framework. One way to address this is to estimate the model specified in Equation 3 but only for each user–friend pair at a time (controlling for other user-specific covariates x_{uk}). A problem with this approach is that the activity levels of a user's friends are often correlated. The bias from omitting the effects of the user's other friends is likely to produce inflated estimates for the influence of each friend. (We actually observe this effect in the simulation results described subsequently.)

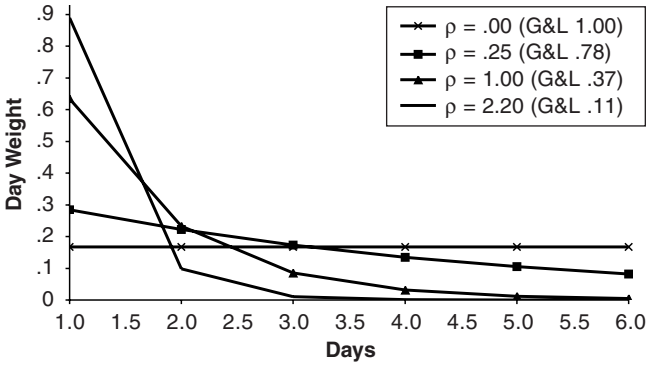
An alternative approach is to use a random-effects framework in which we pool strength across parameter estimates from multiple samples through Bayesian shrinkage. This has been widely done in the analysis of scanner panel data when estimating a given household's response coefficients for marketing variables (e.g., price and advertising). The modeling assumption is that the response coefficients for the various households are drawn from a distribution so that the coefficient values for other households reveal something about the coefficient values for the given household.

In our setting, however, we cannot apply the usual type of Bayesian shrinkage. In the scanner panel situation, the variable set (e.g., price and advertising), whose effects we are trying to determine, is constant across households. In the social network situation, the variables correspond to friends, and different users have different sets of friends. This means that the variable set differs, often completely, across users.

To address this challenge, we propose to shrink not across users but across friends within a user. To do this, we need to choose the across-friends distribution of the β_{uf} terms in Equation 3. A key consideration is that the average user probably tracks just a few other friends. Thus, for a given

[5]Although we can also estimate D as part of the model, the results are not sensitive to changes in D for values greater than 5. We also acknowledge that D could be made user specific; however, this is impractical because it slows down the estimation procedure without providing any significant benefits.

Figure 3
EXAMPLES OF WEIGHTS w(d) DISTRIBUTION ACROSS D LAGS FOR DIFFERENT ρ_u



Notes: G&L = Guadagni and Little (1983).

[6]We thank an anonymous reviewer for making this important point.

user, most of the β_{uf} coefficients in Equation 3 are likely to be zero. This makes it prudent to choose the across-friends distribution of the β_{uf} terms to have a point mass at zero. For example, we could consider a mixture density in which one component has a point mass at zero and another component is a Gaussian distribution with a mean and variance to be estimated. Alternatively, we could consider a latent class model in which the mixture density consists of multiple point-mass densities with one point located at zero. We elect to pursue the latent class approach here primarily because the estimation algorithm is of high computational efficiency. This also is important for scalability in the applied use of our approach.

To implement a latent class model, the analyst needs to select the number of mass points. Fortunately, this selection can be based on several well-known criteria, including the deviance information criterion (DIC) (Spiegelhalter et al. 2002), the Bayes factor, and cross-validation. In the empirical application, models with two or three point masses perform about the same (as we detail subsequently). Thus, we focus the following discussion on the two-class case—that is, we assume that the across-friends distribution of influence coefficients β_{uf} has a mass point at zero and another somewhere else. We need to estimate the location of this other mass point and its weight. We then use this generating distribution to pool strength across friends to estimate the values of the influence coefficients for each specific friend. (In the Appendix, we describe the algorithm for the general case with more than two point masses.)

We can decompose each friend-specific coefficient from Equation 3 into a user's susceptibility to friends' influence, denoted by β_u, and a binary parameter γ_{uf}:

$$\beta_{uf} = \beta_u \times \gamma_{uf}. \tag{6}$$

The binary parameter, γ_{uf}, is 1 if friend f is influential and 0 if otherwise. This approach enables us to parsimoniously capture two phenomena: (1) A friend is either influential or not, so β_{uf} becomes either zero or not, and (2) the susceptibility to friends' influence, β_u, can vary from user to user. All the model parameters can be drawn from the conditional posterior densities given the other parameters' values, and we detail these in the Appendix.

We now describe how γ_{uf} is drawn because it is an unconventional part of our Gibbs sampler:

$$\gamma_{uf,\forall f} \mid \bullet \propto \text{Bin}\left(\gamma_{uf} \mid \frac{c_{uf}}{c_{uf} + d_{uf}}\right), \tag{7}$$

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).

In each iteration of the Gibbs sampler, a friend-specific γ_{uf} is drawn as a Bernoulli random variable, with the success probability based on the ratio of the likelihood with friend f's effect included (i.e., with $\gamma_{uf} = 1$) to model likelihood without friend f (i.e., with $\gamma_{uf} = 0$). From the perspective of variable selection, 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 has a nonzero influence on 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 Beta($1 + IF_u$, $1 + F_u - IF_u$), which has a mean that is approximately equal to the empirical fraction of influential friends.[7]

To apply the approach at an SN site, we might need to estimate millions of ego-centered networks. The proposed decomposition in Equation 6 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, we can rewrite Equation 3 as follows:

$$\log(\lambda_{ut}) = \alpha_{u1}x_{u1t} + \alpha_{u2}x_{u2t} + \ldots + \alpha_{uK}x_{uKt} + \beta_u \sum_{f=1}^{F_u} \gamma_{uf} z_{uft}. \tag{8}$$

To draw α_{uk} and β_u, we use the Metropolis–Hastings algorithm within Gibbs sampling steps. Instead of the usual random walk, we use an independence chain sampler, in which the proposal density is a normal approximation to the posterior density from the Poisson likelihood. The likelihood considered is conditional on the realizations of γ_u on that iteration. This means that the proposal density is adaptive in that it varies from iteration to iteration as a function of the γ_u values. The proposal density for α_{uk} and β_u is a normal distribution centered at the maximum likelihood estimate for Equation 8, with the variance equal to the inverse Fisher information matrix.

ILLUSTRATING THE METHODOLOGY WITH FIELD DATA

We apply the model to data obtained from a major SN site, which wants to remain anonymous. In the 12-week data set, we track daily log-in activities for a random sample of 330 users, their 29,478 friends, and their 2,298,779 friends' friends. We refer to these groups as Level 1, Level 2, and Level 3 networks. For Level 1 and Level 2 network members, 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). For Level 3 users, we have information on log-in activity but no profile information. The average number of log-ins per day in the sample is 2.48. Figure 4 gives examples of log-in time series for four randomly selected users in the sample. Each bar on the graphs corresponds to the number of log-ins on a specific date for a specific user. The examples illustrate how site usage varies considerably from

[7]The model defined in Equation 7 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. As a possible solution, in Equation 7, we could replace p_u, the friend-independent prior probability of friend being influential, with a term p_{uf} that is specific to each friend f. The probability p_{uf} could be a function of γ_{fu}, thus making the stochastic realizations γ_{uf} a function of γ_{fu}, accommodating reciprocity. We plan to address this issue in further research.

Figure 4
LOG-IN TIME-SERIES EXAMPLES FOR FOUR USERS

one user to another (e.g., User 118 logs in an average of 4.3 times per day, and User 12 logs in an average of 1.4 times per day).

Model Estimation

We estimate the model using the previously described Bayesian approach, implemented with Markov chain Monte Carlo (MCMC) methods. To complete the model specification, we introduce priors over the parameters common to all users (see the 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 model in Equation 3, we perform ego-centered estimations for all users in the Level 1 and Level 2 networks. Because of the limited information on the Level 3 network, we present the findings in the following sequence: First, we focus on the 330 Level 1 network users. The objectives are to highlight the importance of peer effects in predicting individual user behavior, to demonstrate variations in the probability of influence across friends, and to show how profile information can be used to explain these variations. Second, we perform ego-centered analyses for Level 2 network users. Note that by construction, members of the

Level 1 network are (1) friends of Level 2 network users and (2) members of the Level 3 network. As a by-product of this analysis, we obtain estimates for the influence of Level 1 users on Level 2 users.

Level 1 network estimation results. In addition to Equation 3, we estimate two benchmark models. Model 1 incorporates self effects only. Model 2 includes the effect of all the user's friends (i.e., all the γ_{uf} are set equal to 1). Model 3 is the proposed model of Equation 3. For model fit and comparisons, we use the DIC. As Table 1 shows, the proposed model provides a significantly better fit to the data than the two benchmarks.

On an individual level, we observe the anticipated heterogeneity among users in terms of the number of influential friends. In Figure 5, we show two users with a similar number of friends, approximately 100 each, but distinct patterns for influential friends. Bars on these graphs correspond to the posterior probability that a particular friend influences the user. For the first user, we observe a few influential friends, and for the second, there are none. Figure 6 shows the empirical distribution of the posterior mean γ_{uf} for the entire sample. The distribution is considerably skewed to the left, which indicates that most of the posterior means γ_{uf} are relatively small. In other words, friends corresponding to these small γ_{uf} have a low probability of influencing site usage.

Table 1
MODEL FITS

Model	*Description*	*Fit (DIC)*
Model 1	User self effects only (none of friends are influential)	76,613.66
Model 2	Self plus all friends' effects (all friends are influential)	76,010.93
Model 3	Self and friends effects with variable selection (some friends are influential)[a]	72,863.45

[a]The DIC for a three point masses model is 72,294.40, which is only slightly better than a DIC of 72,863.45 in a two point masses case. Thus, for expositional ease, we focus the discussion on the two-class case. The Appendix describes the algorithm for two point masses and for the general case of K point masses.

A key construct of interest is $\Sigma_{f=1}^{F_u}\gamma_{uf}$, the total number of influential friends for user u. A good point estimate for this construct is given by computing its expected value over $f(\gamma)$, the posterior joint density of all the gamma terms across friends. The expression for this expected value is as follows:

$$\text{Infl}_u = \int_{\gamma_u} \sum_{f=1}^{F_u} \gamma_{uf} \times f(\gamma) d\gamma_u. \tag{9}$$

This integral is estimated by Monte Carlo using the draws of γ_{uf} over the MCMC iterations. To aid interpretability, we also compute $S_u = \text{Infl}_u/F_u$, which is a point estimate for the fraction of friends who influence user u. A sample average S_u is approximately 22%, and a sample average Infl_u is approximately ten people. On average, about one of five friends within an ego-centered network significantly affects the log-in decisions of ego.

If a user has no friends influencing him or her, Model 2 should be better than Model 3. The DIC (calculated individually for each ego-centered network) shows no decrement in fit for 32% of people in the Level 1 network as we go from Model 3 to Model 2. Behaviorally, this suggests that 32% of users do not have any "influential" friends whose log-in activity on the site would help explain variations in their site usage.

The probability of being influential in a dyad (γ_{uf}) might be explained by static measures available from user profiles. To investigate this, we conduct an exploratory posterior

Figure 5
ESTIMATION RESULTS FOR TWO USERS

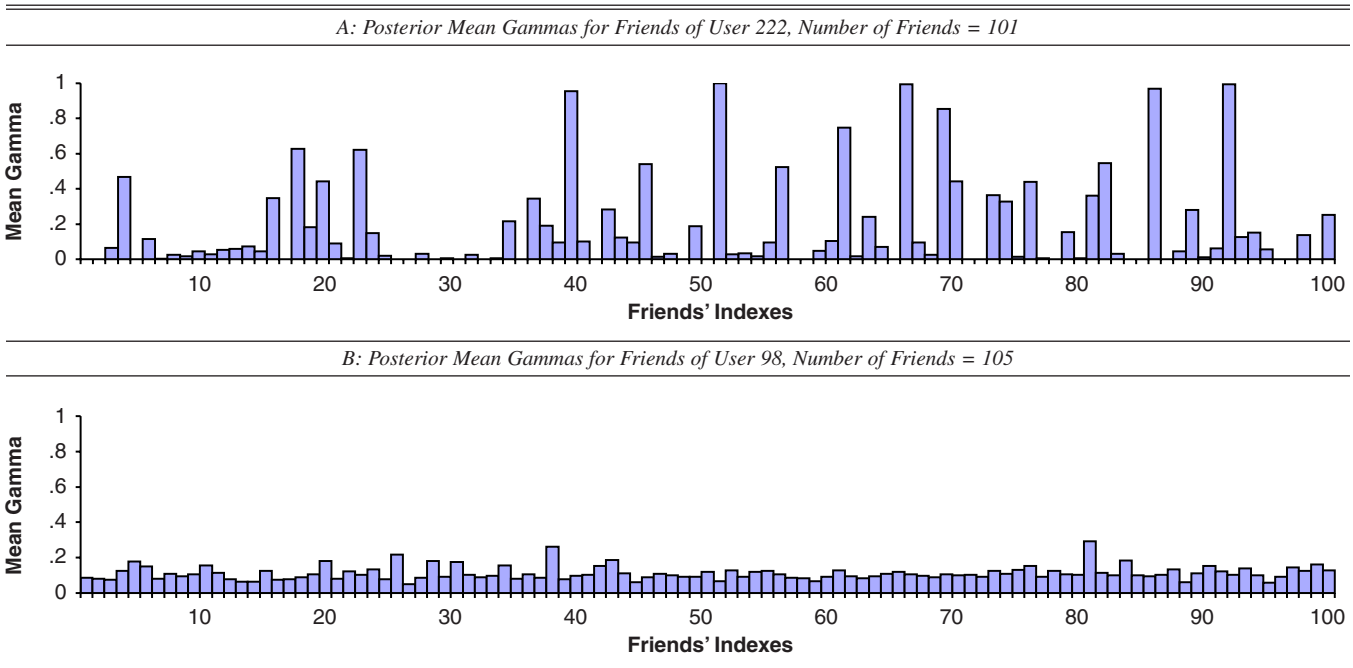
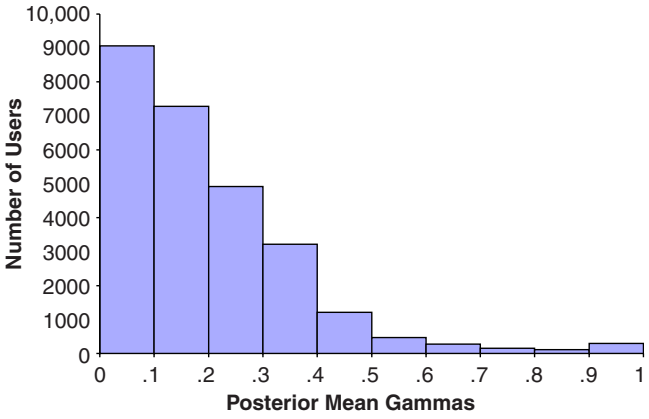


Figure 6
DISTRIBUTION OF POSTERIOR MEAN γ_{uf}



analysis.[8] We regress the logit-transformed posterior values for γ_{uf} on several covariates extracted from the user profiles: gender combination for the dyad (in this case, we chose male user, female friend), months the friend has been a member, same ethnicity for user and friend, user's dating objective, and relative age.

Table 2 gives the results from the regression. We find that female friends tend to influence the site usage of male users more than other gender combinations. Users who have been members of the site longer have more influential member friends. A user is more influenced by a member friend of the same ethnicity. Users with dating objectives have fewer friends who are influential. Finally, friends who are older than the user have less influence.

These exploratory results suggest that posterior analysis based on profile information could help the firm predict the probability of influence for any dyad in the network (because all members provide profile data at sign-up). However, note that the profile data we analyze here explains relatively little of the total variation in influence—the R-square for the regression is .11, and it does not improve with the addition of other profile variables. This indicates that there may be serious limitations to using profile information alone for advertising and retention targeting, as some practitioners have indicated (Green 2008).

[8]We can modify the model defined in Equations 1–4 to incorporate profile data in a hierarchical way by adding priors on γ_{uf}. We leave this extension for further research.

Table 2
EXPLAINING VARIATION IN THE POSTERIOR MEAN VALUES FOR γ_{uf}: THE PROBABILITY OF BEING INFLUENTIAL IN A DYAD

Covariate	*Coefficient*[a]	*t-Statistic*
Gender combination (female friend/male user)	.26	4.72
Months user has been a member	.36	18.02
Friend is of the same ethnicity as user	.25	6.34
User is looking for a date	–.66	–15.82
Friend is older than user	–.08	–2.04
R^2		11%

[a]Left-hand side: $\log[\hat{\gamma}_{uf}/(1-\hat{\gamma}_{uf})]$.

We also examined the distribution of the smoothing parameter ρ_u across users (Figure 7). The empirical distribution has a bimodal shape, with one mass point at approximately 2.2 (corresponding to 90% of weight assigned to lag 1) and the second mass point at approximately .25 (weights slowly decrease with lag).

Variation in ρ_u may be explained, in part, by the log-in pattern of a corresponding user. To investigate this, we regressed ρ_u on the mean and the variance of daily log-ins (calculated on a holdout sample). We find that for users with high (higher means) and stable (lower variances) log-in activities, more weight is assigned to the recent activity.[9] Conversely, less active or irregular users have weights more evenly distributed across lags. Behaviorally, this implies that it may take less time for "regular" users to learn about changes in friends' behavior and to form new expectations regarding content updates.

Level 2 network estimation results. In discussing Google's patent application for ranking influence, Green (2008) raises the notion of computing a user's "Google number"—the sum total of a person's influence on others. In a similar spirit, to estimate a person's influence in an online community, we need to evaluate his or her impact on egos within each ego-centered network of which he or she is a member. Therefore, we first estimate ego-centered networks of all Level 2 users. Then, for each user (u) of a Level 1 network, we calculate the network influence (I_u) as a sum of the marginal impacts he or she has on egos (e) in the Level 2 network (Equation 10):[10]

$$(10)\qquad I_u = \sum_{e=1}^{F_u} \int_{\gamma_{ue}} \int_{\beta_{ue}} \bar{y}_e \times \beta_{ue} \times \gamma_{ue} \times f(\gamma_{ue}, \beta_{ue})\, d\gamma_{ue} d\beta_{ue},$$

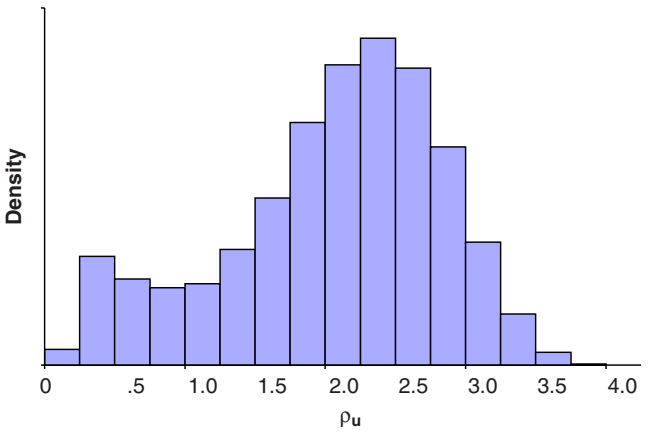
[9]The regression coefficient for the mean of daily log-ins is .05 (t = 5.05), and the coefficient for the variance of daily log-ins is .18 (t = 8.67).

[10]For the Poisson regression, the average response effect to a one-unit change in regressor can be calculated as follows:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial E[y_i|x_i]}{\partial x_i} = \hat{\beta}_j \times \frac{1}{n}\sum_{i=1}^{n} \exp(x_i'\hat{\beta}).$$

For the Poisson regression model with intercept included, this can be shown to simplify to $\bar{y} \times \hat{\beta}$ (Cameron and Trivedi 1998).

Figure 7
DISTRIBUTION OF SMOOTHING PARAMETER ρ_u



where

$\bar{y}_e$ = the average number of daily log-ins of user e,
F_u = the number of ego-centered networks user u enters as a friend (i.e., the number of friends of user u), and
$f(\gamma_{ue}, \beta_{ue})$ = the posterior distribution of γ_{ue} and β_{ue}.

The findings indicate that the majority of users have little impact on the behavior of others. However, some show a significant influence. As we expected, the extent of this influence also varies considerably across users. For example, some users with similar numbers of friends have total network impacts that differ by a factor of eight. These results reflect the ability of our method to identify influential users in the network and to quantify the likely extent of this influence.

ILLUSTRATING THE METHODOLOGY WITH SIMULATED DATA

We also used simulated data to examine the performance of the proposed Bayesian shrinkage approach. Specifically, we are interested in assessing how well the procedure recovers the identity of influential users under varying conditions. We designed the simulated data sets to be in the stochastic neighborhood of the data used in the field application.

The variable set for the simulation consists of two parts: the predictor variables and the response variable. The predictor variables correspond to the lagged activity levels of a user's friends, and the response variable corresponds to the activity level of a specific focal user. The predictor variables in the simulation take values corresponding to the empirically observed lagged activity levels of a user's set of friends. The response variable's values are then simulated from the Poisson model on the basis of the predictor variables and the assumed values for the parameters of the model. The model parameters are γ_{uf}, which represents whether friend f is influential on the focal user u, and β_u, which represents the influence strength of a friend who is influential. The term p_u represents the fraction of friends who are influential on user u. We vary these model parameters over multiple replications. We also vary T, the number of periods observed, and F_u, the number of friends (both influential and noninfluential) that user u has.

The simulation data sets are drawn according to a $3 \times 3 \times 3 \times 3$ design (81 simulation settings in total) produced by manipulating the four factors specified in Table 3. We chose the middle level of each factor to match the value obtained for the corresponding parameter from the field application.

For each cell, we randomly choose 20 users with several friends greater than or equal to the value of F_u in that cell. For each user, we randomly select F_u of that user's friends and take the corresponding activity data to be the predictor variables. The value of T, the number of periods observed, is 85 in the empirical data set. If T in a cell is 40, we randomly subselect 40 of the 85 periods. If T in a cell is 170, half the observations are an exact replication of the 85 periods in the empirical data, and the other half are from a random sampling with replacement.

For each of the 20 users in each of the 81 cells, we applied the estimation procedure to the corresponding simulated data. We ran the samplers for 10,000 iterations, dropping the first 8000 iterations as burn-in.

Table 3
SIMULATION DESIGN

	Level 1	*Level 2*	*Level 3*
Coefficient of friend's f influence on user u β_u	.12	.14	.16
Number of observations T	40	85	170
Number of friends F_u	45	90	160
Fraction of influential friends p_u	.05	.1	.2

Simulation Results

We computed and now discuss the following three performance measures:

1. How well the model differentiates between influential ($\gamma_{uf} = 1$) and noninfluential ($\gamma_{uf} = 0$) users,
2. How well the model recovers the share of influential friends (p_u), and
3. How well the model recovers the strength of an influential friend's influence β_u.

Measure 1: identifying the influentials. We assess how well the model predicts whether a certain friend is influential. The prediction is based on the posterior mean $\hat{\gamma}_{uf}$. Specifically, we predict that friend f is influential if $\hat{\gamma}_{uf}$ exceeds a threshold C. We consider two ways of setting C. The first is to set C to 1/2; we label this the $C_{.5}$ threshold. The other is to set C to minimize the total misclassification error in predicting influentials on a certain holdout sample for that cell; we label this the C_{opt} threshold. In actual practice, we would not be able to set the value of C in the latter way, because we would not observe the true value of γ_{uf}. We considered this threshold only to understand how much worse the simple $C_{.5}$ threshold would be relative to the best possible choice.

Across all 81 cells, the fraction of correctly classified friends varies between 68% and 100% for the $C_{.5}$ threshold and between 77% and 100% for the C_{opt} threshold. The average fraction is 90% for the $C_{.5}$ threshold and 92% for the C_{opt} threshold. Thus, the optimal threshold yields only slightly better performance than the simple threshold of 1/2, and therefore we drop C_{opt} from further discussion.

We also conducted an analysis of variance to explore the differences in the correct classification rate across the 81 cells. The main effects for the first two factors, β_u and T, were not statistically significant (at least for the cells considered in our design). For the other two factors (F_u and p_u), we found strong effects, each significant well beyond the .01 level. For the main effect F_u, number of friends, the mean for the correct classification rate was 95.4% for $F_u = 45$, 90.8% for $F_u = 90$, and 84.9% for $F_u = 180$. Thus, as the number of friends increases, it becomes more difficult to distinguish influential from noninfluential friends. For the fraction of influential friends, p_u, the mean correct classification rate was 96.9% for $p_u = .05$, 92.3% for $p_u = .1$, and 82% for $p_u = .2$. As the proportion of influential friends rises, it becomes somewhat more difficult to distinguish them from noninfluential ones. We chose the specific experiment design to match the field data's structure, and it is important to keep in mind that these patterns may not extend to different designs.

Measure 2: estimating the share of influential friends. We now report how well the methodology recovers the fraction of influential friends. As we discussed previously (see Equation 9), this fraction is estimated as $S_u = Infl_u/F_u$, where $Infl_u = \int_{\gamma_u} \sum_{f=1}^{F_u} \gamma_{uf} \times f(\gamma) d\gamma_u$ and the integration is over the posterior density of the γ_{uf} terms. Across all users in all 81 cells, the correlation between the true fraction and the estimated fraction is .88. The mean absolute value of the difference (MAD) between the actual and the estimated fraction is .05. An analysis of variance shows that the main effects for the four factors are not strongly significant.

Measure 3: estimating the influence strength of influential friends. The β_u parameter represents the influence strength of influential friends. Table 4 summarizes the recovery results for the β_u parameter at each level in the simulation design. We report the mean value of $\hat{\beta}_u$ for each of the three factor levels averaged over all the 27 cells with a particular (true) value for β_u. Table 4 also reports the absolute error over the 27 cells. Dividing the absolute error by the true value gives the relative error; this is 21% on average, a figure we consider reasonably low given the difficulty of the problem. The error is strongly affected by the number of observations, T. The relative error is approximately 29% for T = 40 days, 20% for T = 85 days, and 14% for T = 170 days. This suggests that accuracy in estimating a user's β_u goes up moderately quickly as additional days of activity are recorded and available for analysis.

Comparison with an Alternative Method

We used the simulation to compare the performance of our proposed Bayesian shrinkage approach with a simpler, non-Bayesian alternative. Specifically, we estimate the Poisson regression for each user–friend pair, controlling for other user-specific covariates; we label this the "univariate friend model." In this simpler approach, we use t-statistics to identify the subset of friends whose log-ins significantly explain the focal user's site activity. Using the simulated data, we estimate $3 \times 3 \times 3 \times 3 \times 20 \times F_u$ Poisson regressions by maximum likelihood and calculate t-statistics ($t_{\beta_{uf}}$) for each β_{uf}. We classify friends with $t_{\beta_{uf}} \geq 1.96$ as influential.

Across the 81 simulation designs, the univariate friend model correctly classified the influential friends between 46% and 92% of the time, with an average of 65%. This is substantially worse than the performance of the proposed Bayesian approach, for which the average correct recovery rate is 90% (using the 1/2 threshold). Thus, the simulation results support the use of the more elaborate Bayesian modeling procedure for the purposes of identifying influential users. Taken together, the simulation studies provide encouraging evidence for the performance of the proposed method for estimating influence levels for users in an online social network.

Table 4
RECOVERY OF β_u

	Level 1	*Level 2*	*Level 3*
Actual value of β_u	.12	.14	.16
Mean of $\hat{\beta}_u$ across the 27 corresponding cells	.11	.13	.14
Mean absolute error $\lvert\hat{\beta}_u - \beta_u\rvert$ across the 27 corresponding cells	.024	.029	.035

We also compared the performance of the proposed model with a non-Bayesian alternative, assuming that data simulation follows the latter.[11] Under such a scenario, the "true" model is a fixed-effects model in which each friend is associated with an individual β_{uf} coefficient. Note that, in practice, a full fixed-effects model can be estimated only for network members who have more log-in observations than friends. In the data set, only 60% of users satisfy this condition.

We use the following simulation procedure to generate data from the non-Bayesian alternative model. For all users in the Level 1 network, we estimate a univariate (in number of friends) Poisson regression model (controlling for self effects). We set all β_{uf} with $t_{\beta_{uf}} \geq 1.96$ equal to zero. Next, using these β_{uf} and assuming a full fixed-effects model (multivariate in number of friends Poisson regression), we generate the dependent variable for 85 observations in the actual sample. We simulated another 20 observations from the same empirical distribution to generate a holdout sample. The procedure ensures that the simulated data are in the stochastic neighborhood of the field application but generated according to a non-Bayesian alternative model. Finally, we estimate two models (a univariate non-Bayesian model whose parameters are indicated by superscript NB and the proposed Bayesian model indicated by superscript B) using the first 85 observations and compare in- and out-of-sample performance. The results show that in-sample recovery of β_{uf} is better with the proposed Bayesian model. Initially, this result was surprising because theory predicts that the data-generating model should perform no worse than alternative models, but the theoretical prediction is only an asymptotic result. In small samples, the Bayesian can do better because it is a shrinkage estimator and therefore has a variance advantage, which can offset the "bias" disadvantage that comes from the generating process being different from what is implicitly assumed by the Bayesian model.

The correlation between β_{uf} and $\hat{\beta}_{uf}^{NB}$ is .59, and for β_{uf} and $\hat{\beta}_{uf}^{B}$, it is .70. The MAD between β_{uf} and $\hat{\beta}_{uf}^{NB}$ is .066, and for β_{uf} and $\hat{\beta}_{uf}^{B}$, it is .014. Finally, the MAD calculated on the out-of-sample predictions of y is approximately 7% lower for the Bayesian approach (3.78 versus 4.08). In summary, the simulation shows that the proposed Bayesian model is robust under a data-generation process that assumes heterogeneity in friend effects within a user.

MANAGERIAL IMPLICATIONS FROM THE FIELD DATA ANALYSIS

In this section, we discuss implications for managers of applying the model in practice. First, we ask how well the univariate friend model and some simple descriptors of user activity, such as friend count and profile views, capture the influence levels of the site's most important users. Second, we evaluate scenarios to quantify the financial value of retaining the site's most influential users. Both analyses show that the proposed Bayesian approach offers significant potential benefits to managers concerned with targeting users for advertising and retention.

A practical consideration is whether the Bayesian approach is likely to offer meaningful gains in the identifi-

[11]We thank the associate editor for this suggestion.

cation of the most influential users compared with simple metrics. To address this, we evaluate how well other plausible measures of user importance do in capturing influence. Specifically, we examine the number of friends and the number of profile views ("hits") for each user. For a sample of 330, the correlation between a person's total marginal impact (the estimate of influence as revealed by the model and Equation 10) and his or her total number of friends is .72. The correlation between total marginal impact and profile views is .34. Although the number of friends and profile views both predict influence, considerable variance remains unexplained.

We also note that focusing on correlations alone could be potentially misleading in a practical sense. This stems from the likelihood that managerial interest is focused on members with the highest levels of influence (e.g., the top 5% or 10% of users). Because of their social influence and impact on the site activity of others, these top users are likely to be the most valuable to advertisers and most important for the site to retain as members (Green 2008). We compared how well the univariate friend model and the simple metrics of friend count and profile views predict influence at the top of the list. In Table 5, we present data for the top 10% (33 users) of the sample and show the cumulative impact obtained by proceeding further down the list. We place the values for the top 5% (corresponding to the cumulative impact at User 16) in boldface for discussion purposes. At this point, ranking by friend count yields a substantially lower network impact (43.49 versus 82.53).[12] The ratio is worse for profile views, at 24.37/82.53. Finally, the non-Bayesian alternative produces the best result of three naive rankings, which is still substantially lower than the proposed model (61.15 versus 82.53).[13] According to the foregoing analysis, simpler metrics, such as friend count and profile views, are likely to be inadequate proxies for user influence. As we noted previously, practitioners have been disappointed with the ability of simple metrics to capture influence on SN sites (Green 2008).

The relative payoffs from targeting based on estimated network influence are greatest at the top of the site's member list, but they continue to reach below the top 10%. For example, if we were to extend Table 5 to the entire sample of 330 users, we would find that the top 33% of influencers are responsible for 66% of the total impact. In practice, even narrowly focused targets (e.g., 5% of users) would involve addressing millions of user members at the major SN sites.

We now turn to implications of the model for managing retention efforts. In network settings, customer value to the firm is not solely a function of the cash flows generated by a customer but also a function of the effect of this customer on other customers (Gupta, Mela, and Vidal-Sanz 2006). The negative impact of an influential user leaving the site is not limited to the lost revenues from, for example, ad impressions not served to this particular person. 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 should be part of the valuation. In addition, if there is cost associated with retention efforts (e.g., an incentive for site usage stimulation, such as access to special features and monetary rewards), the firm needs to know whom to target.

A simple approach would be to choose people at random. This is actually similar to what MySpace.com did historically with its "Cool New People" feature—picking users at random and showing their profiles on the site's home page. This gives a popularity boost to the selected profiles. Alternatively, the site may focus on users who have many friends or users with a high number of profile views. Finally, the firm may consider targeting influential users identified by

Table 5
COMPARISON OF CUMULATIVE INFLUENCE CAPTURED

		Cumulative Network Impact of Top k People When Ranked by			
Top k	*Centile*	*Proposed Model*	*Univariate Friend Model*	*Number of Friends*	*Number of Profile Views*
1	.3%	10.44	4.41	4.41	2.60
2	.6%	17.89	8.24	6.63	4.88
3	.9%	24.53	11.57	9.35	6.67
4	1.2%	30.47	13.40	11.09	10.44
5	1.5%	35.87	17.17	15.68	10.47
6	1.8%	40.85	22.56	18.25	11.95
7	2.1%	45.57	25.80	19.72	14.27
8	2.4%	50.16	32.44	24.44	14.84
9	2.7%	54.58	39.89	26.98	15.91
10	3.0%	58.99	43.50	29.82	16.94
11	3.3%	63.16	45.68	32.60	18.48
12	3.6%	67.19	49.85	35.32	20.22
13	3.9%	71.12	52.70	37.01	20.65
14	4.2%	74.95	56.73	39.05	21.39
15	4.5%	78.77	59.42	40.90	23.12
16	**4.8%**	**82.53**	**61.15**	**43.49**	**24.37**[a]
17	5.2%	86.14	62.77	48.47	25.42
18	5.5%	89.47	65.50	49.93	25.56
19	5.8%	92.70	66.66	53.86	26.85
20	6.1%	95.92	68.67	56.46	37.29
21	6.4%	99.03	74.61	60.23	39.46
22	6.7%	102.04	76.09	64.26	41.09
23	7.0%	105.04	78.70	66.05	41.24
24	7.3%	107.96	81.33	70.22	43.27
25	7.6%	110.81	85.93	72.38	43.98
26	7.9%	113.64	88.53	73.97	44.18
27	8.2%	116.47	90.47	75.88	45.01
28	8.5%	119.28	92.51	78.71	46.87
29	8.8%	122.05	94.35	82.53	47.54
30	9.1%	124.78	96.94	85.75	50.14
31	9.4%	127.51	99.51	88.07	50.58
32	9.7%	130.23	101.50	90.80	50.77
33	10.0%	132.92	105.32	94.41	53.54

[a]Read as follows: Cumulative network impact of top 16 users identified by the proposed model is 1.35 times (82.53/61.15), 1.9 times (82.53/43.49), and 3.4 times (82.53/24.37) greater than the total impact of top 16 users identified by the non-Bayesian alternative model, the number of friends, and the number of profile views, respectively.

[12]Because both rankings use the same measure of individual marginal impact to calculate the cumulative top k impact, the first list is by construction greater than the naive ranking for each value of k. Therefore, it is a given that, in Table 5, the numbers in the "Proposed Model" column are greater than the numbers in the "Number of Friends" column.

[13]The procedure closely follows the one for the Bayesian method, except the selection criterion for influential users is based on the t-statistics calculated for each β_{uf} from pairwise Poisson regressions. We classify friends with $t_{\beta_{uf}} \geq 1.96$ as influential. Ranking is based on the total marginal impact, which we calculate as follows:

$$I_u = \sum_{e=1}^{F_u} \bar{y}_e \times \beta_{ue} \times I\left(t_{\beta_{ue}} \geq 1.96\right).$$

some empirical model; for this, we compare the proposed model and a non-Bayesian alternative.[14]

To gauge returns from targeted retention, we examine the potential effects on advertising revenue from impressions. While CPM (cost per mille, or cost per thousand impressions) on some premium sites can be as high as $15, most SN sites have a CPM under $1. Price quotes from several SN sites indicate that $.40 per thousand impressions is a reasonable benchmark. According to the data provided by the anonymous SN site, the average number of pages viewed on a community site by a unique visitor per month is approximately 130. From what we have observed across multiple SN sites, the average page carries approximately two to three advertisements. Thus, the average user contributes approximately $.13 per month or $1.50 a year of revenue from this source. From the data, we observe that users visit the site an average of 2.48 times per day, so each log-in generates approximately $.00175.

We can use the data in Table 5 to develop a numerical estimate of what would happen to the network if top users drop activity levels to zero (i.e., go from 2.48 to 0 log-ins).[15] In the sample of 330 users, the loss of the top 5% of users, ranked by influence, corresponds to a drop of 209.15 log-ins in the network (2.48×84.33). This translates into approximately $.3654 per day. If the drop in network activities persists, the total loss to the firm is approximately $133 a year ($.3654 \times 365$ days). Scaling this number to match the size of the target group of an actual network (approximately ten million users in the case of MySpace.com), the annual financial impact would be $78.5 million. In addition, each user alone (ignoring network effects) contributes approximately $1.50 a year of impression-based ad revenue, which, for ten million users, is $15 million. Adding these gives the payoff from retention actions in this scenario as approximately $93.5 million, given that the firm targets the top 5% of influential users as identified by our model.[16]

We repeat this analysis for the four other targeting approaches. In the case of random selection, the cumulative impact of losing 5% of users is a drop of 57.72 daily log-ins ($2.48 \times 1.369 \times 17$), which translates into approximately $.1008 per day or $36.81 a year ($.1008 \times 365$ days).[17] Scaling up to the size of the actual network and adding users' "stand-alone" value gives $36.65 million in payoff. Targeting on the basis of the number of friends and the number of profile views results in total payoffs of $57.78 million and $38.16 million, respectively. Finally, targeting on the basis of ranking results produced by the univariate friend model yields a total payoff of $72.7 million. All these are substantially below the $93.5 million associated with targeting from the proposed model. In summary, compared with simpler alternatives, application of the Bayesian approach might significantly improve the ability of SN sites to target their most influential members and to design and implement cost-effective retention programs.

[14]None of these approaches indicate how responsive the selected people are to the firm's retention efforts. We suggest that this can be determined through a series of small-scale field experiments with different target groups. We thank the associate editor for this suggestion. In addition, a more comprehensive targeting approach should take into account a person's propensity to leave the site with and without the program.

[15]Because the analysis does not reveal any significant correlation between a number of daily log-ins and strength of network influence, we use a sample average of 2.48 log-ins per day instead of individual-level daily log-in averages.

[16]We acknowledge that our approach is an oversimplification. To accurately infer the network impact caused by losing a customer, the model needs to take into consideration several other factors, which we do not discuss here. The main purpose of this example is to illustrate that different targeting approaches can lead to substantially different financial implications.

[17]The average predicted impact across users in our sample is 1.369; 17 people correspond to approximately 5% of users in the sample.

CONCLUSION

Firms operating SN sites observe an "overt" network of friends, defined according to who added whom as a friend. Most of the links in this network are "weak" in the sense that the relationships do not significantly influence behavior in the network; thus, identifying the "strong" links (i.e., the links corresponding to friends who affect a given user's behavior) is of interest. However, distinguishing weak links from the strong links is a difficult problem for two reasons. First, the number of overt links is large. Second, the firm wants to distinguish the links fairly quickly (e.g., in less than three months), so the number of "observations" available is fairly small. This sets up a challenging $P > N$ problem.

To address this, we develop and test a nonstandard Bayesian shrinkage approach in which the shrinkage is done across predictors within a model. Specifically, we implement a highly scalable Poisson regression model that shrinks influence estimates across friends within users. The primary intended contribution is a methodology for extracting, with limited data, the strong links from a large overt network that has mostly weak links. To the best of our knowledge, existing research, drawing from the literature on variable subset selection, has not yet done this in an application to massive right-hand-side expressions.

We tested the model on field data provided by an anonymous SN site. As expected, we found that relatively few so-called friends are actually significant influencers of a given user's behavior (22% is the sample mean), and substantial heterogeneity across users also exists. We also found that descriptors from user profiles (e.g., gender, stated dating objectives) lack the power to enable us to determine who, per se, is influential—the R-square is 11%. The spirit of this finding is corroborated by Google's recent efforts to better quantify social influence so that it might extract more revenue from targeted advertising on MySpace.com.

We also assessed the model's performance using simulated data. Specifically, we showed that the model performs well in correctly recovering the influential users and other key features of the data. Our Bayesian shrinkage approach also performs much better than a simpler, regression-based alternative model.

Our application to the field data provides a vivid illustration of the set of results that firms could obtain from applying the model in practice. We believe that these also have important implications for SN sites as businesses. In addition to the poor performance of profile descriptors in predicting influence, we showed that friend counts and profile views also fall short of being able to identify influential site members, especially for the most important 5%–10% of users. Examining a user retention scenario, we also illustrate the potential for large gaps in financial returns to the firm from using the model-based estimates of influence versus friend count, profile views, or, as MySpace.com has done, random selection.

Our approach could be readily applied to other data that might also be available to firms operating SN Web sites. For

example, we could augment the data set with richer information on the overt links (e.g., with details of the interactions). With richer information of this kind, the P > N situation is exacerbated, and the primary benefits of the methodology are enhanced. In this article, we chose to illustrate the methodology with log-in data, primarily because these are the data on which the partner SN firm operated. Apart from privacy and storage cost concerns, SN firms focus on own profile page visitation and log-in data because the own profile page is the center point of the typical user's activity and interactivity in the site. To make site usage easy, sites offer ways to minimize the need for a user to explicitly navigate to a friend's page to see the friend's activities. Major sites compile the main updates of a user's friends' profiles and present this compilation directly on the user's own profile page. Nonetheless, should managers or investigators want to extend the model to include additional usage information, the modeling approach we present herein—namely, Bayesian shrinkage across the friend effects within a user—should readily extend.

We also note that our model does not incorporate the potential dynamics in a user's influence over time. To address this, heterogeneity in β_u along the time dimension can be added. In further research, this might be done by using, for example, a hidden Markov approach.

Another potential limitation to this research is shared by most Internet-related studies. It is difficult, if not impossible, for a firm to know comprehensively when and how users interact with one another through other digital media. From interviews with several users, we learned that many maintain profiles on multiple SN sites, as well as digital communications, such as instant messaging, SMS (short message service), and BlackBerry e-mails. Other researchers, such as Park and Fader (2004), note the inherent limitations of models built on behavioral data collected on a single site when users' activities go beyond it. Padmanabhan, Zheng, and Kimbrough (2001) also warn of possibly erroneous conclusions from models that are limited to such data. The offline interactions of users are also potentially relevant and, in general, not available to researchers or managers. In this article, as do many others in Internet marketing, we focus on modeling online activity tracked by a single site and must leave the potential role of other Internet and offline activities for further research.

A final limitation is that we are unable to address how responsive the top "influencers" selected by our approach are likely to be to marketing actions (e.g., targeted advertising, a firm's retention efforts). This means that the managerial implications of the modeling approach are illustrative at this stage. Going forward, we believe that the value of identifying influential users can be established through relatively straightforward (and small-scale) field experiments. The contribution of this research is to address the methodological hurdles involved in suitably identifying such users so that the process of targeting and intervention can advance. (The Web Appendix at http://www.marketingpower.com/jmraug10 includes additional figures and tables omitted from the article for space consideration.)

APPENDIX

The Prior Distributions

1. $\alpha_{uk} \sim$ Normal(0, 100I).
2. $\beta_u \sim$ Normal(0, 100I), truncated at 0 on the left.
3. γ_{uf}
   - •Two point masses case:
     $\gamma_{uf} = 1$ with probability p_u, and $\gamma_{uf} = 0$, with probability $1 - p_u$.
   - •K point masses case:
     $\gamma_{uf} = k$ with probability p_{uk}, where k is a class index.
4. p_u
   - •Two point masses case:
     $p_u \sim \beta(a, b)$ with $a = b = 1$.
   - •K point masses case:
     $p_u \sim$ Dirichlet($a_1, a_2, \ldots, a_K$), where $a_1 = a_2 = \ldots = a_K = 1$.
5. $\rho_u \sim$ Uniform(c_1, c_2), where $c_1 = .01$ and $c_2 = 5$ (ρ_u at value c_1 results in weight being evenly distributed over all D lags, and c_2 is chosen to have more than 99% of weight assigned to the first lag).

The Gibbs Sampler

For each user (ego) u:

1. Generate $\alpha_u, \beta_u|\gamma_u, X_u, Z_{uf}$.

We use an independence Metropolis–Hastings sampler, where the Poisson likelihood function is approximated by a normal distribution with Metropolis correction. We generate parameter values $\alpha_u^{(n)}$ and $\beta_u^{(n)}$ using a normal distribution centered at the maximum likelihood estimate (MLE) for Equation 8 with the variance equal to the asymptotic variance (approximated by inverse of the Hessian H of the log-likelihood). The likelihood considered is conditional on the realizations of γ_u on that iteration. We reject candidates for $\beta_u \le 0$ (note that in K point masses case, β_u is a vector of K – 1 elements because for k = 1, β_{u1} is set to be 0). We accept the new values of $\alpha_u^{(n)}$ and $\beta_u^{(n)}$ with the following probability:

$$\Pr(\text{accept}) = \min\left\{ L_u\left(\alpha_u^{(n)}, \beta_u^{(n)} | \gamma_u^n\right) \times f_\alpha\left(\alpha_u^{(n)}\right) \times f_\beta\left(\beta_u^{(n)}\right) \times \exp\left[-\frac{1}{2}\left(\begin{bmatrix}\alpha_u^{(o)}\\ \beta_u^{(o)}\end{bmatrix} - \begin{bmatrix}\alpha_u^{(MLE)}\\ \beta_u^{(MLE)}\end{bmatrix}\right)' H\left(\begin{bmatrix}\alpha_u^{(o)}\\ \beta_u^{(o)}\end{bmatrix} - \begin{bmatrix}\alpha_u^{(MLE)}\\ \beta_u^{(MLE)}\end{bmatrix}\right)\right] \Bigg/ L_u\left(\alpha_u^{(o)}, \beta_u^{(o)} | \gamma_u^n\right) \times f_\alpha\left(\alpha_u^{(o)}\right) \times f_\beta\left(\beta_u^{(o)}\right) \times \exp\left[-\frac{1}{2}\left(\begin{bmatrix}\alpha_u^{(n)}\\ \beta_u^{(n)}\end{bmatrix} - \begin{bmatrix}\alpha_u^{(MLE)}\\ \beta_u^{(MLE)}\end{bmatrix}\right)' H\left(\begin{bmatrix}\alpha_u^{(n)}\\ \beta_u^{(n)}\end{bmatrix} - \begin{bmatrix}\alpha_u^{(MLE)}\\ \beta_u^{(MLE)}\end{bmatrix}\right)\right], 1\right\},$$

where L_u is an individual likelihood function and f_α and f_β are priors on α_u and β_u, respectively. Otherwise, we keep parameter values from the previous iteration $\alpha_u^{(o)}$ and $\beta_u^{(o)}$.

2. Generate $\gamma_{uf}|\alpha_u, \beta_u, X_u, Z_{uf}$ for all "friends" f of user (ego) u.

In the two point masses case, we draw γ_{uf} as a Bernoulli selecting a friend index f in a random order:

$$\Pr(\gamma_{uf} = 1) = \left\{ \frac{L_u\left(\alpha_u, \beta_u, \gamma_{uf}^{(1)}\right) \times p_u}{L_u\left(\alpha_u, \beta_u, \gamma_{uf}^{(1)}\right) \times p_u + L_u\left(\alpha_u, \beta_u, \gamma_{uf}^{(1)}\right) \times (1 - p_u)} \right\},$$

where $L_u(\alpha_u, \beta_u, \gamma_{uf}^{(1)})$ is an individual likelihood evaluated when $\gamma_{uf} = 1$ and $L_u(\alpha_u, \beta_u, \gamma_{uf}^{(0)})$ is an individual likelihood evaluated when $\gamma_{uf} = 0$.

In the K point masses case, we draw γ_{uf} as a categorical random variable selecting a friend index f in a random order:

$$\Pr(\gamma_{uf} = k) = \left\{ \frac{L_u\left(\alpha_u, \beta_u, \gamma_{uf}^{(k)}\right) \times p_{uk}}{\sum_{j=1}^{K} L_u\left(\alpha_u, \beta_u, \gamma_{uf}^{(j)}\right) \times p_{uj}} \right\},$$

where $L_u(\alpha_u, \beta_u, \gamma_{uf}^{(k)})$ is an individual likelihood evaluated when $\gamma_{uf} = k$ (i.e., friend f is assigned to class k).

3. Generate $p_u|\gamma_{uf \text{ for } \forall f}$.

In the two point masses case, we draw $p_u \sim \beta(a + IF_u, b + F_u - IF_u)$, where $IF_u = \Sigma_{\forall f}\, \gamma_{uf}$ and F_u is the number of "friends" of user u.

In the K point masses case, we draw $p_u \sim \text{Dirichlet}(a_1 + F_{u1}, a_2 + F_{u2}, \ldots, a_K + F_{uK})$, where $F_{uk} = \Sigma_{\forall f}\, 1(\gamma_{uf} = k)$.

4. Generate $\rho_u|\alpha_u, \beta_u, \gamma_{uf}, X_u, Z_{uf}$.

We draw a smoothing parameter ρ_u using a random walk Metropolis–Hastings algorithm. A candidate $\rho_u^{(n)}$ is formed as $\rho_u^{(n)} = \rho_u^{(o)} + \zeta$, where $\zeta \sim \text{Normal}(0, \sigma_\rho)$ and σ_ρ is being adjusted dynamically during burn-in iterations to ensure acceptance rate in the 25%–45% range. It is fixed thereafter. We accept the new value of $\rho_u^{(n)}$ with the following probability:

$$\Pr(\text{accept}) = \min\left\{ \frac{L_u\left(\alpha_u, \beta_u, X_u, Z_{uf}^{(n)}\right) \times f_\rho\left(\rho_u^n\right)}{L_u\left(\alpha_u, \beta_u, X_u, Z_{uf}^{(o)}\right) \times f_\rho\left(\rho_u^o\right)}, 1 \right\},$$

where

$$z_{uft}^{(n)} = \sum_{d=1}^{D} w_{ud}^{(n)} \times y_{f(t-d)},$$

$$w_u^{(n)}(d) = \frac{\exp\left(-d \times \rho_u^{(n)}\right)}{\sum_{k=1}^{D} \exp\left(-k \times \rho_u^{(n)}\right)}, \text{ and}$$

$$f_p = \text{prior on } \rho_u.$$

If a candidate value $\rho_u^{(n)}$ is accepted, we use new weights w_u to calculate Z_{uf}. Otherwise, we keep parameter values from the previous iteration $\rho_u^{(o)}$ and do not update Z_{uf}.

REFERENCES

Cameron, A. Colin and Pravin K. Trivedi (1998), *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Chevalier, Judith A. and Dina Mayzlin (2006), "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research*, 43 (August), 345–54.

Dellarocas, Chrysanthos N. (2005), "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," working paper, Robert H. Smith School of Business, University of Maryland.

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 (3), 241–63.

Godes, David and Dina Mayzlin (2004), "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science*, 23 (4), 545–60.

Green, Heather (2008), "Google: Harnessing the Power of Cliques," *BusinessWeek*, (October 6), 50.

Guadagni, Peter M. and John D.C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2 (3), 203–238.

Gupta, Sunil, Carl F. Mela, and Jose M. Vidal-Sanz (2006), "The Value of a 'Free' Customer," Working Paper No. 07-035, Harvard Business School, Harvard University.

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

Iacobucci, Dawn (1990), "Derivation of Subgroups from Dyadic Interactions," *Psychological Bulletin*, 107 (1), 114–32.

———, ed. (1996), *Networks in Marketing*. Newbury Park, CA: Sage Publications.

——— (1998), "Interactive Marketing and the Meganet: Network of Networks," *Journal of Interactive Marketing*, 12 (Winter), 5–16.

——— and Nigel Hopkins (1992), "Modeling Dyadic Interactions and Networks in Marketing," *Journal of Marketing Research*, 29 (February), 5–17.

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

Leenders, Roger Th.A.J. (2002), "Modeling Social Influence Through Network Autocorrelation: Constructing the Weight Matrix," *Social Networks*, 24 (1), 21–48.

Marketing Science Institute (2006), "2006–2008 Research Priorities," (accessed April 28, 2010), [available at http://www.msi.org/pdf/MSI_RP06-08.pdf].

Narayan, Vishal and Sha Yang (2006), "Trust Between Consumers in Online Communities: Modeling the Formation of Dyadic Relationships," working paper, Stern School of Business, New York University.

Padmanabhan, Balaji, Zhiqiang Zheng, and Steven O. Kimbrough (2001), "Personalization from Incomplete Data: What You Don't Know Can Hurt," in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, 154–64.

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

Phelps, Joseph E., Regina Lewis, Lynne Mobilio, David Perry, and Niranjan Raman (2004), "Viral Marketing or Electronic Word-of-Mouth Advertising: Examining Consumer Responses to Pass Along Email," *Journal of Advertising Research*, 44 (4), 333–48.

Robins, Garry, Philippa Pattison, and Peter Elliott (2001), "Network Models for Social Influence Processes," *Psychometrika*, 66 (2), 161–90.

Scott, John (1992), *Social Network Analysis*. Newbury Park, CA: Sage Publications.

Spiegelhalter, D.J., N.G. Best, B.P. Carlin, and A. Van der Linde (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society: Series B*, 64 (4), 583–39.

Stephen, Andrew T. and Olivier Toubia (2010), "Deriving Value from Social Commerce Networks," *Journal of Marketing Research*, 47 (April), 215–28.

Van den Bulte, Christophe and Stefan Wuyts (2007), *Social Networks and Marketing*. Cambridge, MA: Marketing Science Institute.

Walsh, Mark (2008), "Goldstein at IAB: Marketers Can't Ignore Social Media," *Online Media Daily*, (June 3), (accessed September 30, 2008), [available at http://www.mediapost.com/publications/index.cfm?fuseaction=Articles.san&s=83859].