DOCTORAL DISSERTATION

Essays on Social Media Platforms

Submitted to the
David A. Tepper School of Business
in partial fulfillment for the requirements for the degree of
DOCTOR OF PHILOSOPHY in Industrial Administration

By Yingda Lu

Carnegie Mellon University
Tepper School of Business
Pittsburgh, Pennsylvania 15213

DISSERTATION COMMITTEE:
Param Vir Singh (co-chair)
Baohong Sun (co-chair)
Kinshuk Jerath
Tridas Mukhopadhyay
Sunder Kekre

October 2012

Contents

Chapter 1

Introduction

My dissertation focuses on social media platforms. With the advent of Web 2.0 technologies, various types of social media platforms have prospered in the last few years. However, firms face the challenges of how to optimize the design and management of social media platforms. This requires researchers and managers to understand how individuals contribute and consume content in social media platforms, how individuals interact with each other, and what are the desirable policies that could maximize the value of social media initiatives of an organization. My research investigates these critical research questions in various contexts.

In my first essay, "*The Emergence of Opinion Leaders in a Networked Online Community: A Dyadic Model with Time Dynamics and a Heuristic for Fast Estimation*", I study the drivers of the emergence of opinion leaders in a networked community where users follow each other and share information with peers. I model the formation of opinion leadership by using a dyad-level proportional hazard model with time-varying covariates. To estimate this model, I use Weighted Exogenous Sampling with Bayesian Inference (WESBI), a new methodology that I develop for fast estimation of dyadic models on large network datasets. I find that, in this online review network, both the widely-studied "preferential attachment" effect based on the existing number of inlinks (i.e., a *network-based* property of a node) and the number and quality of reviews written (i.e., an *intrinsic* property of a node) are significant drivers of new incoming trust links to a reviewer (i.e., inlinks to a node). Interestingly, time is an important moderator of these effects—the number of recent reviews written has a stronger effect than the effect of the number of recent inlinks received on the current rate of attracting inlinks; however, the aggregate number of reviews written in the past has no effect, while the aggregate number of inlinks obtained in the past has a significant effect on the current rate of attracting inlinks. This leads to the novel and important implication that, in a network growth setting, intrinsic node characteristics are a stronger short-term driver of additional inlinks, while the preferential attachment effect has a smaller impact but it persists for a longer time. I discuss the managerial implications of the results for the design and organization of online review communities.

In the second essay, "*Learning from Peers on Social Media Platform*", I investigate the knowledge sharing on a social media platform. Nowadays, more and more companies have adopted social media platforms for supporting knowledge sharing among customers and employees, where individuals ask and answer questions among each other. Hence, it is important to understand the

knowledge-sharing behavior of users on these systems. I propose a theoretically-grounded, dynamic structural model with endogenized knowledge-sharing behavior that takes into account "learning by sharing" and "knowledge spillover," which are two salient features that are enabled by social platforms. This model recognizes the dynamic and interdependent nature of knowledge-seeking and sharing decisions and allows them to be driven by knowledge increments and social-status building in anticipation of future reciprocal rewards. Applying this model to a unique panel of data from an expertise-sharing forum used to shore up customer support at a Fortune 500 firm, I illustrate the dynamic interdependency between individual decisions. I show that an individual is more willing to contribute to the community when her peers are more knowledgeable. I further demonstrate how a "core/periphery" knowledge sharing structure emerges, discourages users with low social status from participating, and creates a barrier to knowledge sharing and integration for the company. An exploratory sensitivity analysis shows that hiding the identity of the knowledge seeker breaks the core/periphery structure and improves the knowledge sharing by 20.46%.

Chapter 2

The Emergence of Opinion Leaders in a Networked Online Community: A Dyadic Model with Time Dynamics and a Heuristic for Fast Estimation

1. Introduction

Opinion leaders—individuals who exert a considerable amount of influence on the opinions of others—are an important element in the diffusion of information in a community (Gladwell 2000, Rogers 2003). Motivated by the seminal work by Katz and Lazarsfeld (1955), researchers have contributed to our understanding of opinion leaders by systematically analyzing how individuals emerge as opinion leaders in a community (Watts and Dodds 2007), how they facilitate the diffusion of information by their influence on the opinions of others (Ghose and Ipeirotis 2011, Iyengar et al. 2011, Van den Bulte and Joshi 2007, Stephen et al. 2012), what the characteristics of these individuals are (Chan and Misra 1990, Myers and Robertson 1972), and how to identify them, often with the aim of marketing products through them (Valente et al. 2003, Vernette 2004).

With the advent of Web 2.0, websites where consumers voluntarily contribute product reviews, such as Epinions (www.epinions.com), have prospered in the last few years. By sharing their own opinions on these online forums, consumers influence others' opinions as well. An advantage of such activity being online is that it may be possible to track the flow of influence among the members of the community. For instance, Epinions employs a novel mechanism in which every member of this community can formally include members whose reviews she trusts in her "web of trust." This leads to the formation of a network of trust among reviewers with high in-degree individuals being the opinion leaders. Various other websites which provide forums for user-generated content provide mechanisms of the above nature under which users can extend links to other users whose opinions or content they value (among other reasons for forming such links), thus leading to a networked community. Examples of such websites include The Motley Fool (www.fool.com) and Seeking Alpha (www.seekingalpha.com) for sharing opinions on topics related to financial markets, YouTube (www.youtube.com) for sharing videos, IMDb (www.imdb.com) and Rotten Tomatoes (www.rottentomotoes.com) for sharing opinions on movies, yelp (www.yelp.com) for sharing information on local food and entertainment and, last but not the least, social networks such as Facebook (www.facebook.com).

Among thousands of heterogeneous online reviewers in such communities, which ones emerge as opinion leaders? How do their intrinsic characteristics versus their network-level characteristics influence their statuses as opinion leaders? What are the major factors that influence individuals' consideration of other reviewers' opinions *over time*? In the context of a networked community with links in the network denoting opinion seeking,[1] these essentially become questions regarding the factors influencing the evolution of the network. Therefore, we embed influence through opinion sharing in a network growth paradigm and, using a unique dataset from Epinions, we investigate the emergence and dynamics of opinion leadership in a community.[2]

Several researchers have illustrated that network structure-based factors such as a node's degree, reciprocity and transitivity, have a significant impact on the formation of ties (Barabasi and Albert 1999, Holland and Lienhardt 1972, Jones and Handcock 2003, Merton 1968, Narayan and Yang 2007). A prominent theory is the preferential attachment theory, which suggests that nodes with more existing incoming links, as compared to nodes with fewer existing incoming links, have a higher probability of receiving additional incoming links. However, the effect on network formation of the intrinsic characteristics of the nodes themselves is under-studied (with a few notable exceptions, e.g., Kossinets and Watts 2006, Stephen and Toubia 2009). In our context, characteristics of reviews written serve as natural node characteristics (for instance, is a review written recently, and is it written comprehensively and objectively). A main objective of our paper is to understand how these intrinsic node characteristics influence network evolution.

One of the key features of online review communities is that the network structure and individual behavior are dynamically changing over time. For example, over time, reviewers may receive new incoming trust links and also contribute new reviews, both of which increase their attractiveness to other members of the community. Compared with offline social networks, the cost of changing structural and behavioral characteristics is smaller in online settings, and therefore the dynamic properties may become very salient. As a result, how the time-changing characteristics of

[1] This method of employing the number of incoming links as a proxy of measuring opinion leadership is called the sociometric method and has been used widely before in sociology and marketing (Burt 1999, Iyengar et al. 2010, King and Summers 1970). This method fits our context well, because a larger number of incoming links can lead to overall higher influence. Reviewers with a larger number of incoming trust links are easier to find due to their network position. In addition, they are trusted by more members in the community and this also inspires confidence in the new readers, which makes it more likely that they will influence people who find them. In totality, we can conclude that reviewers who have larger number of incoming links are the ones with higher opinion leadership.

[2] A large literature exists on diffusion of information over an existing network or in a community. Note, however, that our work differs from the above because our focus is on the formation of the underlying network itself.

individuals influence the formation of ties is a question of great importance in understanding how online review communities develop, especially given the recent explosion in user-generated content.

To answer these research questions, we develop a dyad-level proportional hazard model of network growth and estimate it on the network of movie reviewers (in the "Movies" category) at Epinions. We find that while network structure-based factors such as preferential attachment and reciprocity are significant drivers of network growth, intrinsic node characteristics such as the number of reviews written and textual characteristics such as objectivity, readability and comprehensiveness of reviews are also significant drivers of network growth. Interestingly, the recent number of reviews written by a reviewer has a strong impact on the rate of increase of opinion leadership status for the individual, while the past number of reviews written has no statistically significant impact. In contrast, if we also divide the trust-based inlinks for a reviewer into recently-obtained inlinks and past inlinks, we find that both have a statistically significant impact on the rate of increase of opinion leadership status.

Taken together, we find that time is an important moderator of the impact of node-based and network structure-based characteristics on the tie formation process—node-based characteristics are significant short-term drivers of additional inlinks, while the network structure-based preferential attachment effect is a longer-term but less effective driver of additional inlinks. This novel finding provides a deeper understanding of how opinion leaders emerge in online communities, and contributes to the theory of generative models of large networks. This also has important managerial implications for the design of opinion-sharing websites, which we discuss later.

To add to the above substantive findings, we also contribute to the methodology of handling large-scale social network datasets. Review and reviewer characteristics change over the time period of our study, and time-varying covariates need to be taken into account when modeling the growth of the social network. To deal with the overwhelming computational requirements of a dyad-level proportional hazard model with time-varying covariates, we develop a novel Markov Chain Monte Carlo adaptation of the Weighted Exogenous Sampling methodology (Manski and Lerman 1977). Our *Weighted Exogenous Sampling with Bayesian Inference (WESBI)* methodology reduces the time of estimation by an order of magnitude, while still providing valid estimates. Thus, our methodological contribution is the development of a fast hierarchical Bayes inference technique for estimating dyad-level network growth models with time-varying covariates. We also extend the weighted exogenous sampling methodology from binary models to duration models. In a technical appendix to this paper, we report results of a comprehensive simulation study covering a large variety of possible network structures characterized by different parameter values. For each network structure, we show that by sampling a small proportion of the total observations, we can recover the true network generating parameters with high accuracy using WESBI.

The rest of the paper is organized as follows. In Section 2, we develop the theoretical foundations motivating our empirical work. In Section 3, we develop a proportional hazard model with time-varying covariates to estimate the effect of network and reviewer characteristics on social network evolution. In Section 4, we describe the Epinions dataset we constructed and our variable definitions. In Section 5, we develop and explain our novel estimation methodology and present the estimation results of the model on data from the "Movies" category at Epinions. In Section 6, we provide several extensions and robustness checks for our basic model. In Section 7, we conclude by discussing the implications of our study and potential future research.

2. Theoretical Foundations

In this section, we provide theoretical justifications for the various concepts and constructs that we incorporate in our network-based model of opinion leadership.

In the past decade, sociologists, physicists and computer scientists have empirically studied networks in such diverse areas as social networks, citation networks of academic publications, the World Wide Web network, email networks, router networks, etc. A property frequently identified in networks across these domains is the "scale free" property (Dorogovtsev and Mendes 2003). A network is said to be "scale free" if its degree distribution follows a power law at least asymptotically (Barabasi and Albert 1999). Interestingly, we find that the "web of trust" network at Epinions is also a scale free network. The most widely accepted network growth phenomenon that produces a scale free network is the preferential attachment (or "rich get richer") process (Barabasi and Albert 1999). In the context of Epinions, the preferential attachment argument would imply that individuals who already have a high number of inlinks would be proportionately more likely to receive new inlinks. An explanation for why the preferential attachment effect is observed is that individuals who possess social capital can leverage it to receive more social captial (Allison et al. 1982, Merton 1968). In a community of reviewers, high-status reviewers (ones with high in-degree) would be considered more attractive for seeking opinion from (Bonacich 1987, Gould 2002). This implies that people would like to select high-status individuals and this process will be self-reinforcing. Furthermore, by design, Epinions prominently displays the reviews of reviewers with highest in-degrees (i.e., reviews of the reviewers with the most number of followers). This provides higher visibility to such reviewers and hence a greater chance of getting new links (Tucker and Zhang 2010). Motivated by the above arguments, we incorporate the preferential attachment process in our model by assuming that the probability that individual A forms a link with individual B increases with B's in-degree.

In social psychology, another network phenomenon called dyad-level reciprocity has been considered as one of the key drivers of link formation in networks (Fehr and Gachter 2000, Iacobucci and Hopkins 1992). Reciprocity refers to responding to a positive action of another individual by a positive action towards that individual (Katz and Powell 1955). In the context of Epinions, we incorporate reciprocity in our model by assuming that individual A is more likely to put individual B in her web of trust if B has already put A in her web of trust.

While preferential attachment and reciprocity are network-based effects (node-level and dyad-level effects) and have been considered to be important drivers of network evolution, they fail to explain many network dynamics that one observes. For instance, an underlying problem with the preferential attachment framework is that it does not explain why a person could be replaced by another as an opinion leader over time. If the preferential attachment were the only mechanism, we would expect that a person with a large number of incoming links will receive a proportionally larger fraction of new incoming links. In other words, an opinion leader will continue as an opinion leader forever without exerting substantial effort (even though new opinion leaders may emerge). A simple examination of the Epinions data illustrates that this is not the case—specifically, after a popular reviewer becomes inactive for a while, the number of additional incoming links that she obtains in every period decreases dramatically.

We argue that a node's "content" (non-network characteristics) can help us explain such dynamics. For instance, if an opinion leader becomes inactive and stops writing reviews, others will prefer to seek the opinion of a reviewer who is active and provides fresh information. In other words, time is likely to be an important moderator of the impact on opinion leadership of node characteristics such as the number of reviews contributed by an individual. While the total number of reviews should have an impact because more reviews provide more information, recently written reviews are likely to have higher impact because they are more likely to provide new information. For instance, new reviews are likely to be about new items for which few reviews exist, or may provide newer insights on old items. (Stephen, Dover and Goldenberg (2010) suggest similar reasoning in an online diffusion context.) To understand this, we divide the reviews written by every reviewer into "recent reviews" (written in the last time period, which is one month) and "past reviews" (older than one month) and assess their impact separately. To simultaneously understand whether time also moderates the impact of preferential attachment, we divide the trust links obtained by a reviewer into those obtained recently (within the last one month) and those obtained in the past (older than one month). We can expect recent reviews to significantly influence the current rate of incoming links, and past reviews to not. We can also expect preferential attachment to have a significant influence. However, this is still an empirical question (especially the magnitudes of these effects) which we answer using our formal model.

A related stream of literature has established that the attributes of a review such as its readability and comprehensiveness may affect a reader's response to the product and the perception of the reviewer (Ghose and Iperoitis 2011, Kim and Hovy 2006, Liu et al. 2007, Otterbacher 2009, Zhang and Varadarajan 2006). Reviewers may also express their subjective opinions or objective facts, and a mix of both may be most preferred. In other words, textual characteristics of reviews can influence opinion leadership, and we test this formally as well.

Finally, relationships between individuals offline are often characterized by homophily, which refers to a tendency for people who belong to the same demographic or social category, such as age or gender, to be connected to each other (McPherson et al. 2001). There is some uncertainty about the extent to which sharing a demographic or social category produces homophily in an online context (Van Alstyne and Brynjolfsson 2005)—it appears that while similarity in demographic categories does not lead to tie formation in an online context, similarity in certain latent constructs (as measured by expressed characteristics in reviews) leads to tie formation. In the context of Epinions, the expressed characteristics to measure homophily could be the review writing styles. We expect that those pairs of individuals who have similar review writing styles would be more likely to form ties with each other, and we incorporate this into our model.

3. Model Development

We develop a stochastic network growth model conceptualized at the dyad level with directional ties.[3] Since networks evolve over time, network tie formation data is typically right censored. Hence, instead of modeling tie formation as a discrete-choice process, we model it as a timing process by using a proportional hazard model (Greene 2003), i.e., there is a baseline hazard rate for tie formation, moderated by dyad- and direction-specific quantities. We describe this below.

Consider the formation of a directed tie from individual *i* to individual *j*. We use the time period for which both individuals *i* and *j* have been present in the community as the starting point of the timing process for this potential tie, and denote the time from the start to the current time as *t*. The hazard rate for tie formation from i to $\boldsymbol{j}$ is denoted as:

$$\lambda_{ij}(t) = \lambda_0(t)\exp\{V_{ijt}\}.$$

In the above, $\lambda_0(t)$ is the baseline hazard rate at time t, which describes the inherent propensity of two individuals to form a link without considering other factors. We assume that $\lambda_0(t)$ follows a Weibull distribution to allow for a flexible baseline hazard rate:

$$\lambda_0(t) = \alpha_0\alpha_1 t^{\alpha_1 - 1}, where\ \alpha_0, \alpha_1 > 0.$$

The quantity $\exp\{V_{ijt}\}$ increases or decreases the baseline hazard rate for the formation of a directed tie from *i* to *j* at time *t*, based on the values of time-varying dyad- and direction-specific covariates. We interpret V_{ijt} as the "adjustment factor" for the latent propensity of a node *i* to extend a link to node *j* at time *t*, conditional on this not having happened yet. This conditional probability of *i* linking to *j* increases with V_{ijt}, and it incorporates the various covariates that are expected to influence link formation (based on the theory discussed in the previous section). We let $\boldsymbol{z}_{ijt}$ be the set of sender-, receiver- and dyad-specific covariates for the dyad *ij* at time *t*. Then, the above can be written as $V_{ijt} = \boldsymbol{z}_{ijt}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of coefficients for $\boldsymbol{z}_{ijt}$. We discuss in detail the different covariates included in $\boldsymbol{z}_{ijt}$ in Section 4. As an example at this point, note that we can incorporate the preferential attachment process by including the covariate Degree_{jt}, which is the in-degree of node *j* at time *t*. (In other words, if the coefficient for Degree_{jt} is larger, then the probability of *i* extending a tie to *j* at time *t* is higher.)

[3] Some other papers that develop stochastic models for network phenomena include Ansari et al. (2011), Braun and Bonfrer (2011), Handcock et al. (2007), Hoff et al. (2002), Robins et al. (2007) and Snijders et al. (2006).

While the above incorporates observed characteristics, we also need to control for unobserved characteristics in a dyad. For example, the sender nodes could be inherently more active (or passive) and the receiver nodes could be inherently more attractive (or unattractive). To account for this, we incorporate node-specific unobserved effects as $V_{ijt} = \boldsymbol{z}_{ijt}\boldsymbol{\beta} + a_i + b_j$, where a_i is the sender-specific unobserved random effect (that accounts for the "activity rate" of node *i*) and b_j is the receiver-specific unobserved random effect (that accounts for the "attractiveness" of node *j*). The sender- and receiver-specific effects of the same individual are allowed to be correlated with each other as:

$$\binom{a_i}{b_i} \sim MVN\left(0, \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}\right)$$

Furthermore, the extant sociology literature considers homophily as a key driver of link formation in a social network (McPherson et al. 2001), which implies that links are formed between similar individuals. We explicitly incorporate both observed and unobserved homophily in our model. The observed similarity in behavior is captured using dyad-specific variables in $\boldsymbol{z}_{ijt}$, and the unobserved dyad-specific homophily is captured by using a dyad-specific unobserved random effect, d_{ij}, as $\mathrm{V}_{\mathrm{ijt}} = z_{ijt}\beta + a_i + b_j + d_{ij}$, where $d_{ij} \sim MVN(0, \sigma_d^2)$. [4] Furthermore, we assume that the dyad-specific unobserved effects are symmetric, i.e., $\mathrm{d}_{\mathrm{ij}} = d_{ji}$.

We can present V_{ijt} above in a simplified manner by aggregating the random effects together with the corresponding covariates as:

$$V_{ijt} = \left(\boldsymbol{x}_{it}\boldsymbol{\beta}^i + a_i\right) + \left(\boldsymbol{x}_{jt}\boldsymbol{\beta}^j + b_j\right) + \left(\boldsymbol{x}_{ijt}\boldsymbol{\beta}^{ij} + d_{ij}\right),$$

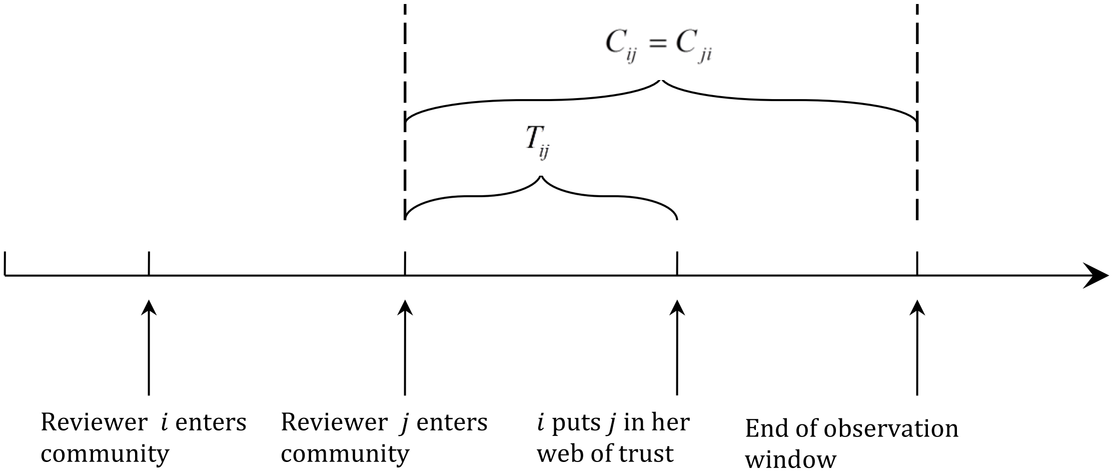
where $\boldsymbol{z}_{ijt} = [\boldsymbol{x}_{it}\ \ \boldsymbol{x}_{jt}\ \ \boldsymbol{x}_{ijt}\ ]$ and $\boldsymbol{\beta}^i$ contains coefficients for sender-specific covariates, $\boldsymbol{\beta}^j$ for receiver-specific covariates and $\boldsymbol{\beta}^{ij}$ for dyad-specific covariates. Therefore, $\boldsymbol{x}_{it}\boldsymbol{\beta}^i + a_i$ is the sender effect, $\boldsymbol{x}_{jt}\boldsymbol{\beta}^j + b_j$ is the receiver effect and $\boldsymbol{x}_{ijt}\boldsymbol{\beta}^{ij} + d_{ij}$ is the dyad effect.

We now derive the conditional likelihood function for the above model. We fix the unit of time in our model as one month. Our data is right censored because we do not observe whether ties are formed or not after the end of our observation time window. Let C_{ij} be the number of time

[4] A richer approach for capturing unobserved homophily is to cluster individuals in multi-dimensional space representing latent characteristics. See Braun and Bonfrer (2011) for an excellent application.

periods for which dyad ij has been observed, and T_{ij} be the length of time from the starting point to the time period when i extends a tie to j. (Note that C_{ij} and C_{ji} are always equal, but T_{ij} is, in general, different from T_{ji}.) We define $\mathbb{I}_{ij} = 1$ if $T_{ij} \leq C_{ij}$ (i.e., if a tie formed within the observation time) and 0 otherwise, and $k_{ij} = \text{floor}\left(\min\{T_{ij}, C_{ij}\}\right)$. We present this graphically in Figure 1.

Figure 1: Illustration of Link Formation Time and Censoring Time Used in the Model (In this figure, $\mathbb{I}_{ij} = 1$ and $k_{ij} = T_{ij}$, and $\mathbb{I}_{ij} = 0$ and $k_{ji} = C_{ji}$.)


$C_{ij} = C_{ji}$
T_{ij}
Reviewer i enters community
Reviewer j enters community
i puts j in her web of trust
End of observation window

Using the notation above, the log-conditional-likelihood function (i.e., conditional on knowing a specific directed dyad's latent parameters) can be written as:

$$\log L = \sum_{i,j \neq i} \left\{ \mathbb{I}_{ij} \cdot \log\left[1 - \exp\left\{- \exp\left[\alpha\left(k_{ij}\right) + \boldsymbol{z}_{ij,k_{ij}}\boldsymbol{\beta} + a_i + b_j + d_{ij}\right]\right\}\right] \right.$$

$$\left. - \sum_{t=0}^{k_{ij}-1} \exp[\alpha(t) + \boldsymbol{z}_{ijt}\boldsymbol{\beta} + a_i + b_j + d_{ij}] \right\} \qquad (1)$$

where $\alpha(t) = \log\{\int_t^{t+1} \lambda_0(u)du\}$. (See Appendix I for the detailed procedure of deriving this expression.)

Before we proceed further, we make a few notes. First, the above model does not account for unobserved heterogeneity either in the baseline hazard rate, or in the coefficients for the covariates. We use this simple (yet still quite rich) model for our basic analysis, and then extend it in Technical Appendix II to include both kinds of heterogeneity above. Second, a key distinction of our model from most other stochastic models of network growth (including the Barabasi and Albert (1999) framework) is that those models typically do not predict which and when two nodes will form a link whereas we model this explicitly. Finally, while our model above shares some commonalities with Hoff (2005) and Narayan and Yang (2007), we extend their models in many ways—most importantly, we incorporate time-varying covariates which they do not.

4. Data

Data Description

Epinions allows reviewers to post reviews, and allows them to put other reviewers whom they trust in their "web of trust." Reviews are organized by product categories, such as Movies, Cars, Books, Music, Electronics, Home & Garden, etc. Reviews in different product categories may have different properties, and communities focusing on different products may have different preferences. For example, reviews that focus primarily on objective details of products may be preferred for electronics but not to the same extent for movies. To avoid mixing the different preferences of people reading and writing reviews in different product categories, we focus on the "Movies" review community. We further restrict our focus on registered members who have at least written one review on any movie, to ensure that the individuals in our dataset indeed have an expressed interest in movies. We relax these constraints on data collection later in Section 6.

To crawl our data on the network of movie reviewers, we first constructed a comprehensive list of feature films released between 1888 and 2008 as listed on http://www.imdb.com/year, and took the intersection of this list with movies reviewed on Epinions. This process gave us 19,851 movie titles. Next, we searched for all reviews written for any of these movies on Epinions, and constructed the list of reviewers who have written these reviews. From this list, we selected reviewers who registered at Epinions between January 2002 and December 2008.[5] For each of these reviewers, we collected data on which others they added in their web of trust and at what time, and

[5] We consider individuals who started their activity only after 2002 because the information about the dates when web of trust ties between individuals were formed is not available for ties before January 2002, which leads to a left-censoring problem.

constructed the full network of trust among these reviewers. In addition, for each reviewer, we crawled the full text of each review she wrote and the date when it was written.

The resulting dataset contained 6,705 reviewers with 2,315 ties among them (out of 44,950,320 dyads) and a total of 27,634 reviews written. We further divided this dataset into a calibration sample and a holdout sample. The calibration sample contained reviewers who entered the movie community between January 2002 and December 2005 (5,180 reviewers who formed 1,906 ties with each other, and wrote 21,049 reviews). The holdout sample, employed to evaluate the model's predictive performance, consisted of reviewers who entered the movie community between January 2006 and December 2008 (1,525 individuals who formed 160 ties with each other, and wrote 6,585 reviews).

Variable Description

As stated before, the variables that we employ can be divided into three categories: receiver-specific covariates, sender-specific covariates, and dyad-specific covariates.

Receiver-specific covariates: This category consists of variables that provide information regarding the intended receiver of a potential tie, and includes the aggregate number of reviews written until time $t-1$, the additional number of reviews written at time t, the total number of incoming links until time $t-1$, the additional incoming links at time t, and the average comprehensiveness, readability and objectivity scores across all reviews written until time t by the receiver. Among these variables, the total number of incoming links until time $t-1$ and the additional incoming links at time t are measures of the opinion leadership status of this receiver at time t. If the preferential attachment effect is prominent in our data, then the coefficients for these variables will be positive and significant. The aggregate number of reviews written until time $t-1$ is used to measure how active a reviewer has been until time $t-1$. The "recency" variable, constructed as the additional number of reviews written in time period t (i.e., in the last one month), measures how active a reviewer was in the most recent period.

We use the text mining tool *Lingpipe* (Alias-I 2008) to process the texts of the reviews, and obtain text properties such as comprehensiveness, readability and objectivity for each review. We use the number of sentences in the text of a review as an indicator of the Comprehensiveness of the review—generally longer texts contain more information, thus are expected to be more comprehensive (Otterbacher 2009).

We measure the Readability of a review by measuring the complexity of its writing style by calculating the Gunning-Fog Index (GFI) of the text of the review. This is a widely used measure in linguistics (DuBay 2004), and is calculated using the following formula:

Readability = GFI = 0.4*(average sentence length + number of hard words for each 100 words),

where a "hard" word is defined as a word with more than two syllables. Note that a larger value of Readability for a review implies that the review is *harder* to read.

To calculate the Objectivity of each review, we follow Pang and Lee (2004) and classify each sentence in the review as an objective or a subjective sentence (automated using a high-accuracy Support Vector Machine classifier pre-trained for movies on a large movie dataset, developed in Pang and Lee (2004)). In this case we follow the standard definition in the Machine Learning community—an objective sentence is one that talks about the plotline of the movie, and all other sentences are classified as subjective. Subsequently, the Objectivity of a review is defined as the total number of objective sentences divided by the total number of sentences in a review.

Epinions designates certain reviewers as "Top Reviewers" and displays this label next to their profile. It is reasonable to expect that reviewers with a rank label will obtain more trust links. We therefore include a covariate "Is Top Reviewer" which indicates the rank of a reviewer. Note that viewers do not know the exact rank of each reviewer and only observe whether the reviewer is a "Top 10," "Top 100," or "Top 1000" reviewer, or not a top reviewer at all. Therefore, we code the values of this covariate as 3, 2 and 1 if the reviewer is in the top 10, top 100 or top 1000, respectively, and 0 if the reviewer is not on the "Top Reviewer" list (i.e., we code the rank variable on a log scale based on the range in which the true rank falls).

Sender-specific covariates: This category consists of variables that provide information on the sender of a potential tie, and include the aggregate number of reviews written until time t, and the total number of outgoing links from this sender until time t. These variables are employed to control for how active a sender is. We would expect that senders who were more active in the past have a higher probability of extending links to other reviewers at a given point in time.

Dyad-specific covariates: This category consists of variables that provide information regarding the dyad in question and include measures for reciprocity, homophily and commonly trusted reviewers between the two individuals in the dyad. In our research, we measure reciprocity as a binary variable. If tie from j to i already exists at time *t*, the reciprocity variable equals 1, and 0 otherwise. We include the absolute differences in average readability, average objectivity and average comprehensiveness of the reviews written by *i* and j as observable measures of homophily. As mentioned earlier, we include a dyad-level term, d_{ij}, to account for unobservable homophily between i and j.

Finally, if the sender and receiver are connected to the same nodes then, as past research has shown, there is a higher chance of a link being formed (Hill et al. 2006). Therefore, we include as a covariate the commonly trusted reviewers between the sender and the receiver. Note that while our core hazard process treats dyads as independent, introducing this covariate relaxes that assumption.

In Table 1, we provide the variable definitions and descriptive statistics for these variables for our data for the "Movies" community.

Table 1: Variable Definitions and Descriptive Statistics

Variables	**Definition**	**Descriptive Statistics***
Receiver Characteristics		
Receiver's PrevAggReview	The aggregate number of reviews written until time $t-1$	1.34 (5.15)
Receiver's CurReview	The additional number of reviews written at time t	0.07 (0.61)
Receiver's PrevAggOpnLeadership	The aggregate number of incoming links until time $t-1$	0.68 (15.61)
Receiver's CurOpnLeadership	The additional number of incoming links at time t	0.02 (0.40)
Comprehensiveness	The average comprehensiveness of reviews until time t	14.41 (17.94)
Objectivity	The average objectivity of reviews until time t	0.21 (0.21)
Readability	The average readability of reviews until time t	14.06 (11.90)
Top Reviewer Label	The rank of the receiver as reviewer on "Top Reviewer" list at time t	0.0101(0.1002)
Sender Characteristics		
Sender's AggReview	The aggregate number of reviews written until time t	1.41 (5.32)
Sender's AggOutgoingLink	The aggregate number of incoming links until time t	0.71 (15.63)
Dyad Characteristics		
Dissimilarity in Objectivity	The absolute difference between average objectivity of reviews by sender and receiver until time t	0.02 (0.08)
Dissimilarity in Comprehensiveness	The absolute difference between average comprehensiveness of reviews by sender and receiver until time t	1.84 (7.95)

Dissimilarity in Readability	The absolute difference between average readability of reviews by sender and receiver until time t	1.79 (9.43)
Reciprocity	Whether the link from receiver to sender exists at time t	0.0003 (0.0160)
Commonly Trusted Reviewers	The number of reviewers trusted by both sender and receiver at time t	0.0022 (0.0697)

*Numbers outside brackets are the means for the "Movies" dataset, and those in brackets are the corresponding standard deviations.

5. Estimation and Results

Estimation Methodology: WESBI

We have 5,180 individuals in our calibration dataset, which generates 26,827,220 dyads. Since we need to calculate the hazard rate for each of 48 time periods for each dyad (January 2002 to December 2005), the total amount of computation is very time expensive. This is a challenge that is often encountered in large scale dyad-level studies of networks (e.g., Braun and Bonfrer 2011).

We develop a new methodology to meet the gap between the huge amount of data that needs to be processed and the limited computing power at our disposal. One of the key characteristics of our dataset is that the proportion of the dyads that actually form a tie is very small—only 1,906 ties are formed out of the nearly 27 million ties possible. To strike a balance between accurate estimation and computation time, we adapt the weighted exogenous sampling maximum likelihood estimator first developed in the choice-based sampling literature by Manski and Lerman (1977) for discrete-choice data. We extend the weighted exogenous sampling concept to timing data and also develop a Bayesian inference procedure for estimation and name our technique as *Weighted Exogenous Sampling with Bayesian Inference (WESBI).*

To employ this method, we collect all of the dyads which actually form ties within the observation time window, and randomly sample from the dyads which do not form a tie within the observation time window. By aggregating these two sets of dyads, we construct a much smaller dataset (we call this smaller dataset as the *sampled dataset*). And then, instead of maximizing the expression in Equation (1), we use the following *weighted* log-conditional-likelihood function for Bayesian inference over our new dataset:

$$\log L = w_1\left(\sum_{(\mathbb{I}_{ij}=1)}\left\{\log\left[1-\exp\left\{-\exp\left[\boldsymbol{\alpha}(k_{ij})+\boldsymbol{z}_{ij,k_{ij}}\boldsymbol{\beta}+a_i+b_j+d_{ij}\right]\right\}\right]\right.\right.$$

$$\left.\left.-\sum_{t=0}^{k_{ij}-1}\exp\left[\boldsymbol{\alpha}(t)+\boldsymbol{z}_{ijt}\boldsymbol{\beta}+a_i+b_j+d_{ij}\right]\right\}\right)$$

$$+w_0\left(\sum_{(\mathbb{I}_{ij}=0)}\left\{-\sum_{t=0}^{k_{ij}-1}\exp\left[\boldsymbol{\alpha}(t)+\boldsymbol{z}_{ijt}\boldsymbol{\beta}+a_i+b_j+d_{ij}\right]\right\}\right). \qquad (2)$$

w_1 and w_0 are the weights of the log-conditional-likelihood functions for the ties that were formed and the ties that were not formed, respectively. Here, $w_0 = \frac{1-Q_1}{1-H_1}$ and $w_1 = \frac{Q_1}{H_1}$, where Q_1 is the fraction of the ties formed in the whole population, and H_1 is the fraction of the ties formed in the sampled dataset.

We estimate the parameters of our model by using a MCMC hierarchical Bayes estimation procedure, using a Gibbs sampler and the Metropolis-Hastings algorithm. The full estimation procedure is provided in Appendix II. In Technical Appendix I, we show using a comprehensive simulation study that the WESBI method can accurately recover model parameters in a wide range of settings. Specifically, we find that sampling 10% of the empty dyads (and using all the dyads that actually formed ties) works well. Therefore, for the Epinions dataset, we sampled 10% of the dyads that did not form a link during our observation window. This final sampled dataset has 1,906 established ties, and 2,682,531 pairs that did not form a tie. While the estimation for the full dataset requires us to compute the likelihood of tie formation for 26,827,220 pairs given parameter values in each MCMC iteration, now we only need to evaluate the likelihood of tie formation for 2,684,437 pairs in the sampled dataset. Commensurate with this reduction in data, we reduce the estimation time by one order of magnitude while still obtaining accurate parameter estimates.

We highlight WESBI as a powerful estimation methodology that can be used for speedy but accurate estimation in other dyad-level network studies as well. Nevertheless, it is advisable for future users of WESBI to test its accuracy in settings which differ widely from those presented in our simulation results.

Estimation Results

We estimated our model in Matlab using the procedure in Appendix II. To reduce the autocorrelation between draws of the Metropolis-Hastings algorithm and to improve the mixing of the Markov chains, we used an adaptive Metropolis adjusted Langevin algorithm (Atchade 2006). We used the first 100,000 draws for burn-in and the last 25,000 to calculate the posterior distributions. To assess the convergence of the Markov chains, we ran multiple chains using a set of over-dispersed starting values and calculated the within-chain variance as well as between-chain variance for the chains for each parameter. The resulting scale reduction factor (Gelman et al. 2003) for each parameter is very close to 1. In the first column in Table 2, we present the posterior means of the coefficients in our model, after we standardize the values of all covariates. We discuss these results below.

Table 2: Parameter Estimates for Networks of Different Communities

Variables	**Movies**	**Expanded Network**	**Cars**	**Home & Garden**
Receiver Characteristics				
Receiver's PrevAggReview	0.1278	0.1094	0.1578	0.1889
Receiver's CurReview	0.8981***	0.5361***	0.5997***	0.5046***
Receiver's PrevAggOpnLeadership	0.4283***	0.3596***	0.4996***	0.3370***
Receiver's CurOpnLeadership	0.3048**	0.2167***	0.3710***	0.2961***
Comprehensiveness	0.3681*	---	0.1668*	0.0920
Objectivity	-0.1706	---	---	---
Readability	-0.1319	---	0.0855	-0.1537
(Comprehensiveness)^2	-0.4609***	---	-0.3571***	-0.3302***
(Objecivity)^2	-0.1147	---	---	---
(Readability)^2	-0.5193***	---	-0.3886**	-0.2408***
Top Reviewer Label	0.1478***	0.1845***	0.1939***	0.1648***
Sender Characteristics				
Sender's AggReview	0.3178*	0.0899	0.1636	0.3315*
Sender's AggOutgoingLink	0.1873	0.2604*	0.2876*	0.1311
Dyad Characteristics				
Dissimilarity in Comprehensiveness	-0.1695*	---	-0.2447*	-0.2875**
Dissimilarity in Objectivity	-0.2079*	---	---	---
Dissimilarity in Readability	-0.0583	---	-0.1866	-0.1683**
Reciprocity	0.3007***	0.1379***	0.3679**	0.3488***
Commonly Trusted Reviewers	0.2059***	0.1705*	0.2884***	0.2224***
Hazard Rate Parameters				

Log(a_0)	-13.7542***	-17.4816***	-13.4542***	-12.5319***
Log(a_1)	-5.0568	-4.8296	-4.4363	-3.8514
σ_d^2	0.1232***	0.1941***	0.2006***	0.3232***
σ_a^2	0.3650***	0.3586***	0.3883***	0.2846***
σ_b^2	0.2615***	0.2205***	0.4325***	0.1823***
σ_{ab}	0.1068***	0.1593***	0.1849***	0.1072***

***, ** and * denote that the 99% credible interval, the 95% credible interval, and the 90% credible interval, respectively, does not include zero.

Receiver-Specific Effects: We find that the coefficients for opinion leadership (both PrevAggOpnLeadership and CurOpnLeadership) are positive and significant. This offers evidence for the traditional preferential attachment argument where individuals with more incoming links have a higher probability of receiving additional incoming links in the current period, given everything else equal. The coefficients for the impact of reviews written by a receiver tell an interesting story. The coefficient of the number of reviews written in the current period (CurReview) is positive and significant, while the coefficient for the total number of reviews written until the previous period is insignificant (PrevAggReview). Intuitively, this indicates that only recent reviews boost a reviewer's reputation and attract other individuals in the community to put her in their respective webs of trust. On the other hand, the reviews written earlier do not influence others' decisions of extending outgoing links to her, and do not contribute to the emergence or the maintenance of opinion leadership. Note, however, that the coefficient for CurReview is larger than the coefficients for both PrevAggOpnLeadership and CurOpnLeadership.

Taken together, these results tell an interesting story—*recent* review activity is a stronger driver of opinion leadership status than preferential attachment, but preferential attachment is a permanent effect while *past* review writing activity does not have a significant effect. This is likely because trust links are not dated and therefore get valued as endorsements even if a long time has passed, while reviews become less valuable as the novelty of information they provide reduces as time passes. Therefore, existing opinion leaders (those who have a large number of inlinks) are at an advantage in terms of maintaining their position in the network. Contributing new content can also boost an individual's opinion leadership status; however, this effect is short lived. If new content leads to new trust links quickly, then these added inlinks will contribute to future opinion leadership increase through the preferential attachment effect.

A review's textual characteristics also have a significant impact on the emergence of opinion leadership. The coefficient for Comprehensiveness is significant and positive, and that of the squared term of Comprehensiveness is significant and negative. This indicates that members of the movie review community have an inverse-U shaped preference where reviews that are somewhat longer than average length are most preferred, while reviews that are either too long or too short are less preferred. The coefficient of the linear term of Readability is insignificant, while the coefficient of the squared term of Readability is negative and significant. This indicates that reviews with an average value of Readability are most preferred, while very simple or naïve reviews and very hard to read reviews are less preferred. The Objectivity of a review does not have an impact, possibly because readers may have varied preferences for objective versus subjective reviews, leading to an overall null effect. We also find that a top-reviewer label has a significant and positive impact on link formation in a dyad.

Sender-Specific Effects: We find that the aggregate number of reviews written by a sender (AggReview) has positive and significant impact on the probability that the sender extends ties to other individuals, which may indicate that there are some reviewers who are more involved in the community—they write reviews as well as develop their web of trust.

Dyad-Specific Effects: We find that reciprocity has a positive and significant impact on the formation of network ties, which is in agreement with many other studies. Our results for the dissimilarity of textual characteristics between two reviewers also support the traditional homophily argument. This is clear from the negative coefficients for dissimilarity of comprehensiveness and objectivity. We also find that the number of commonly trusted reviewers has a significant and positive impact on the formation of a link in a dyad.

Baseline Hazard Rate: From the hazard rate parameters in Table 2, we can see that, as expected, the general tendency of forming links is relatively small in this online community ($\alpha_0 = 1.06 \times 10^{-6}$). Furthermore, we find that the reviewers' baseline hazard rate of forming links decreases over time $(\alpha_1 = 0.0064)$, which is similar to the effect of decreasing activity over time typically observed for individual-level activity in the customer-base analysis literature (e.g., Fader et al. 2005).

Unobserved Random Effects: The fact that $\sigma_a, \sigma_b$and σ_d are significant indicates that random effects at the sender, receiver and dyad levels exist in the community, above and beyond the covariates that we use in our model. Moreover, σ_{ab} is significant and positive, which suggests that reviewers who are intrinsically more attractive are also more active in extending links to others.

Model Performance

To test the performance of our model, we use two alternative models as benchmarks: 1) a time-invariant hazard model with all covariates (as in Narayan and Yang (2007)), and 2) a time-varying hazard model with only network characteristics (and no node-level characteristics) as covariates (i.e., Receiver's PrevAggOpnLeadership, Receiver's CurOpnLeadership, Sender's AggOutgoingLink, Reciprocity and Commonly Trusted Reviewers). Traditional model performance statistics that provide accuracy measures averaged over all dyads cannot serve as good measures because the ties formed in the network are extremely sparse.[6] Hunter et al. (2008) proposed procedures to evaluate how well a model fits real data in a social network context based on key structural properties of the network. Hunter et al. (2008) proposed degree distribution, dyad-wise shared partner distribution, and the distribution of geodesic distances as test statistics to assess the goodness-of-fit of social network data. However, Hunter et al. (2008) proposed these statistics for an undirected network. As we deal with a directed network, we use in-degree distribution, dyad-wise commonly-trusted reviewer distribution, and the distribution of geodesic distances as our model fit statistics. All the test statistics we report in this section are with respect to the holdout sample.

We first calculate the values of the test statistics for the holdout period of the actual network. We then simulate tie formation in the holdout period using our full model and the two benchmark models. We calculate the test statistics for each model by running the simulation 200 times. We compare the distributions obtained from our full model and the two benchmark models with the true distributions in Figure 2. In each figure, the *x*-axis depicts the test statistic, while *y*-axis depicts the percentage of individuals or dyads corresponding to the test statistic in the holdout sample (on a log scale). The solid black dots represent the test statistic from the actual dataset, and the boxes-and-whiskers represent the corresponding statistics across the simulated datasets. The whisker represents the upper and lower limits of the 200 corresponding simulated network statistics. The box represents the 25th and the 75th percentile. If a box is missing for a specific value of a network characteristic, it indicates that there is not even a single corresponding observation across 200 networks. (For example, in the first panel in Figure 2(a), the box for in-degree ≥4 is missing.

[6] Even a naïve model which predicts that no pairs form ties has an accuracy of 99.99% as only 160 ties are formed among 2,324,100 possible pairs in the holdout sample.

This indicates that among the 200 simulated networks for the time-invariant hazard model, no network has a node with in-degree that is ≥4).

From Figure 2(a), we can see that the in-degree distribution from the full model is very close to that for the actual network. In contrast, the time-invariant hazard model shows significant deviations from the observed distribution for in-degree ≥3, and for the model with only network characteristics included, the predicted in-degree distribution differs significantly when the in-degree is ≥2. In other words, our full model performs significantly better than the two benchmark models in predicting the in-degree distribution. From Figure 2(b), we can see that the actual data statistics for commonly trusted reviewer lie within the boxes corresponding to the full model, indicating an excellent fit. In comparison, for time invariant and only network characteristics models, the actual data often lies outside the box or even the whiskers. From Figure 2(c), we can see that our full model outperforms the two benchmark models on accurately predicting the geodesic distance distribution also.

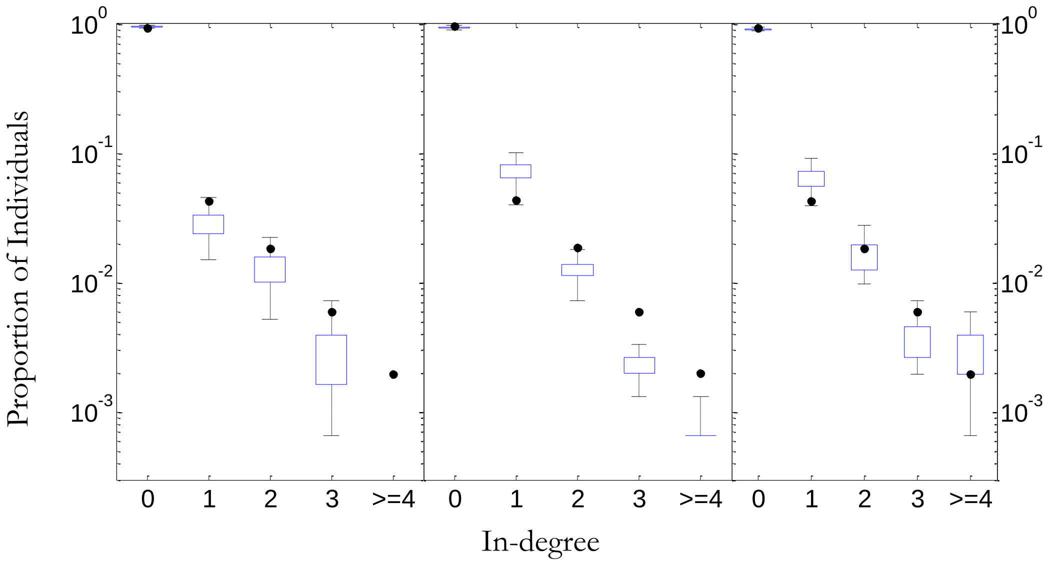
From Figure 2, we can conclude that our full model (time-varying hazard model with all covariates) not only performs well in predicting key network statistics in the holdout sample, but is also superior to the two alternative benchmark models. To illustrate the importance of node-level characteristics, we can see that the performance of the model with only network characteristics is always lower than our model as well as the time-invariant hazard model. This emphasizes that node characteristics are a major driver of link formation in the network evolution process. The time-invariant hazard model is more stable than the model with only network characteristics; however, its performance is also significantly inferior to our full model. The above performance tests strongly indicate that our proposed model performs significantly better than the benchmark models, which shows the importance of incorporating both node characteristics and dynamics into the model.

Figure 2: Performance Tests
(a) In-degree Distribution

Time- Invariant Hazard Model

Only Network Characteristics

Time-Varying Hazard Model

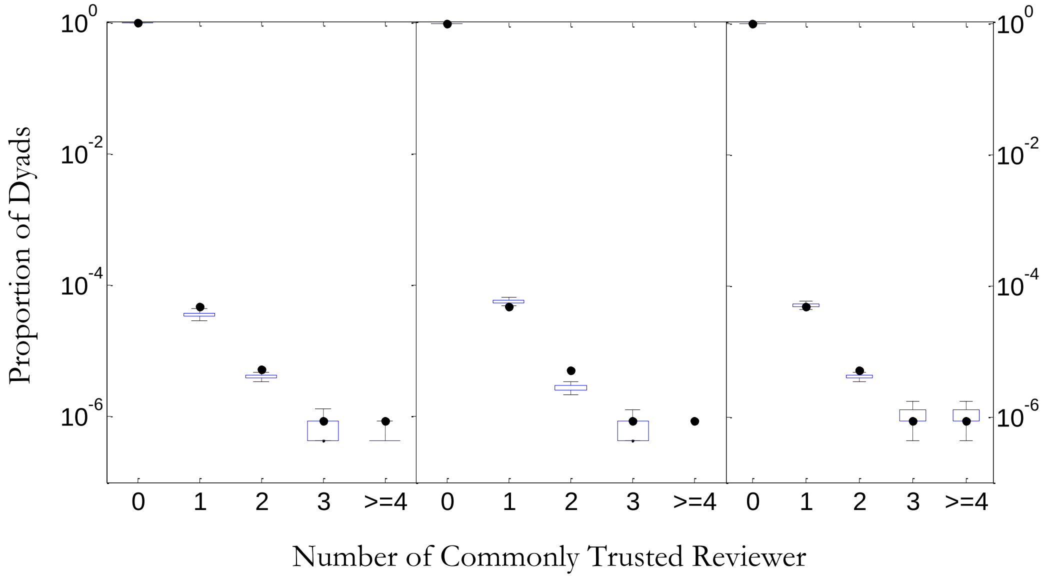


(b) Dyad-wise Commonly Trusted Reviewer Distribution

Time-Invariant Hazard Model

Only Network Characteristics

Time-Varying Hazard Model

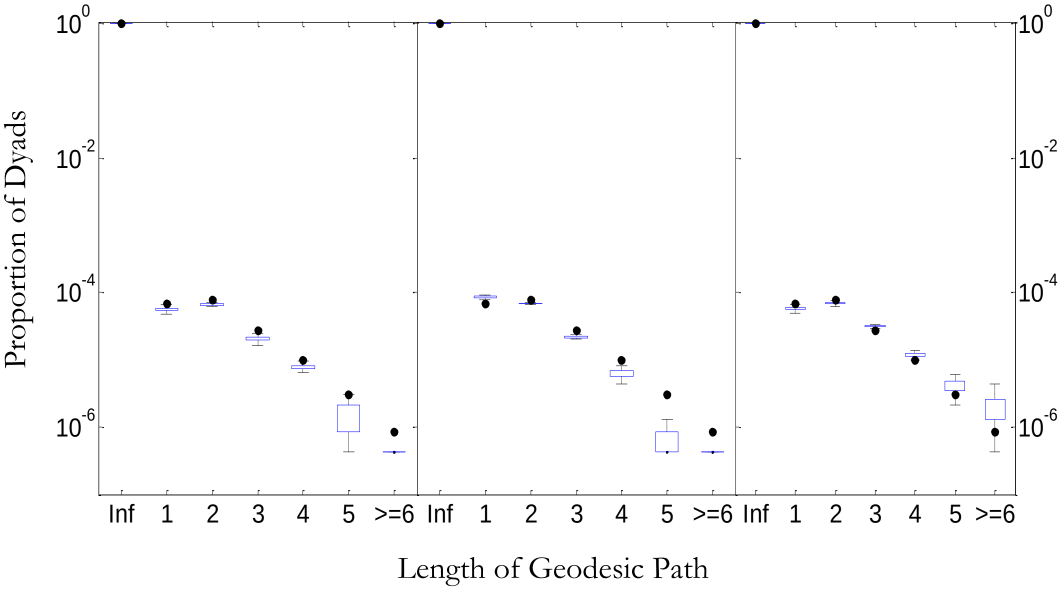


(c) Geodesic Distance Distribution

Time-Invariant Hazard Model

Only Network Characteristics

Time-Varying Hazard Model



6. Extensions and Robustness Checks

In this section, we extend our basic analysis in three different ways. First, we stratify our dataset based on opinion leadership status and find that the strategies of forming links employed by individuals with high opinion leadership statuses are very different from those employed by individuals with low opinion leadership statuses. Second, we consider an expanded network by crawling data independent of categories and also including followers of reviewers who may not have written any reviews. Third, we conduct our analyses in two other product categories.[7]

Analysis with Stratification Based on Opinion Leadership

We use the dataset described in Section 4 and classify all individuals in our sample into two groups based on their opinion leadership statuses. Individuals with <10 incoming links at the end of our calibration period (December 2005) are classified as having low opinion leadership status (LOLS), and the remaining individuals are classified as having high opinion leadership status (HOLS). Based on this, 5,100 and 80 individuals are classified in the LOLS and HOLS categories, respectively. We then stratify all dyads into two groups based on the type of sender. The first group corresponds to all pairs where the tie sender has low opinion leadership status, and the second group corresponds to all pairs where the tie sender has high opinion leadership status. To illustrate how the behavior of

[7] In addition to these extensions, we also estimate a random coefficients model to capture the potential unobserved individual heterogeneity. We find that the impact of preferential attachment and recency are qualitatively the same as in the model with homogenous individuals. Details are available in Technical Appendix II

these two groups of senders differs from each other, we estimate our model for the two samples separately. We report the results in the first two columns of Table 3.

Table 3: Parameter Estimates for the "Movies" Category for Individuals with Different Opinion Leadership

Variables	All Links That Are Formed in Dataset Are Included		Only Links That Are Formed First Are Included	
	Low Opinion Leadership	High Opinion Leadership	Low Opinion Leadership	High Opinion Leadership
Receiver Characteristics				
Receiver's PrevAggReview	0.0347	0.1961*	0.0358	0.1972*
Receiver's CurReview	0.6368***	0.4134**	0.6276***	0.4017**
Receiver's PrevAggOpnLeadership	0.3828**	0.0583	0.3811**	0.0541
Receiver's CurOpnLeadership	0.3533**	0.0420	0.3391**	0.0392
Comprehensiveness	0.4105*	0.1951***	0.4224*	0.2126***
Objectivity	0.1452	-0.1374*	0.1315	-0.1417*
Readability	0.0525	-0.1271*	0.0414	-0.1378*
$(\text{Comprehensiveness})^2$	-0.5152***	-0.1022***	-0.5241***	-0.1216***
$(\text{Objectivity})^2$	0.0436	-0.0896	0.0487	-0.0802
$(\text{Readability})^2$	-0.4960***	-0.2610***	-0.5128***	-0.2602***
Is Top Reviewer	0.1785***	-0.1432**	0.1763***	-0.1491**
Sender Characteristics				
Sender's AggReview	-0.1019***	-0.2410***	-0.0988***	-0.2429***
Sender's AggOutgoingLink	0.0059***	0.0900	0.0062***	0.0816
Dyad Characteristics				
Dissimilarity in Comprehensiveness	-0.2319*	-0.0633	-0.2332*	-0.0602
Dissimilarity in Objectivity	-0.1251***	-0.2205***	-0.1121***	-0.2251***
Dissimilarity in Readability	-0.0468	-0.0006	-0.0438	-0.0008
Reciprocity	0.2447***	0.4094***	---	---
Commonly Trusted Reviewers	0.1643***	0.1909**	0.1629***	0.1948**
Hazard Rate Parameters				
Log(α_0)	-14.7342***	-11.4360***	-15.5124***	-12.4193***
Log(α_1)	-4.7773	-5.3882	-4.2149	-5.3251
σ_d^2	0.1656***	0.1198***	0.1643***	0.1219***
σ_a^2	0.4253***	0.3760***	0.4227***	0.3817***
σ_b^2	0.3728***	0.4617***	0.3721***	0.4642***
σ_{ab}	0.1651***	0.1867***	0.1643***	0.1899***

***, ** and * denote that the 99% credible interval, the 95% credible interval, and the 90% credible interval, respectively, does not include zero.

We uncover an interesting insight into the contrasting strategies for extending trust links employed by individuals with low and high opinion leadership statuses. While low opinion leadership status individuals extend links to others who have high previous and current opinion leadership status and are top reviewers, high opinion leadership status individuals extend links to low-status individuals. One potential explanation for this finding is provided by Mayzlin and Yoganarasimhan (2012): those with a weak network position (LOLS reviewers) want to signal their ability by finding and linking to HOLS reviewers, while those with a strong network position (HOLS reviewers) do not want to promote other strong individuals (HOLS reviewers) as competitors. In addition, LOLS individuals, who can in fact be considered opinion seekers, are seeking access to high-quality reviews for themselves, which individuals identified by others as top reviewers or opinion leaders can provide. In comparison, the HOLS individuals want to retain their followers and gain even higher leadership status by attracting others. Hence, a high opinion leadership status individual would not prefer to extend a link to another high opinion leadership status individual as she may risk losing her followers to the other opinion leader.

We now conduct a robustness check to alleviate the concern that reciprocity drives the results presented above. We estimate our model with the same stratification of the data as above, but, for pairs of nodes that have reciprocated links, we include only those links that are formed first. In other words, if A and B are two nodes with the edges A→B and B→A both existing, and, say, A→B is formed before B→A is formed, then we remove the edge B→A from the data. By artificially removing all the links that could possibly be reciprocated, we completely remove reciprocity as a possible factor in link formation.[8] We provide the results of the model estimated on these data in the last two columns of Table 3. Comparing these estimates with the estimates in the first two columns of Table 3, we find that there is no qualitative difference between the two sets of results.

Analysis for an Expanded, Category-Independent Network with "Followers" Included

In Section 5, we considered only the "Movies" community. In this section, we test our findings on a much larger, category-independent dataset in which we also include individuals who only passively

[8] We thank an anonymous reviewer for suggesting this analysis.

follow other reviewers without themselves writing any reviews. To collect this dataset, in the first step, we use all individuals in the "Movies" community (as described in the *Data Description* section) as the seeds for network crawling. To cover the possibility that some parts of the network are unreachable from the "Movies" community, we further randomly sample 100 individuals from every other product review community, such as "Cars," "Computers and Software," "Home & Garden," etc., and include them as part of the seed group as well in this step. In the second step, we start from this seed group, and collect data on all individuals who are in the webs of trust of the members in the seed group, as well as all individuals who put members in the seed group in their web of trust. These new members are then included in the seed group. We repeat the second step until this crawled network stops expanding. Considering individuals who registered on the website between January 2002 and December 2008, we obtain a network with almost twice the number of nodes as in the calibration data described in Section 4, and includes 10,669 individuals with 3,396 ties. Based on this much larger network, we estimate our model (without considering the textual characteristics of reviews). We present the results in the second column of Table 2. These results show that, in this much larger network as well, the effect of intrinsic node characteristics on the dynamics of network evolution differs from the effect of network-based node characteristics—while the impact of previous opinion leadership carries over into future periods, previous reviews written have no significant impact on the rate of forming ties.

Analysis for Other Product Categories

To check the robustness of our estimation results, we replicated our analysis on the "Cars" and the "Home & Garden" categories. We construct the datasets for these two categories by restricting ourselves to reviewers who entered between January 2002 and December 2008 and wrote at least one review on the topic of the associated community.[9] The resulting "Cars" reviewer community includes 1,059 individuals with 225 ties formed within the community, and the "Home & Garden" community comprises of 1,120 individuals with 457 ties formed within the community. We present the results for the "Cars" and the "Home & Garden" communities in the third and fourth columns of Table 2, respectively.

[9] We used snowball sampling to collect data for this network, which implies that we only detect individuals whom at least one other person has included in her web of trust. For the "Movies" category, we could start with a list of movies for which reviews were written and detect individuals who wrote reviews but were not connected to others. Such an exhaustive list of products for "Cars" and "Home & Garden" is extremely difficult to construct, so we work with this limited dataset for this extension.

As we can see in Table 2, most of the results that we found for the "Movies" community—most notably the result that only recent reviews, and not past reviews, have an impact on opinion leadership status, while both past and recent trust links have an impact—also hold for the "Cars" and "Home & Garden" categories. Note, however, that in both the "Cars" and "Home & Garden" categories, the recency effect is weaker than that in the "Movies" category. One possibility is that readers in the "Movies" community care more about movies that are released recently rather than about old movies, leading to a stronger recency effect. Interestingly, the fact that this effect is salient in both the "Cars" and "Home & Garden" communities, in which more recent products are expected to be less important for consumers than in the "Movies" category, indicates that the recency effect argument is applicable in a wide range of scenarios.

7. Conclusions and Managerial Implications

We model opinion leadership in a community using a social network paradigm. We show that while phenomena highlighted in the extant literature, such as preferential attachment and reciprocity, are important drivers of network growth, intrinsic properties of nodes such as recent activity and the style of writing reviews (objectivity, readability and comprehensiveness) are also very significant drivers of network growth and, in our context, drivers of opinion leadership status. Our study is one of the first to investigate opinion leadership in a longitudinal setting with specific details about the opinion shared also available (such as the time of sharing opinion and the content), and we significantly extend the emerging literature on reputation building in online environments (Forman et al. 2008, Ghose et al. 2009). By incorporating the time dimension into our study, we find the novel and important result that intrinsic node characteristics are a stronger short-term driver of additional inlinks, while the preferential attachment effect has a smaller impact but it persists for a longer time. Our results are robust and hold consistently for the several different communities and network definitions that we consider.

Our findings have several important managerial and design implications for opinion-sharing websites. (While we discuss the managerial implications in the context of Epinions, we believe they will be valid for the numerous other networked online opinion sharing communities as well, such as Motley Fool, Seeking Alpha, IMDB, Yelp, etc.) The manner in which Epinions and most other online review communities are currently designed, the presence of dominant reviewers whom a large number of individuals already trust might hamper the emergence of new high-quality reviewers. This is because preferential attachment has a persistent impact on inlinks received while review generation does not (unless it leads to new inlinks fairly quickly). Therefore, though it is not impossible for new reviewers who write up-to-date and high-quality reviews to become opinion leaders, it is nevertheless quite difficult. A very simple and practical managerial solution to this issue could be to attach a "lifetime" to the trust links, so that these votes of trust can be allowed to "expire" after a certain period of time. This would ensure that reviewers cannot rest on the opinion leadership status that they have earned in the past. They will have to constantly share high-quality opinion, or else have to secede opinion leadership status to new individuals offering high-quality opinions, which will lead to an overall increase in the quality of information available in the community.

Furthermore, in any large online social network such as Epinions, it is a difficult task for users to find relevant individuals among thousands of candidates for relationship formation. Epinions can leverage our results in many ways to help reduce the cost of such search. For example, it could display the list of recently-most-active reviewers along with the reviewers with the highest recent increase in opinion leadership. It could also develop and include a recency score for each reviewer as additional information in its search results ranking algorithm. Epinions can also ask readers to rate reviews on different characteristics such as comprehensiveness, readability, and objectiveness (or automate this process using text mining). It can then use these results to provide an average score for a reviewer on these characteristics. This would help the reader in deciding as to whether or not to read a review, and whether or not to extend a trust link to a reviewer. Epinions could also provide a search tool which could allow users to search reviews for a product based on these desirable characteristics.

Our study contributes not only towards furthering our understanding of how opinion leaders emerge in networked communities, but also underscores the importance of incorporating node-level characteristics in network growth models, a factor that has received limited attention in the extant literature. Our results offer an explanation for why the power law coefficient for the in-degree distribution for the particular online network from Epinions that we work with (having a value of 1.74) is smaller than the values of power law coefficients for in-degree distributions typically predicted by the theoretical preferential attachment literature (between 2 and 4, Barabasi and Albert 1999). (Note that this is true for various other popular online communities as well. For example, Mislove et al. (2007) finds that the power law coefficient for the in-degree distribution is 1.63 for YouTube and 1.74 for Flickr.) Intuitively, if individuals also take inherent node characteristics beyond in-degree (in our case, reviewer and review characteristics) into account when they form ties, and individuals do not extend links to nodes with inferior node characteristics, then superior node characteristics could help individuals attract additional incoming links compared with networks with pure preferential attachment. In this case, the power law coefficient of the in-degree distribution will be smaller, as we find it to be. In fact, differences in the relative importance of node characteristics for tie formation across different networks studied in the extant literature may explain the differences in their power law coefficients. Following the arguments above, communities in which node characteristics are important will have smaller power law coefficients. This suggests that when researchers investigate the evolution of a network, they should not focus solely on network characteristics such as degree, betweenness measures, etc.; they should also take into account how characteristics of individuals can influence the evolution dynamics in a social network. (Note that theories of diffusion over existing networks and formation of networks at a small-scale consider characteristics of individuals. However, the literature, cited earlier, on generative models of large-scale networks has largely overlooked the importance of node characteristics.) Therefore, these findings also contribute to the vast literature on scale-free networks, why their macro-level characteristics may vary across different settings, and why their degree distributions may not always be as skewed as theoretical models based on preferential attachment would predict.

From the methodological perspective, we contribute to the literature on networks by developing a proportional hazard model of network evolution that is able to capture how time-varying covariates can influence the probability of forming a directed tie between two nodes in a network. We extend the weighted exogenous sampling maximum likelihood estimator developed by Manski and Lerman (1977) for binary choice data to duration data. Furthermore, we introduce a hierarchical Bayesian adaptation of the weighted exogenous sampling maximum likelihood estimator as a fast and effective way of dealing with the huge amounts of data that researchers and firms are typically faced with in the estimation of dyadic models on network data. Often, the solution employed is to either simplify the model to be estimated, or randomly sample a small part of the total population to reduce computational requirements. Our method, which involves selective sampling followed by appropriate reweighting of the sampled dyads, will help to reduce the degree to which such compromises need to be made. The results from our simulation show that our proposed method can serve as a very effective heuristic when dealing with large scale network data in a wide range of settings. However, since we do not provide theoretical proofs, we suggest that researchers should check the accuracy of the WESBI method as appropriate for their setting before using it.

Our study can motivate future research in several directions. First, in this study we assume that changes over time in the review writing styles (which are, in fact, minimal in our data) and in the frequency of writing reviews are exogenous. It is possible that a reviewer may learn over time and adjust these factors based on the readers' response to her past reviews. A study that investigates reviewer learning would be influential in understanding the important but understudied review-generation phenomenon. Second, our stratification analysis in Section 6 indicates that reviewers are strategic in extending trust links to other reviewers based on opinion leadership status. It may be interesting to investigate this in future studies. Third, we have only captured link formation and have not looked at link dissolution as the data that would be required are not available to us. Future studies can try to collect such data and study the factors that affect link dissolution. Finally, it may be interesting to consider the impact of product release frequency in a category on review generation and web of trust formation.

Chapter 3

Learning from Peers on Social Media Platform

1. Introduction

The vast and expanding reach of Web 2.0 technology has convinced companies of the potential of social media platforms for knowledge sharing among customers and employees. By engaging customers and employees using social media platforms, companies are able to harness the power of collective intelligence, manage customer relationships and lower operational cost. As early adopters, Microsoft IBM, CISCO, Infosys, Dell, Sun Microsystems, etc. have utilized various types of social media platform[10] within the company to support ideation, crowd sourcing and project management (Bayus 2010)[11].

Recently, a growing number of companies have built internal online forums where customer-support staff can learn from their peers to help resolve customer problems. On the forum, customer-support employees can post questions coming from customer side. At the same time, other employees are encouraged to answer these questions. As such, employees learn from each other, and customer service is improved by providing just-in-time services, assisting customer learning and saving significant customer service costs. Deloitte (2010) reports that OSIsoft saw a 22 percent decrease in the time required for resolving customer support issues due to its use of Web 2.0

[10] Blogs, wikis, micro-blogging, prediction markets, crowdsourcing etc. are other kinds of Web 2.0 platforms that organizations typically adopt for marketing purposes. In this paper, we focus on discussion forums that facilitate knowledge sharing.

[11] A recent survey by the Aberdeen group reports that 300 early-enterprise adopters who widely harnessed social media saw a 36% decrease in the time required to enact key business changes based on customer feedback, while the laggards experienced a 17 percent increase. Further, the best-in-class companies took three months less than the industry average to complete key strategic projects; and took an average of 11 hours to bring a response team together for a key business threat, while the industry average was 113 hours.

platforms. As an example of a discussion forum, Table 1 shows an example of one question followed by answers to it on an online discussion forum.

Table 1. An Example of Knowledge Seeking and Sharing in Online Discussion Forum*

Topic: How to upload an email having attachments in java for clients?

Post Type	Author	Time	Content
Question	**A@xxx.com**	**12/11/2006** **3:06:28 PM**	Hi, We have a client requirement where user needs to upload an email (*.msg) having attachments to a file server. My query is whether it is like any other file upload? Can we view the email and the attachments again? Another requirement is the user can have attachments up to 1 GB in the email. Is it feasible? If not what should be the feasible attachment size. Any thoughts /suggestions are welcome.
Answers:	**B@xxx.com**	12/12/2006 4:51:26 AM	It is generally not a good idea to send large email messages. A lot of mail handlers will not cope with messages larger than a few MB in size. If they don't have enough space to deal with it, your message will likely be returned to you. Your message may need to pass through several mail handlers before it reaches its final destination. If the client do need to send a large file to another user, it is best to set up some other arrangement than to use email. FTP (File Transfer Protocol) is a very good choice for this type of task.
	C@xxx.com	12/12/2006 9:24:35 AM	Yes. When message is saved in .msg (Message Format File), the attachment is also stored as part of the file. So, .msg file is like any other file to be uploaded. This .msg file can be opened back using Outlook. If the .msg file size is huge, it is better to compress and decompress using the files under java.util.zip package.
	D@xxx.com	12/15/2006 9:06:57 AM	I think the concerns raised by Our colleagues are all important. Also it must be noted that the recipient mailbox should also be able to handle the Message size. Under such circumstances its best to look at alternatives. An alternative would be store the file on a server and generate an https secured link for the file to be accessed within a specified period. This kind of solutions require administrative tasks and hence best to get the requirements completely; and then design the solution.

*These asking and answering posts source from a real online discussion forum within an organization.

This type of online forum is also widely used across the boundary of the company as a mean for engaging customers and customer-support staff to enhance customer service. Procter and Gamble, Dell and Oracle are just a few examples among others which gain significant benefits from such social media initiatives. An anecdote suggests that one "diamond member" of Dell's community support forum helped Dell save up to an astonishing $1 million by handling other customers' questions (Li and Bernoff 2009). In another example, with only 14 customer-service employees and no call centers, a telecom company adopted Web 2.0 based systems to encourage employees and customers to answer questions on their online forum. Using this system, not only did customers get their problems resolved within three minutes, but significant costs involved in providing customer service were also saved (Buchanan 2010).

Compared with existing learning models (Erdem and Keane 1996, Erdem et al 2008), there are two salient characteristics of learning that are enabled by a discussion forum: "learning from peers," and "knowledge-spillover." Learning cannot be achieved without knowledge sharing by other users, and users will gain knowledge in online discussion forums only when their peers contribute knowledge to the community. Meanwhile, not only users who ask questions gain knowledge from reading answers contributed by their peers, but everyone else who participates in the community also has 24/7 access to the repository and also learns from the posted answers.

These two characteristics of learning in online social media highlight the important impact of user interdependency on sustaining active participation from members of the community. In traditional learning channels where individuals have direct control over their information updating processes (Erdem and Keane 1996, Erdem et al 2008), incentive schemes of the company have a relatively direct impact on individual decision making in their learning process. However, in online social media platform, individual learning processes critically depends on each others' decisions, thus company policies could only indirectly influence individual learning process through improving user interactions. As a result, it is of great importance of understanding how the user interdependency influences the knowledge sharing on a social media platform. Uncovering this interaction mechanism will help the company design desired incentive structures and corporate policies which can maximize the return of investment of an online social media platform initiative.

To understand the interdependency among individual decisions, we draw on marketing, economics and social psychology theories in the context of employee internal usage of enterprise 2.0 system and present a dynamic structural model. In this model, users decide whether to ask a question, and whose question to answer to maximize a long-term utility that depends on knowledge, social status, and the cost of actions. The proposed model recognizes "learning from peers" by allowing decisions of the users to depend on how all of their peers will respond. It also takes into account "knowledge spill-over" by allowing each user's action to update the state variables (knowledge and social status) of everybody in the community. Thus, the knowledge seeking and sharing decisions of all the users are allowed to be inter-temporally dependent. By allowing the users to decide whose questions to answer, our model also treats the formation of the network as an endogenous decision that is driven by knowledge accumulation and social-status building within the community. This model is in the same spirit as the multi-agent dynamic game with imperfect information described by Ericson and Pakes (1995), Benkard (2004) and Bajari et al. (2007). Based on Ericson and Pakes (1995), we focus on the Markov Perfect Equilibrium as the solution concept for this dynamic competitive game. We estimate the dynamic competitive model by adapting two-step approach of Bajari et al. (2007) to the case of continuous state variables (Bajari et al. 2008). [12]

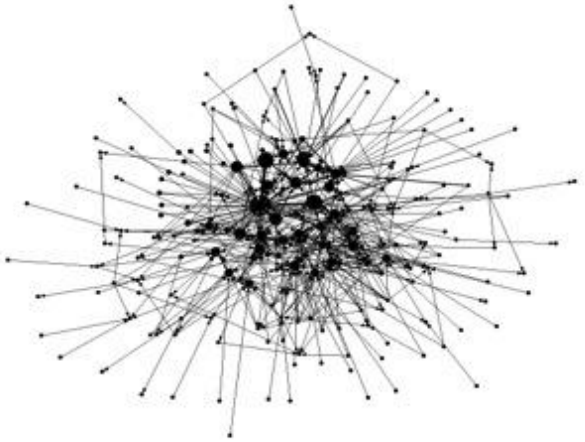
[12] We considered several approaches and employ the two-step estimation developed by Bajari et al (2007) because of two reasons: 1) We have very large number of individuals in our research context. Two-step estimation allows us to feasibly recover individual policy from observe data as in Ericson and Pakes (1995); 2) Given our research focuses on examining how formation of network affects knowledge sharing, we have to explicitly incorporate individual decisions on dyad-level into our model, where utility is indirectly obtained through peers' decisions (whose question to answer, and

In this research, we illustrate two mechanisms in which decision interdependency among members influence the efficiency of knowledge sharing within an organization. First, we find that an individual is more willing to contribute to the community when her peers are more knowledgeable. That is, the users were less willing to contribute when the community needed help. We show that this effect can be explained by a dynamic and interdependent decision-making process. An individual is more likely to receive a reciprocated reward by a more knowledgeable community and thus is more willing to share his/her knowledge with others in the community. Second, we find that the community revolves around a set of central actors who are well connected with each other, leading to the formation of a core/periphery network structure[13]. Figure 1 documents the core/periphery structure of peer interactions among the discussion forum adopters in our research setting. We demonstrate how the dynamic, interdependent decision-making process among individuals results in cohort formation among the centrally located users, who tend to answer questions from each other. We further show that the existence of this "privileged" circle discourages users outside of the privileged circle from participating, and creates a barrier to knowledge sharing and integration for the company.

Figure 1. The Core/Peripheral Network Structure[a]

how my contribution will be reciprocated from the community. Oblivious Equilibrium framework is not applicable in our context, because this method assumes that agents only take into account aggregate state of peers and that agents play long-run equilibrium strategies.

[13] Core/periphery structures have been documented in the sociology literature (McPherson et al 2001, Borgatti and Everett 2000) and in a number of Web 2.0 settings: open source software (Singh and Tan 2010), blogs (Obradovic and Baumann 2009), and micro-blogging (Huang et al. 2010). Central actors are active contributors to the community and are connected to both central and peripheral actors. Peripheral actors, by contrast, are connected to the central actors but not to each other. Such a social network is referred to as a core/periphery network.



a. Individuals are represented by spheres. Lines connecting two individuals represent the presence of a knowledge-sharing relationship between them. The arrow heads point towards the individual who answers the question. More active participants are indicated by larger spheres.

Our dataset is provided by a multinational IT service and consulting firm. It includes the complete history of the customer-support employees using the social media platform to ask and answer the questions that are generated when the focal company provides IT consulting service to its client companies. Based on the dynamic and interactive decision process, we run several analyses to explain the following phenomena: 1) the higher likelihood of knowledge sharing by individuals when the community is more knowledgeable; 2) the formation of a cohort that discourages participation; and 3) the greater effectiveness of proactive learning by asking questions compared to reactive learning by knowledge acquisition through reading. We also conduct an exploratory sensitivity analysis to show that hiding the names of the knowledge seekers (but not the sharers) breaks the core/periphery structure of the community and increases knowledge acquisition by approximately 20%.

Our research contributes to the marketing literature in the following ways. First, this is the first paper to examine online peer learning and explicitly model the dynamic and interdependent decision process to investigate the key factors driving user participation in a social media knowledge sharing platform. While previous literature focus on “learning by doing” and treat customers as isolated, we explicitly model the dynamic interactions among individuals. Second, our study endogenizes the formation of a social network and allows its evolving structure to affect the users' knowledge-sharing decisions. This concept is in contrast to most of the existing literature, which treats the social network as an antecedent to knowledge sharing. Third, we advance the learning literature by treating knowledge sharing as a consequence of dynamic strategic interactions between individuals. This approach is different from traditional learning models that take the atomistic view that individuals learn either from consumption experiences or quality signals, such as price and advertising. Managerially, we are one of the first papers to investigate user-participation decisions in social media platforms. The results provide insights for managers who want to evaluate their social-media policy and platform designs.

2. Literature Review

Our paper is related to the marketing literature on user decision making in online communities, and consumer learning. First, our paper is related to the marketing literature on customer behavior in online communities. It has been shown that online social media have a significant impact on consumer purchasing decisions in various industries, such as television (Zhang and Wedel 2008, Chevalier and Mayzlin 2006), movies (Chintagunta et al. 2010), and publishing (Godes and Mayzlin 2004), etc. However, researchers have only recently started to investigate the dynamics of social communities (Katona and Sarvary 2008). Mayzlin and Yoganarasimhan (2008) propose an analytical model analyzing how individual heterogeneity affects the ability to post breaking news and how the ability to find news in the blogs of others influences the bloggers' link-formation decisions and their strategic links with their competitors. Stephen and Toubia (2010) find that sellers in an online social-commerce marketplace derive significant benefit from connection with peers, and this benefit primarily comes from the accessibility enhancement of the network. Narayan and Yang (2007) model the decision of one individual trusting another whose reviews are found to be consistently helpful in an online review community.

The question of why people contribute to online social media has received increased attention in the marketing literature only in recent years. Lurie et al. (2009) suggest that user identities, such as expertise, social connections and symbolic incentives (forum points, in this case), can affect individual contributions to the community. Trusov et al. (2010) show that users' activities on Facebook are significantly influenced by a proportion of their friends' activity levels. Kumar et al. (2010) is among the pioneers to employ a dynamic structural model and rationalize that individuals contribute to connected goods primarily because of self-expression, social-status competition and consumption utility from peers. Being the first to endogenize link formation in an empirical framework, Ma et al. (2010) simultaneously investigate the content creation and link-formation processes in an online review community and find that reviewers with more content and low network status are more likely to contribute to online social media. Substantively, the most pertinent research was conducted by Bayus (2010), who examined the contribution of ideas from users on crowd-sourcing platforms and found that productive individuals are likely to have creative ideas and are unlikely to repeat their early creative successes once their ideas are recognized.

The majority of the existing research on social communities treats the social network as an antecedent to an outcome of economic interest and takes a frequentist perspective. Only recently have researchers started to investigate social network formation (Katona and Sarvary 2008) and employ dynamic structural models to better illustrate the dynamics of individual decision making within social networks (Hartmann 2010, Kumar et al. 2010, Huang et al. 2010, Ma et al. 2010). Our research builds on these findings and provides a more theoretically-grounded understanding of network evolution than that which currently exists in the literature. We treat the social network as a consequence of the strategic utility-maximizing actions of individuals.

Our work is also related to the marketing literature on consumer learning, which focuses on understanding how individuals learn about the quality of a product through consumption (Erdem and Keane 1996), information gathering (Erdem et al 2005) exposure to quality signals contained in the price, advertising, branding (Erdem 1998; Erdem et al. 2008; Narayanan and Manchanda 2009), and peer choices (Zhang 2010, Iyengar et al. 2008, Van den Bulte and Lilien 2001). The traditional learning models take an atomistic view of an individual and assume that customers cannot share information about the product. By comparison, we investigate peer learning on a public forum, which is an alternative learning mechanism that is characterized by sharing and the externality of learning. This highly distinct learning mechanism inherently implies that any user's decisions cannot be made independently of the others (that is, it implies interdependence) and that there is a long-run spillover of knowledge throughout the community (that is, it implies dynamic and independent decision making process).

The research on learning from peers has primarily been developed outside the marketing literature. This line of research focuses on the role of facilitating transfer mechanisms, conduits or agents through which the transfer of knowledge takes place within a company (Benkard 2000, Argote et al. 1990, Argote 1999, Levitt and March 1988, Olivera et al. 2008, Darr et al. 1995). The transfer mechanism that is particularly relevant in the present study is interactions with peers. Several studies have found that knowledge is shared through peer interactions (Singh et al. 2010, Ingram and Simons 2002). In general, the literature suggests that the increased use of transfer mechanisms is associated with increased levels of knowledge transfers. While the findings in the organizational-behavior literature provide a theoretical background for the formulation of our model, our study advances these findings in the following ways. First, without directly observing the knowledge exchanges among peers, the organizational-learning literature measures learning from peers through a positive impact on productivity when individuals work together over time. By comparison, our dataset allows us to directly observe the knowledge exchanges among peers. Hence, we can directly capture knowledge sharing among peers. Second, the extant organizational-learning literature assumes that knowledge exchange among employees is an exogenous factor. By comparison, we treat the knowledge seeking and sharing decisions as endogenous. Furthermore, we emphasize the role of the network position of the individual and his/her peers in determining the formation of knowledge-sharing relationships.

Methodologically, our research is related to the emerging literature on dynamic-competition games. Many studies have developed models to incorporate strategic interactions among forward-looking actors in various contexts: firm entry/exit (Bajari et al. 2007; Weintraub et al. 2008, Aguirregabiria and Mira 2007), product repositioning in a differentiated product market (Sweeting 2007), technology adoption (Ryan and Tucker 2008), product adoption (Kumar et al. 2010) and etc. In this paper, we apply this framework to the context of an online social media and illustrate how individuals take into account their peers' decisions on a public learning platform.

3. Model Specification

3.1 Industry Background

Online discussion forums have been widely adopted to support peer learning among employees and customers of companies, and to enhance customer services. In this study, we focus on an internal discussion forum adopted by a firm to support peer learning among customer-support staff. In such practices, online discussion forums are commonly integrated as a major component of employee working environment, which facilitates employees to access to the forum while they are working with clients. By embedding internal discussion forums into employee work processes, every employee is guided to seek and share knowledge with others within the firm. Employee company email address is automatically assigned as her identifier (or user name) on the forum. And employee profile (containing employee basic information) is displayed along with the user name whenever the user logs on. By clicking on the user name, other users can find her personal information and all the history of asking and answering questions. When a user has a question to ask, she can post the question to the public forum. Anybody on the forum, regardless of location, can choose to answer the question. When multiple answers are provided (often by different people), they are listed as a queue sequenced according to time. As more and more forums adopt various types of feedback and reward mechanisms, knowledge seekers and sharers are encouraged to provide high-quality questions and answers. In certain cases where a user's true identity is publicly available, revealing the true identity of both the knowledge seeker and the knowledge sharer acts as a guarantee of quality because the users are likely to provide answers in a professional manner.

This type of online discussion forum is also widely adopted for customer support. On this type of customer support forums, registration is also open to all customers of the company, and user profile is also available most of the time. Customers can post questions about the products that they bought, and other customers and employees are encouraged to answer these questions. As we can see, user decisions in customer support forum is very similar to the context in our research, and the "learning from peers" and "knowledge spillover" are presented in customer support forums as well.

In this research, we focus on the basic features of a discussion forum and examine the fundamental drivers of asking and answering decisions by users. Many new features, such as ratings for questions and users, the number of viewings, and virtual rewards, have been gradually introduced to social media platforms, especially on discussion forums that are open to customers. However, as a first study, we focus on the fundamental features that are common in almost all types of discussion forums. We leave the examination of how these features can be incorporated into the model and how they modify the main findings for future research.

3.2 Decision Variables

Assume that there are a total of *N* individuals who have the option to participate in a public forum. During each of the time periods $t \in \{1,2,\dots,T\}$, every individual $i \in \{1,2,\dots,N\}$ can make two types of decisions: to ask a question and to answer a question. For the decision to ask a question, the individual first decides whether to ask a question in a period; if the answer is "yes", she must then decide whether to ask an easy question or a hard question. More specifically,

$$a_{it} = \begin{cases} 2, & if\ individual\ i\ asks\ a\ difficult\ question\ at\ time\ t \\ 1, & if\ individual\ i\ asks\ an\ easy\ question\ at\ time\ t, \\ 0, & otherwise. \end{cases} \tag{1}$$

We allow users to ask questions with different difficulty levels to model the possible differential effects of asking an easy versus asking a hard question on knowledge acquisition and reputation building. While questions could be traced back to the interaction between customers and employees, whether employee need to ask this question (indicating she is able to handle this question herself), or will ask this question (indicating she expects benefit from asking the question) is endogenously determined by her states as well as peers' states. As a result, not all questions arise from customer-employee interaction will be asked on the forum.

We use the dummy variable s_{ijt} to denote the binary decision of an individual i deciding to answer a question posted by individual j at time t (the difficulty level of the question is assumed to be known).

$$s_{ijt} = \begin{cases} 1, & if\ i\ answers\ a\ question\ from\ j\ at\ time\ t \\ 0, & otherwise \end{cases} \tag{2}$$

As we will see in the following section, while we do not explicitly distinguish answers to hard questions from answers to easy questions, this difference is inherent in the type of questions these answers correspond to. Notice that s_{ijt} is both *i* and *j* specific. This specificity means that we consider the source of the question and allow the user to decide whose question to answer. Recognizing the dyadic nature of answering decisions permits us to endogenize the formation of the network and to investigate its fundamental drivers. Users may answer multiple questions during the same period. We use the vector $\boldsymbol{s_{it}}$ to represent the set of answering decisions for individual i at time *t*. Note that when an individual chooses not to ask or answer questions and instead stay as an observer on the forum, our model treats it as a choice not to ask or answer, and a_{it} and s_{ijt} remain zero.

3.3 Per-Period Utility Function

We assume that the utility function of an individual i at time t is affected by her knowledge, her social status (which indicates how active she is as a community contributor), and the cost of asking and answering questions (Levitt and March 1988, Darr et al. 1995, Singh et al. 2010, Lakhani and Von Hippel 2003). To be more specific, this per-period utility function can be written as

$$(3) \qquad U_{it}(\boldsymbol{K}, \boldsymbol{R}, \boldsymbol{X_i}, a_{it}, \boldsymbol{s_{it}}, \boldsymbol{\varepsilon_{it}}) = \alpha_1 K_{it} + \alpha_2 R_{it} - C(a_{it}, \boldsymbol{s_{it}}, \boldsymbol{X_i}) + \varepsilon_{it}(a_{it}, \boldsymbol{s_{it}}),$$

where K_{it} is the knowledge level accumulated by individual i at time t. Individuals on the forum obtain utility from their accumulated incremental knowledge levels because a higher knowledge level is associated with higher productivity levels, which can indirectly lead to monetary incentives or more free time for other activities. According to the existing findings in the organizational-behavior literature, the knowledge gained through interactions with peers increases productivity and job performance for three reasons (Singh et al. 2010, Reagans et al. 2005, Argote et al. 2003). First, these interactions allow opportunities for resource pooling and sharing alternative interpretations of problems (Reagans et al. 2005). Second, these interactions help in coordinating the effort, which may minimize effort duplication (Singh et al. 2010). Third, interactions with peers provide opportunities for an individual to apply her efforts or knowledge to different but related problem domains (in which her peers may be having problems) and, in the process, to develop a deeper cognitive understanding of her field (Schilling et al. 2003).

R_{it} is the social status level for individual i at time t. Individuals may derive psychological or economic utility from building social status within a community (Lakhani and Von Hippel 2003) because higher social status brings social recognition and increases value to the community (Kilduff and Krackhardt 1994). Implicitly or explicitly, many companies use the participation levels on social media platforms to identify experts in different areas (McAfee 2006). Being identified as an expert within the community provides indirect incentives, such as job opportunities, salary increases, promotions, etc.

Finally, the individual incurs a cost from asking and answering questions. When she asks a question, she needs to invest time in posting the question on the forum and in carefully phrasing it so that people in the community can correctly understand it. When she answers a question, she needs to first think about the answer and then express it on the forum in an organized and clear manner. Both of these two processes are time consuming. Let $\boldsymbol{X_i}$ denote $i's$ individual characteristics, such as gender, that affect the cost of asking and answering questions. This factor accounts for potential heterogeneity in costs across individuals. Let ε_{it} denote the private shock that is only observable by the individual in question. We assume that ε_{it} has a type-I extreme-value distribution and that private shocks are *iid* across participants and periods.

Both knowledge and social status are endogenously determined by a participant's decisions on whether to ask and answer questions in the current period and by everyone else in the community. As we stated in *Industry Background* section, these two state variables are public information for all individuals on the forum, as an individual's complete history of asking and answering questions are revealed by clicking on her profile. We will describe how these two variables are updated according to individual decisions in the following sections.

3.3.1 Knowledge Updates

We use the term *knowledge* to represent an individual's expertise in a professional discipline, which is usually specific to a working project, rather than a measure of an individual's overall knowledge level[14]. We use K_{it} to denote the knowledge level of individual i that at the beginning of period t. It can be updated according to the following process:

$$(4) \qquad K_{i(t+1)} = K_{it} + \sum_{j \in N, j \neq i} k_s^w s_{jit} + \sum_{j,m \in N, j,m \neq i} k_x^w s_{jmt},$$

where $w \in \{E, D\}$ represents the difficulty of the question, easy (E) or hard (D). The dummy variable a_{it} represents the specific question that individual i has asked at time t. Under the assumption that the knowledge level is additive, the term$\sum_{j \in N, j \neq i} k_s^w s_{jit}$ represents the total amount of knowledge from answers provided by all of the other individuals to individual *i*'s question[15]. Similarly,$\sum_{j,m \in N, j,m \neq i} k_x^w s_{jmt}$ represents the total amount of knowledge from answers provided by individuals other than i to questions asked by individuals other than i during time *t* [16].

[14] We assume that it is the knowledge level from this online forum that enters the individual utility functions. Apparently, individuals can obtain knowledge from alternative channels, such as prior education, offline communication, learning by doing, etc., and we acknowledge that we do not have information on offline activities. However, we think it is a reasonable assumption because the participants in the online forum do not know each other in real life. They also don't have significant information about the true knowledge levels of their peers, and the only information source for this knowledge level is the online forum. As a result, when they make decisions based on others' knowledge levels, the decisions are made based on the proportion of knowledge coming from the online forum. Furthermore, in our context, questions are generated when the focal company provides IT consulting service to its client company. Thus, individuals equipped with more knowledge can better solve the problems resulting from this cooperation, improve service quality and consequently improve their performance.

[15] Note that the individuals participating in this setting are problem solvers. The literature in this area states that when multiple solutions are provided to a problem, the individuals learn different ways of solving the problem (Singh et al. 2010). The individuals are likely to use the tricks and tools from these solutions to solve other problems. Hence, multiple answers to a question provide greater knowledge increment than a single answer.

[16] We make three simplifying assumptions in the knowledge updating process. First, we assume that each answer provided to a question increases the knowledge level by same amount. It will be interesting to modify this updating rule in future research by allowing quality weighting of the unit knowledge increment. Second, we assume that the knowledge is additive in the sense that the marginal knowledge increment is independent of the knowledge level. This assumption can be relaxed by allowing the knowledge increment to decrease with the number of answers provided. Third, we assume that everybody reads all the answers posted online. This assumption is realistic, given the small number of questions and answers that are posted on the forum at each period of time in our research setting. In addition, the forum archive acts as knowledge repository where both

The last two terms represent two ways of gaining knowledge on the public forum. First, the learning can be initiated by an individual who posts a question and collects answers from others that directly address her problem. Second, she can still gain knowledge without actively soliciting answers by reading the responses to questions asked by others. Consequently, she can increase her knowledge level even if she is not the one who asks the question.

k_s^w is a coefficient to be estimated that measures the marginal knowledge increment for one additional answer provided for the question asked by the focal user i with question type w, and k_x^w measures the corresponding increment for questions proposed by someone else. Intuitively, we expect that individual *i* can gain more knowledge from reading answers to her own question because those answers are essential and are a better fit to the knowledge she urgently requires. By comparison, while *i* gains knowledge from reading answers to questions asked by *j*, she may already possess that piece of information or that piece of information may not perfectly fit her needs. Therefore, we expect that $k_s^w > k_x^w$ (i.e., "active" learning is more effective than "passive learning). In addition, we allow the marginal knowledge gains to differ between difficult and easy questions and expect the answers to hard questions to yield higher knowledge gains.

As suggested by the terms $\sum_{j\in N, j\neq i} k_s^w s_{jit}\, I\big(s_{jit} \in \{a_{it}\}\big)$ and $\sum_{j\in N, j\neq i} k_x^w s_{jit} I\big(s_{jit} \notin \{a_{it}\}\big)$, user *i*'s decisions are not independent of the decisions of her peers. When deciding whether to ask a question at each period *t*, for example, individual i needs to predict the number of answers that will be provided to her question. With everything else being equal, she will ask a question only when the expected number of answers to her question is large enough and the knowledge increments can justify the cost.

the questions and answers are stored; it is accessible to everyone within the company at all times. We leave it for future research to incorporate variations in browsing behavior.

The knowledge updating rule also implies that a single user's decision alters the knowledge levels of all her peers. Whenever an answer is provided, the knowledge level of everybody in the community is updated according to equation (4). When making decisions on both asking and answering questions, an individual needs to consider that both her own knowledge level and that of her peers in the community will increase as a result of the answers posted to the public forum. The increase in knowledge levels throughout the community may change the asking/answering decisions of her peers in the future.

This observation implies that an individual may expect to be rewarded in the future because of the increases in the knowledge of the whole community. When the whole community becomes more knowledgeable, more answers will be provided to any question asked in the future. Thus, a knowledge seeker can expect higher knowledge increments from questions she asks in the future because her peers are the knowledge sharers. The anticipation of possible future rewards that are reciprocated by her peers can potentially change her current decisions about asking/answering questions, which is especially interesting when she decides whose question to answer; we discuss this issue next.

3.3.2 Online Social Status Updates

The literature has shown that people tend to contribute to a community to build up their social status (Kollock 1999, Lakhani and Von Hippel 2003). Typically, social status in online communities is determined not by possessions but by contributions. This effect leads to a culture where the social status of a user is primarily determined by his/her contributions to the community. This consideration is especially salient in online community settings because contributions are transparent to every member of the community (Lampel and Bhalla 2007). There are various reasons why people may contribute. First, economic incentives may directly motivate participants to contribute when competing to achieve a higher relative social status. Second, it may be attributed to deeply held values. For example, an individual may be motivated to give back to the community to reciprocate support that he/she may have received in the past (Lampel and Bhalla 2007). Third, individuals may expect that in the future others may respond back in kind as a result of their present contributions (Lakhani and Von Hippel 2003).

An individual can build her social status by frequently providing answers to questions posted on the public forum. The higher the frequency is, the greater the perception that she is an active contributor to the community. Moreover, extensive sociology and psychology literature has shown that building social status goes far beyond a simple count of the questions answered. For example, Bonacich (1987) shows that entities wield power or benefit from being central in a community and that entities closely connected with other central participants are considered to have high status. In our context, this finding means that those who answer questions posted by high-status participants obtain a higher perceived social status themselves, because answering questions from central individuals indicates a closer connection with colleagues who are more informative, active, and resourceful. This consideration requires us to explicitly incorporate the dyadic relationship between individuals, which is defined for any pair of individuals by how many answers they provide to each other.

To approximate this thinking process, we adopt *eigenvector centrality*, which measures an individual's position within the discussion forum based on both the intensity and direction of the interactions among all of the users. This approach is the most commonly used measurement of social status in sociology (Bonacich 1987) and recently adopted by Katona and Sarvary (2008), Kumar et al. (2009, 2010) and Ma et al. (2010) in marketing; in which higher centrality indicates a higher frequency of contribution and/or a more centralized position in a network. In this measure, not only are the individuals with high contribution levels relatively central in the network, but the ones who are connected with other central members also have a relatively high status in the network. A_t is defined as an $N \times N$ *adjacency matrix* that measures the intensity of interactions between i and j (in our context, it is the number of questions answered) up to time *t*. It is defined by

$$A_t = A_{1,t} + rA_{2,t}, \tag{5}$$

where the $(i,j)th$ element of $A_{1,t}$ is the number of easy questions asked by *i* that are answered by j up to time t , and the $(i,j)th$ element of $A_{2,t}$ is the number of hard questions from *i* that are answered by j up to time t. r is a coefficient to be estimated, and a value greater than one indicates greater improvement in social status from answering hard questions than from answering easy questions.

Let x_i denote the *eigenvector centrality* of individual i. It is defined as the sum of the interaction intensities between the focal user and her peers weighted by the peers' centrality in the network:

$$x_{ti} = A_{t,1i}x_{t1} + A_{t,2i}x_{t2} + \cdots + A_{t,Ni}x_{tN}. \tag{6}$$

Thus, the eigenvector centrality of *i* depends on the frequency of her answering and the position of the users whose questions she has answered. If individual m has a higher centrality score than individual n (if $x_{tm} > x_{tn}$), then answering a question proposed by m is more beneficial for individual i than answering a question proposed by n ($A_{t,mi} = 1\ and\ A_{t,ni} = 0$ increases x_i by a higher margin compared to the case in which $A_{t,mi} = 0\ and\ A_{t,ni} = 1$). In other words, a user can improve her social status more by answering a question asked by a user who is more centrally located. The calculation of eigenvector centrality captures the commonly observed phenomena that individual status in a network is increased by connecting to others who are themselves well connected. We can rewrite it in matrix form for everyone in the community as

$$x_t = {A_t}^T x_t. \tag{7}$$

However, equation (7) has no non-zero solution unless A_t has an eigenvalue of one. One solution to this problem is to instead use alpha-centrality, which is a commonly applied measurement of asymmetric networks that is widely used in sociology, economics, management and computer-network research (Bonacich and Lloyd 2001).

(8) $$x_t = \alpha {A_t}^T x_t + y_t$$

Here, y_t is an exogenous factor that influences individual status in the network; it is assumed to be a vector of ones in our context. α captures the relative importance for determining social status of individual social position in the online discussion forum compared to exogenous factors[17].

One may argue that individual social status in online discussion forums is not solely determined by her contributions to the online discussion forum; asking hard questions can also improve social status for two reasons. First, hard questions can be inspiring in the sense that they encourage people to think about important issues. Second, asking hard questions may indicate the sophistication of the knowledge seeker because hard questions are usually generated from careful thinking about a problem and require a deep understanding of the topic. Thus we construct social status using equation (9), in which we assign a weight to the social status generated by contributing to the discussion forum:

(9) $$x_t = \alpha {A_t}^T x_t + (y_t + \alpha' y'_t),$$

where y'_t is the number of hard questions that employees ask up to time t and thus captures the improvement in social status achieved by asking hard questions. α' is the weight of asking hard

[17] In estimation, we fix the value of α at $\alpha = 0.01$ to guarantee that the values of all the calculated social status scores are positive. If the value of α is greater than the largest eigenvalue of the adjacency matrix, some of the calculated social statuses will become negative. We tested different value of α, and found that the specific value of α did not influence the conclusions of the paper.

questions when individuals evaluate their peers' social status. Then we can solve the equation to get centrality score:

(10) $$x_t = (I - \alpha {A_t}^T)^{-1}(y_t + \alpha' y_t').$$

However, eigenvector centrality is an absolute measure of social status, while in reality individuals care more about their relative rank in the community. To adapt to this observation, we define the *social status score* in our context by

(11) $$R_{ti} = \frac{x_{ti} - \min(x_{ti})}{\max(x_{ti}) - \min(x_{ti})} + 1.$$

R_{it} can be viewed as a score that summarizes the relative frequency of contribution (with differential rates of improvement for answering easy and hard questions), the centrality of the other users whose questions she has answered, and the number of hard questions she has asked. Appendix 1 provides an example to intuitively explain how the social status score is calculated based on the interaction history among peers.

The relative ranking of social status (equations 5-11) implies that the decisions of all users are interdependent because one user's contribution decision changes her social status and inevitably affects that of her peers. An individual who answers one more questions gains a higher social status, which inevitably decreases the relative ranking of her peers. Asking questions also changes an individual's social score by offering others an opportunity to answer the question. When an individual's question is answered by her peers, their social status improves and her social status score decreases correspondingly. Thus, the focal user's decisions about asking and answering questions change the social-status rankings and hence the decisions of her peers. The changes in social ranking alter the future decisions of all the users in the community.

The competition for social ranking modifies the dynamic and interactive decision making process regarding knowledge sharing. First, when competing for relative social ranking, users are making strategic decisions about asking and answering questions that directly affect the rate at which an individual's knowledge increases. It will be interesting to examine whether the competition for social ranking helps the knowledge accumulation of the entire community. Second, competition for higher social status makes knowledge sharers more selective in deciding whose question to answer. If greater centrality-score improvements can be obtained by answering a question from a more centralized user compared to a less centralized user, it is likely that questions from high social status members are more likely to be answered in general. Endogenizing network formation allows us to examine whether this helps or hurts knowledge accumulation.

Note that the social status updating process is different from the knowledge updating process for the following four reasons. First, users improve their social status scores mainly by answering questions. However, they improve their knowledge mainly by asking questions and reading the answers posted by others. Their knowledge cannot be improved without contributions from their peers. Second, the knowledge updating rule depends on the questions that are answered on the forum but not on who answered them. However, the social status updating rule accounts for the dyadic relationships. That is, the social status updating rule considers whose question an individual answered, and who answers this question. Third, an individual's social status can increase or decrease depending on the action taken by everybody in the community, while the knowledge can only go up. Fourth, while both knowledge and social status imply that individuals are interdependent, the mechanisms are different. Individuals care about the absolute knowledge level and do not need to compete for knowledge. However, individuals do compete for a higher rank in the sense that their own social status increase at the cost of the social status of others. As a result, users with higher social status may not necessarily be the ones with more knowledge. For example, a user who has been actively seeking knowledge by asking lots of questions without answering questions can have a very low social status but a high knowledge level. Similarly, a user who only answers questions can build up high social status but not necessarily accumulate a high level of knowledge.

3.4 Costs of Asking and Answering Questions

While knowledge and social status are influenced by the decisions of whether to ask and answer questions, there are also costs associated with each one of these decisions. For asking and answering questions, an individual needs to invest time and effort to clearly frame and explain her problems and answers to others. So we can write the costs for each time period as

$$C(a_{it}, \boldsymbol{s_{it}}) = C_a(a_{it}) + C_s(\boldsymbol{s_{it}}), \tag{12}$$

where $C_a(a_{it})$ is the cost function of asking question and $C_s(\boldsymbol{s_{it}})$ is the cost function of answering questions. We assume that the cost may depend on the knowledge seeker's personal characteristics, such as organizational position and gender. For example, a higher position may be associated with more experience in the subject area and therefore different costs for asking and answering questions. Males may have lower costs for asking questions because they are more aggressive, a characteristic which is likely to be applicable to online forums as well. To be consistent, we assume that the cost of asking/answering questions can be written as a linear function of the two observed user characteristics:

$$C_a(a_{it}) = \sum_w I(a_{it} \in w) * \left(c_{a,0}^w + c_{a,1}Gender + c_{a,2}Position\right), \tag{13}$$

$$C_s(\boldsymbol{s_{it}}) = \sum_{j \in N} I\left(a_{jt} \in w\right) * s_{ijt} * \left(c_{s,0}^w + c_{s,1}Gender + c_{s,2}Position\right). \tag{14}$$

The coefficients $c_{a,1}$ $c_{a,2}$ $c_{s,1}$ and $c_{s,2}$ measure how the costs are modified by gender and organizational position, and $c_{a,0}^w$ and $c_{s,0}^w$ are two constant terms that are allowed to be different across question type $w \in \{E, D\}$. Here, we cannot distinguish altruism from the cost of answering questions in this model. Whenever individuals contribute to the forum, they get certain psychological benefit from this altruistic behavior, as well as incurring a cost. Because these two effects happen at the same time, we can't disentangle them.

3.5 Users' Dynamic Problems, Intertemporal Tradeoffs and Estimation

As we discussed earlier, knowledge seeking cannot be achieved without the contributions of peers. Even though a user cannot control the decisions of her peers, she can always make her own knowledge seeking and sharing decisions and thereby influence the knowledge level and network position of others, which alters the future decisions of everybody in the community. Such behavior inherently requires individuals to be forward looking so that they are willing to incur the costs of asking questions and sharing knowledge now to gain reciprocal knowledge increments from their peers in the future. This assumption is consistent with the descriptive results from prior research on social media that show participants of a social media platform tend to help solve the problems of others to gain a higher social status, which helps them get help from the community in the future (Lakhani and von Hippel 2003). Accordingly, we assume that each individual is forward looking and maximizes her long term utility:

$$E[\sum_{\tau=t}^{T} \gamma^{\tau-t} U_i(\tau)|\boldsymbol{S_t}], \tag{15}$$

where γ is the discount factor indicating how much the individual values future utility. In this model setup, the state at time period t, denoted as $\boldsymbol{S_t}$, is the collection of individual cumulative knowledge levels ($\boldsymbol{K_t}$) and social status levels of the individual and her peers ($\boldsymbol{R_t}$): $\boldsymbol{S_t} = (\boldsymbol{S_{1,t}}, \dots, \boldsymbol{S_{N,t}})$, where $\boldsymbol{S_{i,t}} = \{K_{i,t}, R_{i,t}\}$. Both state variables are endogenously determined by user decisions. Individuals make decisions to maximize their discounted lifetime utility based on the information available to them at time t about their own knowledge and social status and that of their peers. Realizing that their states and decisions are interdependent, all the users can anticipate the possible responses from their peers when making decisions about asking and answering questions that maximize their own long-term utility.

The proposed model implies several inter-temporal tradeoffs under equilibrium conditions. First, when asking a question, an individual incurs a cost from framing the question and sacrificing her social status. However, she benefits from reading the answers to her question. Meanwhile, her peers read the posted answers and thus improve their knowledge. The knowledge increments of her peers imply that more answers will be provided to her questions in the future. Thus, she is more willing to ask the question when the anticipated future benefits reciprocated by her peers dominate the current utility she has to sacrifice. Similarly, when answering a question, a user incurs a cost of writing the answer. Even though she can build up her relative social status, the direct impact is to increase the knowledge level of all her peers. As before, a higher community knowledge level implies more answers provided to her future questions. Because the impacts of state changes persist into the future throughout the community, an individual's utility in the future improves and compensates for the costs incurred in the current period. In both asking and answering decisions, therefore, an individual sacrifices short-term utility for long-term knowledge gains from the contribution of her peers.

Traditionally, a dynamic game model is estimated by explicitly solving for equilibrium (e.g., Pakes and Mcguire 1994). However, the curse of dimensionality is one of the obstacles to estimating our model due to the high dimensionality of the state space. To circumvent the computational burden of iteratively approaching the equilibrium strategy, we estimate our dynamic game model using the two-step approach specified in Bajari et al. (2007)[18].This two-step approach also helps

[18] While there is an emerging literature that takes into account unobserved heterogeneity (Arcidiacono and Miller 2010), their methods require either strict assumptions on the state space or a large number of observations. While this is a caveat of our research due to the constraints on data and methodology, we believe that unobserved heterogeneity is unlikely to change our main results.

circumvent multiple equilibria concerns because we empirically recover policy function in first stage of the estimation instead of solving the for equilibria. Furthermore, because employees make decisions within a single community, the observations are generated from a single equilibrium (Ryan and Tucker 2008); thus, the second stage parameter estimations are consistent. Following Ericson and Pakes (1995), we focus on the pure strategy Markov perfect equilibrium (MPE), in which every individual is assumed to rely on the state variables and unobservables of the current period and to adopt an equilibrium strategy that maximizes her lifetime utility. The individual also expects her peers to use the publicly known equilibrium strategy based on the observable states and their private information to make their own decisions. Hereafter, we use $\boldsymbol{\sigma_i}$ to denote individual $\boldsymbol{i}$'s decisions to ask and answering different types of question as a function of the state variables and the private shock: $\boldsymbol{\sigma_i}\colon \boldsymbol{S} \times \boldsymbol{\varepsilon_i} \longmapsto \boldsymbol{A_i},$ where $\boldsymbol{A}_i$ is the set of all actions individual $\boldsymbol{i}$ can take. Then a strategy profile $\boldsymbol{\sigma}^* = \{\boldsymbol{\sigma_1^*}, \dots, \boldsymbol{\sigma_N^*}\}$ is a Markov-perfect strategy solution to a MPE if there is no incentive to deviate from this strategy, given others' fixed strategies:

$$V_i(\boldsymbol{S}|\boldsymbol{\sigma_i^*}, \boldsymbol{\sigma_{-i}^*}) \geq V_i(\boldsymbol{S}|\boldsymbol{\sigma_i'}, \boldsymbol{\sigma_{-i}^*}), \forall i, \boldsymbol{S}, \boldsymbol{\sigma_i'}\,. \tag{16}$$

For example, one may argue that the control of the (unobserved) dyad-level heterogeneity is needed. Individuals may know each other offline, which may increase the probability that they will answer each others' question in the online forum. However, if that is true, we should observe the formation of many cohorts in the online social network because individuals who have offline connections communicate more often with each other rather than with other people on the forum. This prediction is contrary to the core/periphery structure that we observed in the real network structure, in which core individuals communicates with everyone else in the forum, and peripheral members didn't communicate with other peripheral members. As a result, we expect the influence of offline connections to be limited. However, to illustrate the robustness of our model to unobserved offline connections, we incorporates information about employee location and their department in the initial stage to partially take into account the possible influence on knowledge sharing decisions when the knowledge sharer and knowledge seeker are closely located physically. We also incorporate observed individual characteristics, such as gender and age, in the cost function to control for individual characteristics having an impact on the decisions that they are making.

In the first stage, we empirically recover the participant's equilibrium strategies from the observed individual decisions and states[19]. However, Bajari et al. (2007) require the state space to be discrete, while our focal state variables, individual knowledge and social status are continuous. As a result, we adapt the method of Bajari et al. (2007) to allow for continuous state variables (Bajari et al. 2008) by using a "sieve logit". We construct a series of basis functions that can approximate individual decision rules and regress a logit/ordered logit model using these basis functions. In the second stage, we simulate the individual value functions under different policy rules and estimate the structural parameters by comparing the value functions from the first-stage decision rule with those from a perturbed policy. As stated previously, individuals maximize their discounted lifetime utility:

(17) $$\max_{(a_{i\tau},s_{i\tau})_{\tau=t}^{T}} E[\textstyle\sum_{\tau=t}^{T} \gamma^{\tau-t}\, U_i(\boldsymbol{S_t}, a_{i\tau}, \boldsymbol{s_{i\tau}}, \boldsymbol{\varepsilon_{it}})|\boldsymbol{S_t}].$$

We can rewrite this lifetime utility maximization problem as a Bellman equation in which the value function is calculated in terms of the payoff in current period and the value of the remaining decision problems given the initial state:

(18)

$$V_i(\boldsymbol{S}, \boldsymbol{\sigma}^*; \theta) = \max_{\boldsymbol{\sigma_i(K,R)}} U_i(\boldsymbol{S}, \boldsymbol{\sigma_i}\ (\boldsymbol{S}), \boldsymbol{\sigma}^*_{\boldsymbol{-i}}) + \gamma \int V_i(\boldsymbol{S'}|\boldsymbol{S}, \boldsymbol{\sigma}^*_{\boldsymbol{-i}}) dP\big(\boldsymbol{S'}\big|\boldsymbol{S}, \boldsymbol{\sigma}^*_{\boldsymbol{-i}}, \boldsymbol{\sigma_i}\ \big),$$

where $\boldsymbol{\sigma_i^*}$ is individual $\boldsymbol{i}$'s Markov-perfect strategy for the decisions to both ask and answer questions, as has been stated previously. The vectorθ is the aggregation of all the parameters: $\theta = \{\alpha_1, \alpha_2, c_{a,1}, c_{a,2}, c_{s,1}, c_{s,2}, c_{a,0}^{w}, c_{s,0}^{w}\}$. Notice that the state-transition process is deterministic in

[19] As a robustness test for the length of training period, we also allow for training period with different length to initialize individual states in the first stage estimation: $T_1 = 0$, $T_1 = 20$ and $T_1 = 40$. In each case, we initialize individual state variables using data from the first T_1 periods, and estimate individual decision rules using data from remaining periods. Results of the three cases are very similar. Here, we only report the estimation results for the case where $T_1 = 0$.

our setting; the integration is performed over unobservable terms. Details concerning identification and estimation are provided in Appendix 2 of this paper.

4. Empirical Result

4.1 Data Description

The global IT service and consulting corporation in our research advises their clients on how to optimize IT systems to meet their objectives, as well as designs, implements and administers IT systems. Over the years, the company has rapidly expanded to 168 cities in eight nations and more than 80 thousand employees. During this rapid growth, little attention was paid to channelizing knowledge flow between locations. As with many other companies, location-based knowledge silos emerged in the firm, with little flow from one silo to another. Taking advantage of the recent advances in social media technology, the firm integrated an online discussion forum as a major component of employee working environment, and embedded the online discussion forum into employee working process. This design grants employees direct access to the knowledge sharing platform at all times. This platform is mainly used internally, and managers actively monitor activities on the forum. Almost all of the posts are client-related technical questions that are specific to certain areas (for example, "*How to upload an email (design an email system) having attachments in java for clients?*").[20] Our discussions with the top management revealed that they use the forum to identify experts in any area within the firm. Consequently, individuals who make numerous contributions on the forum have a higher probability of receiving a bonus or promotion at the annual salary and promotion evaluation. The firm also found a strong correlation between the participation level of a user on this

[20] On most Web 2.0 applications used within a company (usually called enterprise 2.0), the participants' true identities are revealed, and managers are actively monitoring individual performances on the forum. As a result, the participants usually remain highly professional by asking relevant questions and providing thoughtful answers. In our data, almost all the questions are specific to the IT services that this firm provides to its client. Answers to these questions cannot be easily found outside the company. Thus, when the employees face problems in their work, their first choice is to seek help from their colleagues.

forum and her speed of resolving customer problems. Overall, the firm found that customer satisfaction has increased and customer service costs have decreased since the adoption of this forum.

The data contains detailed information about all of the activities related to asking and answering questions by 2954 employees over 73 weeks (511 days) between April 2006 and August 2007. As stated, this online discussion forum was integrated as a main component of employee working environment, thus all employees are utilizing this forum from the date when it was implemented. During the observation period, a total of 19948 questions were asked, and 58089 answers were provided to these questions. On average, 39 questions were asked and 113.7 answers were provided every day. Furthermore, 11.1 users posted questions and 18.3 users provided answers every day. The majority of the answers (76.1%) were posted the same day as the corresponding question was asked. These posts can be divided into 127 subject categories, such as ".Net Framework", "J2ee" and "Development." These sub-communities are relatively isolated. Few individuals are involved in more than one sub-community; thus, there is not much overlap across the sub-communities from the participants' perspective. Furthermore, the structure of this online forum is designed in such a way that individuals who are browsing one category of the posts are not able to simultaneously observe posts in other categories. This restriction effectively prevents individuals from making decisions based on information from other sub-communities and from simultaneously making decisions in multiple sub-communities. This sub-community isolation allows us to treat each sub-community as a separate network and to focus on a single community without worrying about spillovers from other sub-communities.

Table 2. Data Description

	.Net Framework
Total Number of Questions	652
Total Number of Answers	1676
Number of Participants	329

Percentage of Employees Ever Asking Questions	44.07%
Average Number of Questions Asked per Employee	1.9818
Average Number of Questions Asked per Week	13.04
Percentage of Employees Ever Answering Questions	83.59%
Average Number of Questions Answered per Employee	5.09
Average Number of Answers Provided per Week	33.52
Mean of Gender	0.7428
Standard Deviation of Gender	0.4378
Mean of Position	0.4534
Standard Deviation of Position	0.7933
Mean of Tenure	0.8294
Standard Deviation of Tenure	1.3477

Our calibration sample focuses on the subcategory "*Net Framework*", which is a representative category in the sense that it is one of the major programming platforms in industry. We select individuals who have asked or answered at least one question on the .*Net Framework* sub-forum, and track back their history until the first day when forum was established. We end up with 329 individuals in our dataset. There were 652 questions posted by 145 users and 1676 corresponding answers posted by 275 users. Table 2 provides some sample statistics from the calibration sample. In the *Net Framework* sub-community, 44.07% of the users asked at least one question, and 83.59% answered at least one question. Each individual asked an average of approximately two questions during the study period, and 13 questions were asked every period.

4.2 Estimation Result

Table 3. Parameters Estimates

Variable	Coefficient
Knowledge Updating Rule	
Knowledge increments from own easy question (k_s^o)	0.5401***
Knowledge increments from others' easy question (k_x^o)	0.0036***
Knowledge increments from own hard question (k_s^D)	1.1703***
Knowledge increments from others' hard question (k_x^D)	0.0036***
Reputation Updating Rule	
Reputation increments from asking hard question (α')	-0.0201
Contribution level increments from answering hard question (r)	6.8783***
Impact of individual department on social status (α_0)	0.0027*
Impact from Knowledge (α_1)	0.2805***

Impact from Social Status(α_2)	3.6399***
Cost of asking a question	
Constant for asking an easy question	5.0030***
Constant for asking a hard question	8.8917***
Position	-0.0256*
Gender	-0.8052***
Cost of answering question	
Constant for answering an easy question	7.5703***
Constant for answering a hard question	12.5224***
Position	-0.1411***
Gender	-0.5539***

*** The 99% confidence interval does not include zero.
** The 95% confidence interval does not include zero.
*The 90% confidence interval does not include zero.

The estimation result is presented in Table 3. We fixed the discount factor $\gamma = 0.95$ in the estimation. One of the observations is that $k_x^O < k_s^O$ and $k_x^D < k_s^D$. This result is consistent with our hypothesis that an individual obtains much more knowledge from reading answers to her own question than from reading answers to others' questions. As expected, the knowledge increment is much higher from hard questions than from easy questions (k_s^D > k_s^O). However, the knowledge increment does not differ significantly across question types ($k_x^D \approx k_x^O$).

Relative social status plays an important role in driving user knowledge seeking and sharing decisions as well. While asking hard questions does not necessarily increase a knowledge seeker's online social status (α' is insignificant), providing answers to hard questions will help improve the knowledge sharer's social status. In fact, the effect of answering a hard question on an individual's social status is almost the same as that of answering seven easy questions ($r = 6.8783$).

As expected, both knowledge and social status increase a user's utility, which confirms both major findings of previous studies (e.g., Argote et al. 2003, Reagans et al. 2005, Kilduff and Krackhardt 1994, and Lakhani and von Hippel 2003) and the firm's observation that the speed at which customer problems are solved and the level of customer satisfaction increase with employee participation. Meanwhile, as we expected, the costs of answering questions are higher than those of asking questions. The costs of asking and answering a hard question are higher than those of asking and answering an easy question. The costs also differ across individual characteristics. They are significantly higher for females than for males. Interestingly, the cost of answering questions is lower for users with higher organizational positions. This finding is intuitive because individuals with higher position tend to have more expertise in their field; thus, they are proficient at solving problems. The costs of asking questions are higher for users with higher organizational position, but this effect is barely significant.

It is important to notice that whenever an individual asks or answers a question, the additional utility from the knowledge and/or social status increments in the current period generally cannot compensate for the incurred cost. This sacrificial behavior can be justified when the individuals are allowed to make participation decisions in anticipation of future reciprocal rewards, as we will discuss below.

4.3 Dynamic and Interdependent Decision Making

We now report the policy functions that represent the equilibrium decision rules resulting from the users' dynamic interactions. We focus on describing how asking and answering decisions are driven by the knowledge of the knowledge seekers and the community (Figures 2A and 2B), whose question to answer (Figure 2C), and whether to ask and answer questions given her own particular level of knowledge and social status (Figures 2D and 2E).

Figure 2. Equilibrium Policy Functions

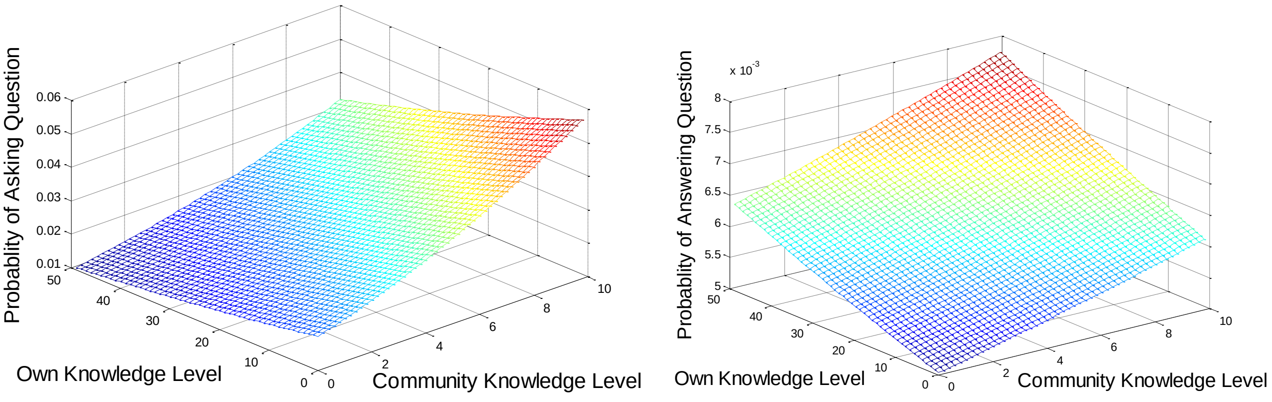


Figure 2A. The Probability of Asking Questions

Figure 2B. The Probability of Answering Questions

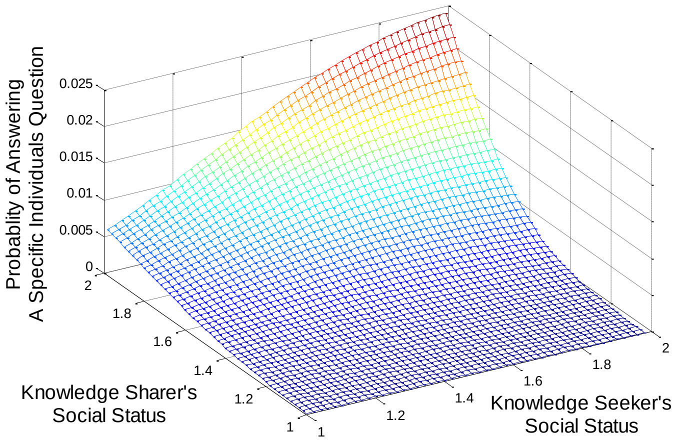
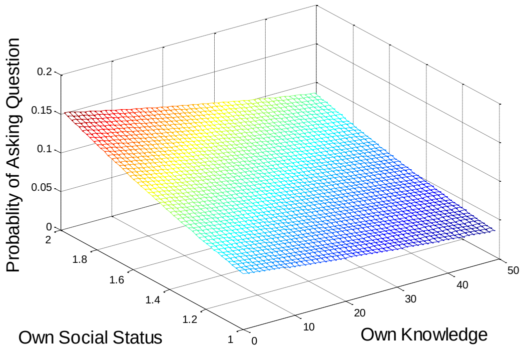


Figure 2C. The Probability of Answering a Specific individual's Question



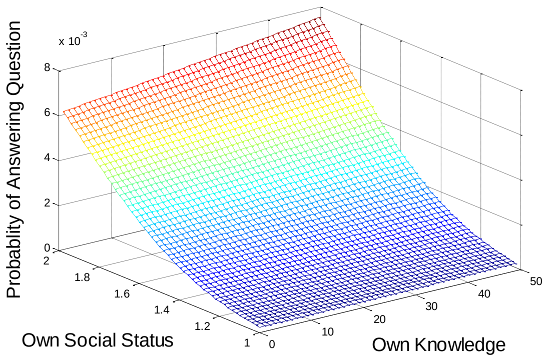


Figure 2D. The Probability of Asking Question Given One's own State

Figure 2E. The Probability of Answering Question Given One's Own State

Reciprocal Rewards Depend on the Knowledge of Peers

Figures 2A and 2B show how the probabilities of asking questions and answering questions are driven by the individual knowledge levels and the mean knowledge levels of the peers. This value is obtained by averaging other state variables for each individual. We list a few interesting findings. First, and not surprisingly, the probability of asking a question decreases and the probability of answering a question increases with the focal user's knowledge level. Second, the probability of asking a question increases with the mean peer knowledge level. This finding can be explained by the dynamic, interdependent decision process: when seeking knowledge on a public forum with higher knowledge, an individual expects a higher probability for her question to be answered and hence a higher incremental knowledge increase. When the anticipated future reward is high enough to justify the immediate cost of writing the question and the decrease in network position, she will ask the question.

Third, it is surprising to find that the probability of answering a question also increases with mean peer knowledge level. This finding is counter-intuitive and cannot be explained by the conventional altruism view, which suggests that individuals should be more willing to help those with low knowledge levels or should at least be indifferent to their peers' knowledge levels when answering questions. However, it can be explained by our dynamic and interdependent decision process: when the population is more knowledgeable, more answers are likely to be offered to each posted question. As a result, an individual can expect more help from the community when she posts a question in the future. In other words, she expects a higher reciprocal reward when contributing to a more knowledgeable population. It is the greater future reciprocal reward from the community that motivates her to prefer contributing to a more knowledgeable audience.

These results shed some light on the incentives for individual contribution to the community from a dynamic perspective. While previous literature on incentive of individual contribution focuses more on static reasons such as altruism, we show that there is another layer of incentives involving the dynamic interaction among all the users and the future payoffs reciprocated by the community. This observation is consistent with the concept of "reciprocal altruism" that is established in the social psychology literature and recently in marketing literature (e.g. Kumar 2010 and Ma 2010). Trivers (1971) states that "altruism, defined as an act of helping someone else although incurring some cost for this act, could have evolved since it might be beneficial to incur this cost if there is a chance of being in a reverse situation where the person whom I helped before may perform an altruistic act towards me."[21]

Whose Questions to Answer

Figure 2C shows how the probability of answering a question is driven by the social status of knowledge seeker and sharer. It can be observed that, in general, the higher the social status of the knowledge seeker, the more likely her question will be answered. Interestingly, her question is mostly answered by other high social-status users. By contrast, users with lower social-status rakings rarely get help from the community. This effect is observed because answering a question posted by a more-central peer will increase an individual's network position more than answering a question asked by a less centralized peer. In expectation of a higher future reward, an individual chooses to be associated with better connected peers. This observation implies that users are selective when deciding whose questions to answer.

Under the assumption of a Markov Perfect Equilibrium, the above can be viewed as the decision rules followed by an individual' peers and can serve as the basis for the decision maker forming expectations about her peers' reactions. These expectations describe a user's anticipation of whether her question will be answered. We next examine how she decides whether to ask or answer questions based on her own current states.

Whether to Participate

[21] Scientists have documented reciprocal altruistic behavior among vampire bats (Wilkinson 1988). The bats are found to feed each other by regurgitating blood. To qualify for reciprocal altruism, the benefit to the receiver would have to be larger than the cost to the donor. This effect seems to hold because the bats usually die if they do not find a blood meal two nights in a row. Putting the concept into the form of a strategy in a repeated prisoner's dilemma would mean to cooperate unconditionally in the first period and behave cooperatively (altruistically) as long as the other agent does as well. If the chances of meeting another reciprocal altruist are high enough or the game is repeated for a long enough, this form of altruism can evolve within a population.

Figures 2D and 2E demonstrate how the probabilities of posting a question and answering a question are driven by an individual's current knowledge and social status. Consistent with expectation, the general trend is that the lower the knowledge, the higher the probability of asking a question and the lower the probability of answering a question. It can also be observed that the probability of answering questions increases with social status regardless of one's own knowledge level. This effect is observed because individuals need to contribute to the community to maintain their social status. A user with a relatively high social status needs less effort to infiltrate the "core" group, within which her future questions will receive substantially more answers. This consideration gives her extra incentive to contribute to the community. By contrast, individuals with low status know that it will take much more effort to get into the "core" group, and the benefit cannot compensate for their cost of contribution. As a result, they will be reluctant to contribute.

It is interesting to observe that a user is more likely to ask a question when she is more centrally connected. This finding can be explained by her anticipation of future reciprocal rewards from her peers, who follow the decision rules described in Figure 2C: questions posted by individuals with higher network positions tend to attract more answers. Expecting a higher probability of receiving an answer from their peers, the users realize that the future benefits from knowledge improvements dominate the immediate cost of writing the question and of lowered social status. Hence, they are more likely to seek knowledge from the public forum when they are central in the network. When they are not centrally located, however, they expect a lower probability of getting help from their peers and are less likely to decide to seek knowledge from them.

4.4 Results Analysis

Based on the decision rules described in the previous session, we now present some analysis results from the second-stage estimation that explain the observed adoption and knowledge sharing patterns, and we explore how to better encourage users to share their knowledge on social media.

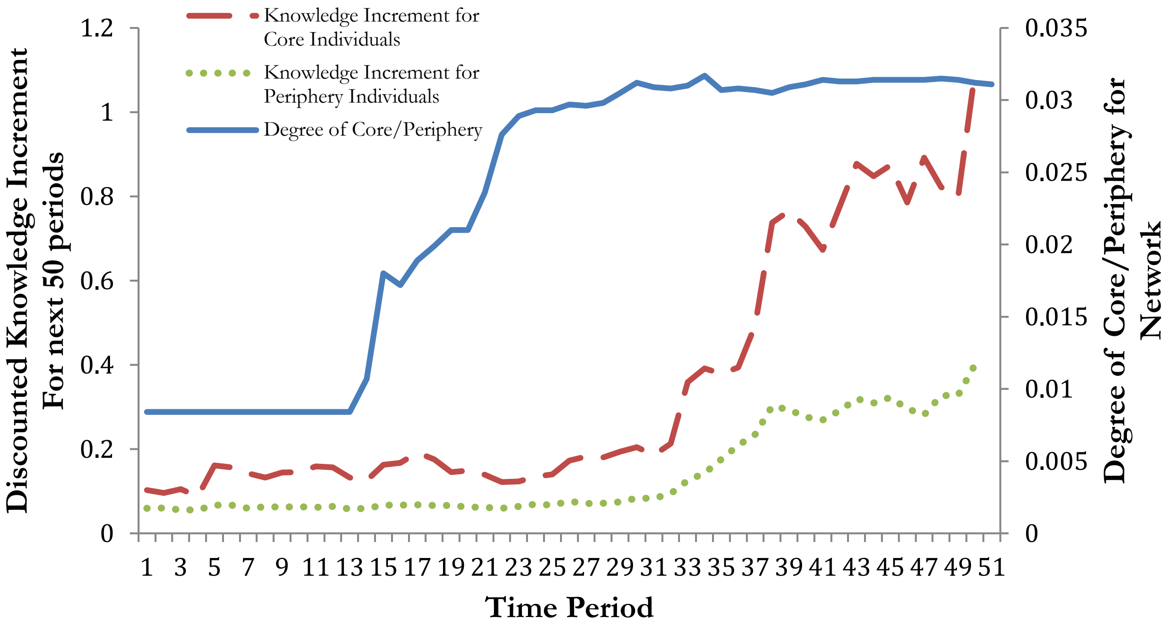
The Formation of the Core/Periphery Structure and the Efficacy of Knowledge Sharing

The dynamics shown in Figure 2C suggest that due to the formation of cohorts in the online social network, everyone tends to answer the questions proposed by high social-status users. However, this effect works differently for users at different positions in the network. A group of users with high social status answer each other's questions while leaving knowledge seekers with low social-network positions in disadvantageous situations. Once a cohort appears, it reinforces itself through the pattern of future interactions among its participants. Over time, this decision process will result in a small inner circle within which the users have the privilege of answering each other's questions, while the questions posted by users outside the circle are likely to be ignored. This effect is similar to the offline silos that are detrimental to knowledge sharing. Thus, even though the adoption of a discussion forum eliminates the location-based knowledge silos, the strategic interaction among users creates another kind of silo that is based on social status.

As a result, it is in the best interest of the peripherally located users not to participate when they anticipate a lower probability of their questions being answered. Instead, they wait for other individuals to ask questions and learn reactively from reading the answers. This situation creates "free riding" behavior in the sense that they learn from reading the answers without asking or answering questions themselves. As a result, most of the activities are generated by users who are in the privileged core group.

Figure 3. The Formation of a Core/Periphery and the Speed of Knowledge Increments

This figure illustrates how the degree of core/periphery structure in the network evolves over time and how the discounted knowledge increments for the next 50 periods of individuals located at different position in the network change over time.



To examine whether the formation of cohorts affects knowledge accumulation, we compare the growth of knowledge for users who are within the cohort and to that of those outside it (Figure 3). The solid line depicts how the degree of core/periphery structure of the online social ne0074work evolved over time in our data (see Appendix 3 for details on the measurement). The dotted line represents the knowledge increments for the individuals who rank in the social-status top 30 at the end of our observation periods (these individuals are almost always in the top 30 across most time periods). The dashed line represents the remaining 301 individuals who are located at the periphery of the network. From this graph, we can see that when the degree of core/periphery structure in the network becomes salient, the rate of knowledge increments is faster among the users who are within the privileged core group and is much slower among the rest. In other words, the core individuals benefit more from the community, and the peripheral individuals benefit less. These results imply that the endogenously formed cohort impedes effective knowledge sharing within it. In practice, individuals who are of low social status are likely to be the newcomers who need more help. However, they are much less likely to receive help from the community.

Proactive Learning versus Reactive Learning

To investigate the differential effects of asking and answering questions on knowledge increments, we compare the knowledge increments when a user asks one question at time *t* to those from answering one question at time *t*. For both actions, we also compare the knowledge increments for the focal user and the whole community. To be more specific, we compare the expected knowledge increments for all the subsequent periods of the two alternatives (asking versus not asking a question and answering versus not answering a question) for the focal individual and for the whole community. Given the expected knowledge increments from each individual's choice over every time period, we obtain the average knowledge increments from asking an additional question (Table 4A) and answering an additional question (Table 4B). Our measurement considers the decisions of the entire community due to the increase in the overall knowledge level and the changes in relative social status.

Table 4A. A Decomposition of the Knowledge Increments from Answering a Question

Period	% Change of Asking Questions	% Change of Answer Questions	% Knowledge Increment of User i	% Knowledge Increment of Community
$t = 1$	0.0037%	0.0313%	0%	0.5377%
$t = 2$	0.0035%	0.0296%	0.3689%	0.0054%
$t = 3$	0.0032%	0.0278%	0.3584%	0.0049%
$t = 4$	0.0031%	0.0270%	0.3289%	0.0047%
….	…..	……	……	…...
Cumulative[a]	0.0722%	0.6025%	7.308%	0.6407%

Table 4B. A Decomposition of the Knowledge Increments from Asking a Question

Period	% Change of Asking Question	% Change of Answer Question	% Knowledge Increment of User i	% Knowledge Increment of Community
$t = 1$	0.0019%	0.0244%	11.02%	0.2579%
$t = 2$	0.0019%	0.0235%	0.0461%	0.0039%
$t = 3$	0.0018%	0.0218%	0.0444%	0.0037%
$t = 4$	0.0016%	0.0210%	0.0409%	0.0035%
….	…..	……	……	…..
Cumulative	0.0351%	0.4788%	11.93%	0.3347%

When a user answers a question at time *t*, her own knowledge does not increase, while that of her peers increases by 0.5377%. In the next period, the improved knowledge level and competition for social reputation make everybody in the community alter their probabilities of asking and answering questions, which leads to a total knowledge increment of 0.3689% for the focal user and of 0.0054% for the entire community. Then, the dynamic process continues. At the end of our observation period, the focal user improves her knowledge by a total of 7.308% and the entire community improves by 0.6407%.

Similarly, when a user asks a question, she obtains an 11.02% knowledge increment from the answers provided directly to her question. During the same period, all of the answers to her question will be read by the entire community, which results in a 0.2579% increase in knowledge for the community. In the second period, the increased knowledge level of everybody in the community allows them to provide more answers to the questions raised. When everybody reads these additional answers, the community knowledge level further increases. There is a 0.0461% knowledge increment for the focal user and an average 0.0039% knowledge increment for all of her peers. Then, this process continues. At the end of our observation period, the knowledge increment of the focal user is 11.93% and that of the whole community is 0.3347%. Again, the focal user benefits more than her peers in the long run.

It is interesting to note that the focal user benefits more from asking a question than answering a question (11.93% versus 7.308%). This effect occurs because (as was discussed previously) the focal user, being an active knowledge seeker, benefits the most from reading the answers that are provided to her own questions. It is even more interesting to note that for both sharing and seeking knowledge, the focal individual benefits significantly more than the community. This finding further confirms that users anticipate future reciprocal rewards from the community when making asking and answering decisions.

Breaking the Cohort: A Sensitivity Exercise

Based on our understanding of the fundamental drivers of knowledge seeking and sharing decisions and the formation of the network, we next conduct a sensitivity analysis that requires the knowledge seekers to hide their identities while allowing the knowledge sharers to build their reputations.[22] All of the users are still motivated to contribute and compete for their reputation. Without knowing the source of the questions, however, the knowledge sharers cannot selectively answer the questions asked by the central users. In other words, we do not allow the existing social structure to influence users' decision on whose question to answer.

More specifically, we assume that the change occurs at the end of our observation window. The existing social status and knowledge levels are preserved. After this time, however, whenever an individual asks a question, they are forced to post it anonymously. In this setting, individuals still gain social status from answering questions. Without knowing the source of the questions, however, they will not strategically answer questions to increase the probability that their questions will be answered in the future. Due to the large number of individuals in our dataset, the sensitivity exercise is conducted based on the notion of oblivious equilibrium introduced by Weintraub et al. (2008).[23] The algorithm developed in Weintraub et al. (2009) is employed to calculate the value functions and decision rules. We simulate individual behavior for the subsequent 50 periods, starting from the last period in our dataset.

Table 5. Hiding the Identity of the Knowledge Seeker: A Sensitivity Analysis

	Probability of Asking Questions			Probability of Answering Questions			Degree of Core/ Periphery	Mean Community Knowledge		
	Core[a]	Periphery	Total	Core	Periphery	Total		Core	Periphery	Total
Benchmark	0.0675	0.0291	0.0326	0.1832	0.1229	0.1284	0.0501	9.0895	3.5119	4.0205

[22] Here, we acknowledge the limitations of the BBL framework for conducting policy simulation. Restrictive assumptions are required for the decision rule to avoid dealing with multiple equilibriums and to guarantee that the solved equilibrium under new policy is not off the chart. Given the limitations of counter-factual analysis, we only show one modification to the current design. Other interesting analyses include resetting the individual reputation periodically, encouraging individuals to answer low social-status members question by giving them financial incentives, etc. We will leave these interesting counterfactual analyses for future research.

[23] When deciding on whether to answer the question, individuals do not know who the knowledge seeker is. Thus we do not need to model the interaction between the individuals and the specific network structure. This simplification allows us to use an oblivious equilibrium to solve the equilibrium.

Anonymity	0.0344	0.0399	0.0394	0.1786	0.1461	0.1491	0.0438	8.6272	4.4632	4.8429

a.Here, core individuals are selected as a cohesive group of thirty individuals who are closely communicate with each other, and periphery individuals are the remaining ones who may loosely connected with someone in the core group, but have rare connection with other periphery ones (Borgatti and Everett, 2000).

Table 5 shows the percentage change in the number of questions asked and answered, the percentage change in the community knowledge increments under the alternative design, and compares them with those under the original design. These numbers are also reported separately for users with central and peripheral locations. We observe that with this minor change to the design of the forum, users are more likely both to ask questions (a 20.86% increase) and to answer questions (a 16.12% increase). These increases occur because without knowing the source of the question, users treat all peers and their questions equally. As a result, otherwise low-status users, who consist the majority of the users on the forum, are more likely to obtain help from the public forum, thus they are more likely to seek knowledge as well as contribute to the community..

More importantly, we can see from Table 5 that the degree of core/periphery structure decreases from 0.0501 to 0.0438 (a 12.57% decrease). The total amount of knowledge accumulated at the end of the observation period increases from 4.0205 to 4.8429 (a 20.46% improvement) on average. This suggests that this slight modification of the existing design encourages knowledge sharing within the community. Thus, breaking the cohort helps encourage participation and increases the amount of knowledge shared within the community, which is the primary goal of adopting social media platform. We can further differentiate between the core and peripheral individuals in terms of the effect of the design change on their knowledge improvement from the last column. The peripheral members receive a 27.09% knowledge-increment improvement (from 3.5119 to 4.4632) at the slight cost of an average 5.09% knowledge-increment loss (from 9.0895 to 8.6272) for the privileged core individuals.

4.5 Model Fit

To evaluate the fit of our model, we recover the empirical CCPs using non-parametric methods in first stage and compare the simulated moments calculated from the equilibrium policy with the moments from the real data.

Table 6. Model Fit

	Data Moments	Simulated Moments from Full Model	Simulated Moments from Baseline Model
Total Number of Questions	652	635.45	742.80
Total Number of Answers	1676	1556.7	2253.74
Average Number of Questions per Week	13.04	12.71	14.85
Average Number of Answers per Week	33.52	31.13	45.07
Average Knowledge Level	1.3123	1.2786	2.4899
Average Social Status	1.0364	1.0401	1.3497

As we can see from first and second columns in Table 6, the simulated moments are very close to moments from real data, indicating that our model can explain the data well. Here, we also compare the fit of our model with a baseline model. In the baseline model, we estimate a reduced form model where the individuals only consider their own current state when they make decisions. That is, we assume away the interdependence of the member's decisions in the baseline model. Then, we employ estimated parameters to simulate the network and calculate the corresponding moments. The results are shown in the third column in Table 6 above. By comparing the results from our full model with the baseline model, we can see that the model with interdependency built in is superior. This result indicates that interdependency assumption fits data better than a model with atomistic view, and individuals in this community incorporate peers' decision into account when they make knowledge seeking and sharing decisions.

5 Conclusions, Managerial Implication, and limitations

As more and more firms are adopting social media platforms for knowledge sharing, idea generation, project management, customer service, and identifying sales and marketing opportunities, it is important to understand the fundamental drivers of user behavior to increase the return on investment (ROI). Understanding the dynamics behind the individual participation decisions becomes even more critical with the fast development of social CRMs. A typical social CRM scenario occurs when customers want to communicate their problems (i.e., customer support) or desires (i.e., future product development requirements) to a company and when company involves more employees and customers to solve customer service related issues. According to a report by Gartner, spending on social CRMs is predicted to exceed $1 billion in 2010 which is approximately 8% of all the CRM spending in that year.

Based on existing theories from economics, marketing, and social psychology, we recognized the dynamic and interdependent decision-making process and built a dynamic structural model to investigate the users' knowledge-seeking and knowledge-sharing decisions. Applying the model to data provided by an IT service consulting company, we found the following results. (1) Knowledge seeking and sharing on public social platforms are driven by the knowledge and social status of both the users themselves and their peers in the community. We showed that sharing knowledge with peers can be better explained by dynamic, interactive decision making in anticipation of future reciprocal rewards from the community. This result was supported by our further findings that users are more likely to share their knowledge when their peers are more knowledgeable and that the users who initiate knowledge seeking and sharing actions benefit significantly more than their peers in the community. (2) The formation of the cohort results from the strategic interactions described above. The users strategically choose to answer the questions asked by the more centrally located users to improve their social status and hence to obtain greater future reciprocal rewards. (3) The users located within the cohort have the advantage of obtaining help from each other and meanwhile exclude other users from participating. Thus, the "free-riding" behavior of the inactive contributors may be an equilibrium result because the existence of a cohort discourages low-status users from participating. (4) Interestingly, a decomposition analysis revealed that active learning by asking questions is much more effective for improving knowledge than reactive learning by reading answers. (5) A sensitivity analysis found that breaking the cohorts by hiding the knowledge seeker's identity can improve knowledge sharing by 35.7%.

These results suggest that it is important for management to recognize the conspicuous nature of platform adoption. The adoption can be accelerated by collective action from the entire community, such as a "Knowledge Sharing Day." Management should design platform features that encourage competition for social status, which has been shown to effectively motivate users to share and seek knowledge. However, it is important to understand that the cohort formed during the process of competition excludes remotely located users (who are usually newcomers) from participating. Features should be introduced to prevent users from being selective about whose questions they answer. Otherwise, the offline knowledge silos to be broken (the original purpose of adopting social media platforms) may appear again online. In addition, the social media platforms should be viewed more as knowledge-seeking rather than knowledge-donation platforms. Thus economic incentives should also be linked with knowledge-seeking to encourage users to ask questions to actively learn from the community.

Our research has some limitations, which open exciting avenues for future research. First, as stated previously, we made simplified assumptions about the knowledge updating process. We did not consider the quality of the answers. The proposed model can be modified to consider information quality, by ratings for questions/answers or users, for example. In addition, future research could relax the assumption that each user reads all the postings and incorporate information on user-browsing behavior to more accurately measure knowledge increments. Furthermore, future research can allow diminishing increments of knowledge as the number of answers to the same question increases. Second, we did not have information on employee productivity and job performance and therefore could not explicitly link knowledge to these measurements. It will be interesting to incorporate these variables in future research, which may help better measure the knowledge increments. Third, the two-stage estimation approach allowed us to explicitly recognize the dyadic nature and endogenize formation of the network. However, these benefits were acquired at the cost of not being able to consider the unobserved heterogeneity and inflexibility of running policy simulations. Alternative estimation methods, such as the one proposed by Aguirregabiria and Mira (2007) and the one suggested by Arcidiacono and Miller (2010), can be adopted in other research contexts when unobserved heterogeneity and policy simulations are more important. Fourth, it will be interesting to apply the model to the B2C and C2C settings, in which users have more freedom to express themselves.

References:

Aguirregabiria, Victor and Pedro Mira. 2007. "Sequential Estimation of Dynamic Discrete Games". *Econometrica* Vol. 75, No. 1, pp 1-53.

Allison, P., S. Long and T. Krauze. 1982. "Cumulative Advantage and Inequality in Science". *American Sociology Review*. 47(5), 615-625.

Ansari, A., O. Koenigsberg and F. Stahl. 2011. "Modeling Multiple Relationships in Social Networks". *Journal of Marketing Research*. 48(4), 713-728.

Arcidiacono, Peter and Robert Miller. 2010. "CCP Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity." Working Paper. Duke University.

Argote, Linda. 1999. "Organizational Learning: Creating, Retaining and Transferring Knowledge". Kluwer Academic Publishers.

Argote, Linda, Sara Beckman and Dennis Epple. 1990. "The Persistence and Transfer of Learning in Industrial Settings." *Management Science*. Vol. 36, No. 2, pp. 140-154.

Argote, Linda, Bill McEvily and Ray Reagans. 2003. "Managing Knowledge in Organizations: An Integrative Framework and Review of Emerging Themes". *Management Science*. Vol. 49, No. 4, pp. 571-582.

Atchade, Y. 2006. "An Adaptive Version for the Metropolis Adjusted Langevin Algorithm with a Truncated Drift". *Methodology and Computing in Applied Probability*. 8(2) 235-254.

Bajari, Patrick, Lanier Benkard and Jonathan Levin. 2007. "Estimating Dynamic Models of Imperfect Competition". *Econometrica*. Vol 74, No. 5, pp 1331-1370.

Bajari, Patrick, Victor Chernozhukov, Han Hong and Denis Nekipelov. 2008. "Nonparametric and Semiparametric Analysis of a Dynamic Game Model". Working Paper. University of Minnesota.

Barabasi, A. and R. Albert. 1999. "Emergence of Scaling in Random Networks." *Science*. 286(5439) 509-512.

Barsky, N. P. 1999. A Core/Periphery Structure in a Corporate Budgeting Process. Connections 22(2) 22–29.

Bay L. Bayus. 2010. "Crowdsourcing and Individual Creativity over Time: The Detrimental Effects of Past Success." Working paper, University of North Carolina.

Benkard, Lanier. 2000. "Learning and Forgetting: The Dynamics of Aircraft Production". *American Economic Review*. Vol. 90, No. 4, pp. 1034-1054.

Benkard, Lanier. 2004. "A Dynamic Analysis of the Market for Wide-Bodied Commercial Aircraft". *Review of Economic Studies*. Vol 71, No. 3, pp 581-611.

Bonacich, P. 1987. "Power and Centrality: A Family of Measures". *American Journal of Sociology*. 92(5) 1170-1182.

Bonacich, Phillip and Paulette Lloyd. 2001. "Eigenvector-like Measures of Centrality for Asymmetric Relations". *Social Networks*. Vol. 23, No. 3, pp. 191-201.

Borgatti, Stephen and Martin Everett. 2000. "Models of Core/Periphery Structures". *Social Networks*. Vol. 21, No. 4, pp. 375-395.

Braun, M. and A. Bonfrer. 2011. "Scalable Inferences of Customer Similarities from Interactions Data using Dirichlet Processes." *Marketing Science*, 30, 513-531.

Buchanan, Laurence. 2010. "GiffGaff-A Case Study of Customer in Control". http://thecustome-revolution.blogspot.com.

Burt, R. 1999. "The Social Capital of Opinion Leaders". *The Annals of the American Academy of Political and Social Science*. 566(1) 37-54.

Cattani, Gino and Simone Ferriani. 2008. "A Core/Periphery Perspective on Individual Creative Performance: Social Networks and Cinematic Achievements in the Hollywood Film Industry". *Organization Science*. Vol. 19, No. 6, pp.824-844.

Chan, K. and S. Misra. 1990. "Characteristics of the Opinion Leader: A New Dimension". *Journal of Advertising*. 19(3) 53-60.

Chen, Pei-Yu and Lorin Hitt. 2002. "Measuring Switching Costs and the Determinants of Customer Retention in Internet-Enabled Businesses: A Study of the Online Brokerage Industry". *Information System Research*. Vol. 13, No. 3, pp. 255-274.

Chevalier, Judith and Dina Mayzlin. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews". *Journal of Marketing Research*. Vol. 43, No. 3, pp. 345-354.

Chintagunta, Pradeep, Shyam Gopinath, Sriram Venkataraman. 2010. "The Effect of Online User Reviews on Movie Box-office Performance: Accounting for Sequential Rollout and Aggregation across Local Market". *Marketing Science. Forthcoming.*

Cummings, J., R. Cross. 2003. Structural Properties of Work Groups and Their Consequences for Performance. *Soc. Networks* 25(3) 197–210.

Darr, Eric, Linda Argote and Dennis Epple. 1995. "The Acquisition, Transfer and Depreciation of Knowledge in Service Organizations: Productivity in Franchise". *Management Science*. Vol. 41, No. 11, pp. 1750-1762.

Deloitte. 2010. *Webinar for the Enterprise 2.0 Conference*. Deloitte Center for the Edge.

Dorogovtsev, S. and J. Mendes. 2003. "Evolution of Networks: From Biological Nets to the Internet and WWW". Oxford University Press, USA.

DuBay, W. 2004. "The Principles of Readability". http://www.impact-information.com/.

Erdem, Tulin. 1998. "An Empirical Analysis of Umbrella Branding". *Journal of Marketing Research*, Vol. 35, No. 3, pp. 339-351

Erdem, Tulin and Michael Keane. 1996. "Decision Making Under Uncertainty: Capturing Dynamic Brand Choice Process in Turbulent Consumer Good Market". *Marketing Science*. Vol. 15, No. 1, pp. 1-20.

Erdem, Tulin, Michael Keane, T. Sabri Oncu and Judi Strebel. 2005. "Learning About Computers: An Analysis of Information Search and Technology Choice," *Quantitative Marketing and Economics*. Vol.3, No. 3, pp. 207-246

Erdem, Tulin, Michael Keane and Baohong Sun. 2008. "A Dynamic Model of Brand Choice when Price and Advertising Signal Product Quality". *Marketing Science*, 27(6), 1111-1129.

Ericson, Richard and Ariel Pakes. 1995. "Markov-Perfect Industry Dynamics: A Framework for Empirical Work". *The Review of Economic Studies*. Vol. 62, No. 1, pp. 53-82.

Fader, P., B. Hardie and K. Lee. 2005. "RFM and CLV: Using Iso-value Curves for Customer Base Analysis". *Journal of Marketing Research*. 42 (4) 415-430.

Fehr, E. and S. Gachter. 2000. "Fairness and Retaliation the economics of reciprocity". *The Journal of Economic Perspective*. 14(3) 159-181.

Forman, C., A. Ghose and B. Wiesenfel. 2008. "Examining the Relationship between Reviews and Sales: the Role of Reviewer Identity Disclosure in Electronic Markets". *Information System Research*. 19(3) 291-313.

Gartner. 2011. "Gartner Says Spending on Social Software to Support Sales, Marketing and Customer Service Processes Will Exceed $1 Billion Worldwide By 2013". http:// http://www.gartner.com/-it/page.jsp?id=1541415

Gladwell, M. 2000. "The Tipping Point". Little, Brown and Company.

Gelman, A., J. Carlin, H. Stern and D. Rubin. 2003. "Bayesian Data Analysis". Chapman and Hall/CRC.

Ghose, A, P. Iperotis and A. Sundarajan. 2009. "The Dimensions of Reputation in Electronic Markets". NYU Working Paper.

Ghose, A. and P. Ipeirotis. 2011. "Estimating Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics." *IEEE Transactions on Knowledge and Data Engineering*, 23(10), 1498-1512.

Godes, David and Dina Mayzlin. 2004. "Using Online Conversation to Study Word-of-Mouth Communication*". Marketing Science*. Vol. 23, No. 4, pp. 545-560.

Goodman, L. 1961. "Snowball Sampling". *The Annals of Mathematical Statistics*. 32(1) 148-170.

Greene, W. 2003. "Econometric Analysis", 5th Edition. Prentice Hall.

Gould, S. 2002. "The Structure of Evolutionary Theory." Harvard University Press.

Handcock, M., A. Raftery, J. Tantrum. 2007. "Model-based clustering for social networks." *Journal of the Royal Statistical Society*. Series A 170(2) 301–354.

Hartmann, Wesley. 2010. "Demand Estimation with Social Interactions and the Implications for Targeted Marketing". *Marketing Science*. Vol. 29, No. 4, pp. 585-601.

Hill, S., F. Provost and C. Volinsky. 2006. "Network-based Marketing: Identifying Likely Adopters via Consumer Networks." *Statistical Science* 22(2): 256-276.

Hinchcliffe, Dion. 2009. "14 Reasons Why Enterprise 2.0 Projects Fail". http://www.zdnet.com.

Hoff, P. 2005. "Bilinear Mixed-Effects Models for Dyadic Data." *Journal of the American Statistical Association*. 100(469) 286-295.

Hoff, P., A. Raftery, M. Handcock. 2002. "Latent Space Approaches to Social NetworkAnalysis." *Journal of the American Statistical Association* 97(460) 1090–1098.

Holland, P. and S. Leinhardt. 1972. "Reply: Some Evidence on the Transitivity of Positive Interpersonal Sentiment". *The American Journal of Sociology*. 77(6) 1205-1209.

Huang, Yan, Param Singh and Anindya Ghose. 2010. "A Structural Model of Employee Behavioral Dynamics in Enterprise Social Media". Working Paper. Carnegie Mellon University.

Hunter, D., S. Goodreau and M. Handcock. 2008. "Goodness of Fit of Social Network Models." *Journal of the American Statistical Association* 103(481) 248–258.

Iacobucci, D., and N. Hopkins. 1992. "Modeling Dyadic Interactions and Networks in Marketing." *Journal of Marketing Research*. 29(1) 5-17.

Ingram, Paul and Tal Simons. 2002. "The Transfer of Experience in Groups of Organizations: Implications for Performance and Competition". *Management Science*. Vol. 48, No. 12, pp. 1517-1533.

Iyengar, R., C. Van den Bulte and T. Valente. 2011. "Opinion Leadership and Social Contagion in New Product Diffusion". *Marketing Science*, 30(2), 195-212.

Jones, J. and M. Handcock. 2003. "An Assessment of Preferential Attachment as a Mechanism for Human Sexual Network Formation". *Proceedings of the Royal Society*. 270 (1520) 1123-1128.

Katona, Zsolt and Miklos Sarvary. 2008. "Network Formation and the Structure of the Commercial World Wide Web". *Marketing Science*. Vol. 27, No. 5, pp. 764-778.

Katz, E., P. Lazarsfeld. 1955. "Personal Influence: The Part Played by People in the Flow of Mass Communications". Glencoe, IL: Free Press.

Katz, J. 1998. Luring the Lurkers. Available at http://slashdot.org/features/98/12/28/1745252.shtml.

Katz, L., and J. Powell. 1955. "Measurement of the Tendency Towards Reciprocation of Choice". *Sociometry*. 18(4) 403-409.

Kilduff, Martin and David Krackhardt. 1994. "Bringing the Individual Back In: A Structural Analysis of the Internal Market for Reputation in Organizations". *Academy of Management Journal*. Vol. 37, No. 1, pp. 87-108

Kim, S. and E. Hovy. 2006. "Automatic Identification of Pro and Con Reasons in Online Reviews". *Annual Meeting of the ACL, Proceedings of the COLING/ACL on Main Conference Poster Sessions*. Sydney, Australia.

King, C. and J. Summers. 1970. "Overlap of Opinion Leadership across Consumer Product Categories". *Journal of Marketing Research*. 7(1) 43-50.

Kollock, Peter. 1999. "Communities in Cyberspace". Routledge Publisher.

Kossinets, G. and D. Watts. 2006. "Empirical Analysis of an Evolving Social Network". *Science*. 311, 88-90.

Kumar, Vineet, Ramayya Krishnan, and Baohong Sun, "Measuring Dynamic Effect of Promotion through Social Network." Working paper, Carnegie Mellon University.

Kumar, Vineet, Kannan Srinivasan and Baohong Sun. 2010. "Why do Consumers Contribute to Connected Goods? A Dynamic Game of Competition and Cooperation in Social Networks". Working Paper. Carnegie Mellon University.

Lakhani, Karim and Eric von Hippel. 2003. "How Open Source Software Works: "Free" User-to-User Assistance". *Research Policy*. Vol. 32, No. 6, pp. 923-943.

Lampel, Joseph and Ajay Bhalla. 2007. "The Role of Status Seeking in Online Communities: Giving the Gift of Experience*". Journal of Computer Mediated Communication*. Vol. 12, No. 2, pp. 434-455.

Levitt, Barbara and James March. 1988. "Organizational Learning". *Annual Review of Sociology*. Vol. 14, pp. 319-340.

Li, Charlene and Josh Bernoff. 2009. "Helping the Groundswell Support Itself: Customers Supporting Customers Through Social Technologies". Harvard Business School Press.

Alias-I. 2008. Lingpipe Home Page. http://alias-i.com/lingpipe/index.html

Liu, J., Y. Cao, C. Lin, Y. Huang and M. Zhou. 2007. "Low-quality Product Review Detection in Opinion Summarization". *Joint Conference on Empirical Methods in NLP and Computational NLP*, 334-342.

Lurie, Nicholas, Hai Che and Allen Weiss. 2009. "Helping Strangers: Who Contributes to Online Communities, How Much Do They Contribute, and Why". Working Paper. Georgia Institute of Technology.

Ma, Liye, Baohong Sun and Kannan Srinivasan. 2010. "A Dynamic Competitive Analysis of Content Production and Link Formation of Internet Content Developers". Working Paper. Carnegie Mellon University.

Manski, C. and S. Lerman. 1977. "The Estimation of Choice Probabilities form Choice Based Samples". *Econometrica*. 45 (8) 1977-1988.

Mayzlin D. and H. Yoganarasimhan. 2012. "Link to Success: How Blogs Build an Audience by Promoting Rivals". *Management Science*, forthcoming.

McAfee, AP. 2006. "Enterprise 2.0: The Dawn of Emergent Collaboration". *MIT Sloan Management Review*.

McPherson, M., L. Smith-Lovin and J. Cook. 2001. "Birds of a Feather: Homophily in Social Networks". *Annual Review of Sociology*. 27(August) 415-444.

Merton, R. 1968. "The Matthew Effect in Science". *Science*, 159(3810) 56-63.

Mislove, A., M. Marcon, K. Gummadi, P. Druschel and B. Bhattacharjee. 2007. "Measurement and Analysis of Online Social Networks". IMC'07, San Diego, CA, USA.

Moe, W.W. and Fader, P.S. 2001, "Uncovering Patterns in Cybershopping", *California Management Review*, Vol. 43 No. 4, pp. 106-17.

Myers, J. and T. Robertson. 1972. "Dimensions of Opinion Leadership*". Journal of Marketing Research*, 9 41–46.

Narayan, V. and S. Yang. 2007. "Modeling the Formation of Dyadic Relationships Between Consumers in Online Communities". Working Paper. Cornell University.

Narayanan, Sridhar and Puneet Manchanda. 2009. "Heterogeneous Learning and the Targeting of Marketing Communication for New Products*". Marketing Science*. Vol. 28, No. 3, pp. 424-441.

Nonnecke, Blair. 2000. "Lurking in Email-based Discussion Lists". In: SCISM. London: South Bank University.

Nonnecke, Blair and Jenny Preece. 2001. "Why Lurkers Lurk". *American Conference on Information System*.

Obradovic, Darko and Stephan Baumann. 2009. "A Journey to the Core of the Blogsphere". *Advances in Social Network and Mining*. July, pp. 1-6.

Olivera, Fernando, Paul Goodman and Sharon Tan. 2008. "Contribution Behaviors in Distributed Environments". *MIS Quarterly*. Vol. 32, No. 1, pp. 23-42.

Otterbacher, J. 2009. "'Helpfulness' in Online Communities: a Measure of Message Quality". Proceedings of the *27th International Conference on Human Factor in Computing Systems*.

Pakes, Ariel and Paul McGuire. 1994. "Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model". *The RAND Journal of Economics*. Vol. 25, No. 4, pp. 555-589.

Pang, B. and L. Lee. 2004. "A Sentimental Education". *Annual Meeting of the ACL. Proceedings of the 42nd Annual Meeting on Association for Computing Linguistics*, Barcelona, Spain.

Reagans, Ray, Linda Argote and Daria Brooks. 2005. "Individual Experience and Experience Working Together: Predicting Learning Rates from Knowing Who Knows What and Knowing How to Work Together". *Management Science*. Vol. 51, No. 6, pp. 869-881.

Reichheld, F.F. 1996, The Loyalty Effect. Harvard Business School Press, Boston, MA.

Reinartz, Werner and V. Kumar 2000, "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing*, 64 (October), 17-35

Reinartz, Werner, Manfred Krafft, Wayne D. Hoyer. 2004. "The Customer Relationship Management Process: Its Measurement and Impact on Performance". *Journal of Marketing Research*. Vol. 41, No. 3, pp. 293-305

Robins, G., T. Snijders, P. Wang, M. Handcock and P. Pattison. 2007. "Recent Developments in Exponential Random Graph (p*) models for social networks". *Social Networks*. 29(2007), 192-215.

Rogers, E. 2003. Diffusion of Innovation. Free Press.

Ryan, Stephen and Catherine Tucker. 2008. "Heterogeneity and Dynamics of Technology Adoption". Working Paper. Duke University.

Schilling, Melissa, Patricia Vidal, Robert Ployhart and Alexandre Marangoni. 2003. "Learning by Doing Something Else: Variation, Relatedness, and the Learning Curve". *Management Science*. Vol. 49, No. 1, pp. 39-56.

Singh, Param and Yong Tan. 2010. "Developer Heterogeneity and Formation of Communication Networks in Open Source Software Projects." *Journal of Management Information System*. Forthcoming

Singh, Param, Yong Tan and Vijay Mookerjee. 2010. "Network Effects: The Influence of Structural Social Capital on Open Source Project Success". *Management of Information System Quarterly*. Forthcoming.

Snijders, T., P. Pattison, G. Robins, and M. Handcock. 2006. "New Specifications for Exponential Random Graph Models". *Sociological Methodology* Vol. 36, No. 1, pp.99-153.

Stephen, A., Y. Dover, L. Muchnik, and J. Goldenberg. 2012. "The Effects of Transmitter Activity and Connectivity on Information Dissemination Over Online Social Networks". Working Paper. University of Pittsburgh.

Stephen, A. and O. Toubia. 2009. "Explaining the Power-Law Degree in a Social Commerce Network". *Social Networks*. 31(4) 262-270.

Sweeting, Andrew. 2007. "Dynamic Product Repositioning in Differentiated Product Markets: The Case of Format Switching in the Commercial Radio Industry". Working Paper. Northwestern University.

Trivers, Robert. 1971. "The Evolution of Reciprocal Altruism". *The Quarterly Review of Biology*. Vol 46, No. 1,pp 35-57.

Trusov, Michael, Anand Bodapati and Randolph Bucklin. 2010. "Determining Influential Users in Inernet Social Network*". Journal of Marketing Research*. Vol. 47, No. 4, pp. 643-658.

Tucker, C. and J. Zhang. 2010. "Growing Two-Sided Networks by Advertising the User Base: A Field Experiment." *Marketing Science*. 29 (5) 805-814.

Valente, T., B. Hoffman, A. Ritt-Olson, K. Lichtman and A. Johnson. 2003. "Effects of a Social-Network Method for Group Assignment Strategies on Peer-Led Tobacco Prevention Programs in Schools". *American Journal of Public Health*. 93(11) 1837-1843.

Van Alstyne, M., and E. Brynjolfsson. 2005. "Global Village or Cyber-Balknas? Modelling and Measuring the Integration of Electronic Communities". M*anagement Science*. 51(6), 851-868.

Van den Bulte, C. and Y. Joshi. 2007. "New Product Diffusion with Influentials and Imitators". *Marketing Science*. 26(3) 400-421.

Van den Bulte, Christophe and Gary Lilien. 2001. "Medical Innovation Revisited: Social Contagion versus Marketing Effort". *American Journal of Sociology*. Vol. 106, No. 5, pp. 1409-1435.

Vernette, E. 2004. "Targeting Women's Clothing Fashion Opinion Leaders in Media Planning: An Application for Magazine". *Journal of Advertising Research*. 44(1) 90-107.

Watts, D. and P. Dodds. 2007. "Influentials, Networks and Public Opinion Formation". *Journal of Consumer Research*. Vol. 34 441-458.

Weintraub, Gabriel, Lanier Benkard and Benjamin Van Roy. 2008. "Markov Perfect Industry Dynamics with Many Firms". *Econometrica*. Vol. 76, No. 6, pp. 1375-1411.

Weintraub, Gabriel, Lanier Benkard and Benjamin Van Roy. 2009. "Computational Methods for Oblivious Equilibrium". Working paper. Columbia University.

Wilkinson, G. 1988. Reciprocal Altruism in Bats and Other Mammals. *Ethology and Sociobiology*, 8, pp. 85-100.

Zhang, Jie and Michel Wedel (2009). "The Effectiveness of Customized Promotions in Online and Offline Stores," *Journal of Marketing Research,* 2009, 46, 190-206.

Zhang, Juanjuan. 2010. "The Sound of Silence: Observational Learning in the U.S. Kidney Market". *Marketing Science*. Vol. 29, No. 2, pp. 315-335.

Zhang, Z., and B. Varadarajan. 2006. "Utility Scoring of Product Reveiws". *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. Arlington, Virginia, USA.

Appendix I: For Chapter 2

Appendix I.A: Derivation of the Log-Conditional-Likelihood Function

For the basic proportional hazard model:

$$\lambda_{ij}(t) = \lambda_0(t)exp\{\boldsymbol{z}_{ijt}\boldsymbol{\beta}\}$$

the probability that the tie from i to j is not formed at time $t+1$, conditional on the fact that it is not formed yet at time t is:

$$P\big(T_{ij} \geq t+1 \big| T_t \geq t\big) = \exp\left(-\int_t^{t+1} \lambda_{ij}(u)du\right)$$

$$= \exp(-\exp(\boldsymbol{z}_{ijt}\boldsymbol{\beta})\int_t^{t+1} \lambda_0(u)du)$$

Here, we require the value of $\boldsymbol{z}_{ijt}$ to be invariant between t and $t+1$. The conditional probability above can be rewritten as:

$$P\big(T_{ij} \geq t+1 \big| T_{ij} \geq t\big) = \exp(-\exp(\boldsymbol{z}_{ijt}\boldsymbol{\beta} + \alpha(t)))$$

where $\alpha(t) = \log\{\int_t^{t+1} \lambda_0(u)du\}$.

Let C_{ij} be the length of time for which dyad ij has been observed, and T_{ij} be the length of time from the starting point to the time period when i extends a tie to j. Thus the log-conditional-likelihood function for a dataset with N individuals in this basic model is:

$$\log L = \sum_{i,j\neq i}\left\{\mathbb{I}_{ij} \cdot \log\left[1 - \exp\left\{-\exp\left[\alpha\big(k_{ij}\big) + \boldsymbol{z}_{ij,k_{ij}}\boldsymbol{\beta}\right]\right\}\right] - \sum_{t=0}^{k_{ij}-1} \exp[\alpha(t) + \boldsymbol{z}_{ijt}\boldsymbol{\beta}]\right\}$$

where $\mathbb{I}_{ij} = 1$ if $T_{ij} \leq C_{ij}$ (i.e., if a tie formed within the observation time) and 0 otherwise.

Appendix I.B: MCMC Inference for the Time Varying Hazard Model

The steps below provide the details of the estimation process for the time-varying hazard model with homogeneous consumer preferences. The procedure of estimating the model with heterogeneous consumer preferences is very similar to the one illustrated below; we discuss the heterogeneous case in Technical Appendix II. For the procedures below, letters with superscript u represent the values of the updated corresponding parameters.

Step 1: Estimating $\boldsymbol{\gamma}$

$$\boldsymbol{\gamma}^u | \boldsymbol{\beta}, a_i, b_i, \alpha_0, \alpha_1, d_{ij}, \text{data}$$

$$f\left(\boldsymbol{\gamma}^u \middle| \boldsymbol{\beta}, a_i, b_i, \alpha_0, \alpha_1, d_{ij}, \text{data}\right)$$

$$\propto \left|\Sigma_{\boldsymbol{\beta 0}}\right|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma_0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma_0}})\right] L(\boldsymbol{Y})$$

where $\overline{\boldsymbol{\gamma_0}}$ and $\Sigma_{\boldsymbol{\gamma 0}}$ are diffused priors. Because there is no closed form for this, we use the Metropolis-Hastings algorithm to draw from this conditional distribution of $\boldsymbol{\gamma}^u$. The probability of accepting $\boldsymbol{\gamma}^u$ is:

$$\Pr(\text{acceptance}) = \min\{\frac{\exp\left[-\frac{1}{2}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma_0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma}^u - \overline{\boldsymbol{\gamma_0}})\right] L(\boldsymbol{Y}|\boldsymbol{\gamma}^u)}{\exp\left[-\frac{1}{2}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma_0}})'\Sigma_{\boldsymbol{\gamma}0}^{-1}(\boldsymbol{\gamma} - \overline{\boldsymbol{\gamma_0}})\right] L(\boldsymbol{Y}|\boldsymbol{\beta})}, 1\}$$

We define diffuse priors by setting $\overline{\boldsymbol{\gamma_0}}$ to be a vector of zeros and $\Sigma_{\boldsymbol{\gamma 0}} = 30I$.

Step 2: Generate a_i^u, b_i^u :

$$f(a_i^u, b_i^u | \boldsymbol{\beta}^u, \alpha_0^u, \alpha_1^u, d_{ij}, \text{data})$$

$$\propto \mathrm{N}\left(\left(a_i^u, b_i^u | \boldsymbol{\beta}^u, \alpha_0^u, \alpha_1^u, d_{ij}\right), \Sigma_{ab}\right) L(\boldsymbol{Y})$$

$$\propto |\Sigma_{ab}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right] L(\boldsymbol{Y})$$

Because this distribution does not have a closed form, we use the Metropolis-Hastings algorithm to draw from the conditional distribution of a_i, b_i: a_i, b_i is the draw of the random effect from the previous iteration, and we draw a_i^u, b_i^u by $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix} = \begin{bmatrix} a_i \\ b_i \end{bmatrix} + \Delta \begin{bmatrix} a \\ b \end{bmatrix}$, where $\Delta \begin{bmatrix} a \\ b \end{bmatrix}$ is a draw from $\mathrm{N}(0, \Delta^2\Lambda)$, and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006). The probability of accepting this $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix}$, the updated value for $\begin{bmatrix} a_i \\ b_i \end{bmatrix}$ is:

$$\Pr(\text{acceptance}) = \min\{\frac{\left[\exp\left(-\frac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right)\right] L(\boldsymbol{Y}|a_i^u, b_i^u)}{\left[\exp\left(-\frac{1}{2}(a_i \ , b_i \ )\Sigma_{ab}^{-1}(a_i \ , b_i \ )'\right)\right] L(\boldsymbol{Y}|a_i \ , b_i \ )}, 1\}$$

Step 3: $\Sigma_{\mathrm{ab}}^{\mathrm{u}}|a_i^u, b_i^u$

$$(\Sigma_{\mathrm{ab}}^{\mathrm{u}}|a_i^u, b_i^u) \sim \mathrm{IW}_2\left(7 + N, G_0^{-1} + \sum_{i=1}^{N}(a_i^u, b_i^u)(a_i^u, b_i^u)'\right),$$

where IW$_2$ denotes the inverse-Wishart distribution.

Step 4: $d_{ij}^u d_{ij}^u, d_{ji}^u|\alpha_0^u, \boldsymbol{\beta}^u \ , a_i, b_i, \alpha_1^u, \sigma_d^2$,data

$$f(d_{ij}^u, d_{ji}^u|\alpha_0^u, \boldsymbol{\beta}^u \ , a_i, b_i, \alpha_1^u, \sigma_d^2, \text{data} \ )$$

$$\propto N\left((d_{ij}^u, d_{ji}^u|\alpha_0^u, \boldsymbol{\beta}^u \ , a_i, b_i, \alpha_1^u), \sigma_d^2\right) L(\boldsymbol{Y})$$

$$\propto \sigma_d^{-1} \exp\left[-\tfrac{1}{2}\left(d_{ij}^u + d_{ji}^u\right)^2 \sigma_d^{-2}\right] L(\boldsymbol{Y})$$

We use the Metropolis-Hastings algorithm to draw from this conditional distribution of d_{ij}^u and d_{ij}^u: d_{ij} and d_{ji} are the draws of the unobservable similarity effects from the previous iteration, and we draw d_{ij}^u and d_{ji}^u by $\begin{bmatrix} d_{ij}^u \\ d_{ji}^u \end{bmatrix} = \begin{bmatrix} d_{ij} \\ d_{ji} \end{bmatrix} + \Delta\boldsymbol{d}$, where $\Delta\boldsymbol{d}$ is a draw from N(0,$\Delta^2\Lambda$), and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006). The probability of accepting $\begin{bmatrix} d_{ij}^u \\ d_{ji}^u \end{bmatrix}$ is:

$$\Pr(\text{acceptance}) = \min\{\frac{\left[\exp\left(-\frac{1}{2}(d_{ij}^u + d_{ji}^u)\sigma_d^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}^u, d_{ji}^u)}{\left[\exp\left(-\frac{1}{2}(d_{ij} + d_{ji} \ )\sigma_d^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}, d_{ji} \ )}, 1\}$$

Step 5: Generating σ_d^u

$$(\sigma_d^u |\ d_{ij}^u, d_{ji}^u) \sim IW_1(1 + N(N-1), 1 + \sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}(d_{ij}^u + d_{ji}^u)^2),$$

where IW$_1$ denotes the inverse-Wishart distribution.

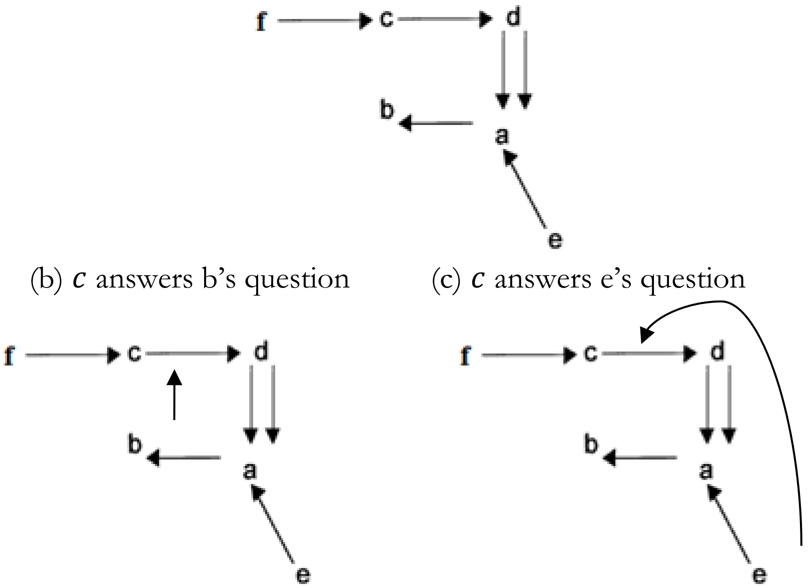
Step 6: If convergence is not reached, go to Step 1.

Appendix II: For Chapter 3

Appendix II.A: An Example Illustrating the Properties of Eigenvector Centrality and Relative Social Ranking[24]

Figure A1. Example of Constructing Social Status

(a): original link



(b) c answers b’s question (c) c answers e’s question

The graph in Figure A1(a) demonstrates the original network structure with $N = 6$ individuals. A link from i to j indicates that j answers a question proposed by i. That is, all the users but b ask at least one question in the past, and *d* asks two questions. In the past, individual *a* answered three questions, *b, c* and *d* answered one question each, and *e* and f did not answer any questions. The double arrow from *d* to *a* indicates that in the past *a* answered two questions asked by *d.* The adjacency matrix of the graph shown in Figure A1(a) can be written as:

$$A_t = \begin{bmatrix} 0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 2\,0\,0\,0\,0\,0 \\ 1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \end{bmatrix}$$

Take individual *a* as an example. Those who answer her question are shown in row 1. $A_{t,12} = 1$ indicates that *b* answered a question from *a*. Those whose questions *a* answered are shown in column 1. We can see that $A_{t,41} = 2$ and $A_{t,51} = 1$ means that *a* answers questions proposed by *d* and *e*. The sum of the i^{th} column is the number of answers individual i provided, and the sum of the i^{th} row is the number of questions individual i asked.

We illustrate how the measurement of social status derived from social structure captures the notion that answering different individuals’ questions has different impact on an individual's social status. Consider a scenario where c decides whether to answer a question from b or e.

[24] For illustrative purposes, we only show an example where no one asks hard question. The calculation of cases with answers to hard questions is similar to this example.

Table A1: Higher Improvement of Social Status by Answering Question from a more Centrally Located User

	Baseline	**c answers Question from b**	**c answers Question from e**
a	2.0000	2.0000	2.0000
b	1.5646	1.5451	1.5525
c	1.2646	1.6537	1.5051
d	1.3439	1.4412	1.4040
e	1.0000	1.0000	1.0000
f	1.0000	1.0000	1.0000

In Table A1, we present the social status scores resulting from the two cases: 1) c answers a question from b; and 2) c answers a question from e. This table reveals several important insights into the evolution of the social status score. First, we can see from the baseline column, representing the social status scores before c answers a question, that individual a possesses the highest social status because she answered three questions, while e and f have the lowest social status because they answered no question at all. Second, we can see that after c answers b' s question, b' s social status decreases as a result of an increase in c' s social status. Third, the social status score is higher if she answers a question from b, who has a higher centrality score than that from e. Put into our context, it means that an individual gains a higher social status if she answers a question proposed by a higher social-status user than one posed by a lower social-status user.

Appendix II.B. Identification and Estimation

The variations in knowledge seeking and sharing behavior with respect to the state of their own knowledge and that of the other people in the community allow us to identify the effect of knowledge level and social-status level. First of all, the updating rules for knowledge and reputation are different given an individual's actions; thus, the variation across the knowledge and reputation level can help us identify the effect of knowledge and reputation. An individual's knowledge-seeking decision will increase her knowledge level but decrease her social status level because her peers' social status increases. Individual knowledge-sharing decision will increase her social status while her knowledge level remains the same. Second, the knowledge seeking and sharing decisions critically depend on the knowledge increments from other people sharing their knowledge, which depend on the state of other individuals. If peers answer questions, the focal person's knowledge level increases, and her reputation decreases. This also helps identify the model.

The timeline of the decisions at time *t* is as follows:

1) everyone observes their own states and the states of everyone else in the community;
2) everyone receives their private shocks for the decision of asking question;
3) everyone makes predictions of their peers' decisions based on equilibrium strategy given their information on others' states in current period, and using this prediction everyone simultaneously makes decisions on whether they are going to ask a question;
4) everyone observes the questions asked (i.e., they know who asked questions in current period);
5) everyone receives their private shocks for the decision of answering questions;

6) given the information on who asked questions in current period and the predictions on others' decisions of answering questions, everyone simultaneously makes decisions on whether to provide an answer for each of the questions proposed;
7) the state variables, accumulated knowledge K_{it} and social status R_{it} , are updated.

In the first step, we non-parametrically recover the conditional choice probability as a function of the state of the individual and the state of other individuals in the same community and calculate the corresponding choice-specific value function $v_i(a_i, \boldsymbol{s_i}, \boldsymbol{S})$:

$$v_i(a_i, \boldsymbol{s_i}, \boldsymbol{S}) = E\left[U_i(a_i, \boldsymbol{s_i}, \boldsymbol{\sigma}^*_{-\boldsymbol{i}}(\boldsymbol{S}), \boldsymbol{S}) + \gamma \int V_i\left(\boldsymbol{S}^{\prime} \mid \boldsymbol{\sigma}^*\right) dP\left(\boldsymbol{S}^{\prime} \mid \boldsymbol{S}, \boldsymbol{\sigma}^*_{-i}, \boldsymbol{\sigma}^*_{\boldsymbol{i}}\right)\right].$$

It follows that

$$v_i\left(a_i^{\prime}, \boldsymbol{s}_{\boldsymbol{i}}^{\prime}, \boldsymbol{S}\right) - v_i(a_i, \boldsymbol{s_i}, \boldsymbol{S}) = \ln\left(\Pr\left(a_i^{\prime}, \boldsymbol{s}_{\boldsymbol{i}}^{\prime} \mid \boldsymbol{S}\right)\right) - \ln(\Pr(a_i, \boldsymbol{s_i} \mid \boldsymbol{S})),$$

where $\Pr(a_i, \boldsymbol{s_i} | \boldsymbol{S})$ is the probability of observing choice $a_i, \boldsymbol{s_i}$ given state $\boldsymbol{S}$ (Hotz and Miller 1994). Consequently, we derive the policy function given the states and private shock: $\boldsymbol{\sigma}_{\boldsymbol{i}}^*(\boldsymbol{S}, \boldsymbol{\varepsilon_i}) = \{a_i, \boldsymbol{s_i}\}, if$

$$v_i\left(a_i^{\prime}, \boldsymbol{s}_{\boldsymbol{i}}^{\prime}, \boldsymbol{S}\right) + \varepsilon\left(a_i^{\prime}, \boldsymbol{s}_{\boldsymbol{i}}^{\prime}\right) \leq v_i(a_i, \boldsymbol{s_i}, \boldsymbol{S}) + \varepsilon(a_i, \boldsymbol{s_i}), \forall\, a_i^{\prime}, \boldsymbol{s}_{\boldsymbol{i}}^{\prime}.$$

In the second stage, we first simulate the value function (given the policy function that we derived from first step) according to the following steps. The value function is

$$V_i(\boldsymbol{S}; \boldsymbol{\sigma}; \boldsymbol{\theta}) = E\left[\sum_{t=1}^{T} \beta^t U_i(\boldsymbol{S_t}; \boldsymbol{\sigma}; \boldsymbol{\theta}) | \boldsymbol{S_0} = \boldsymbol{S}\right].$$

Then we can simulate the value function for time period $t = 1, \ldots, T$ as follows:

1) draw private shocks $\boldsymbol{\varepsilon_{it}}(a_i, \boldsymbol{s_i})$ for each individual;
2) for a given policy function $\boldsymbol{\sigma_i}(\boldsymbol{S}, \boldsymbol{\varepsilon_i})$, we calculate the optimal action given individual state and private shock;
3) calculate the current period utility given the optimal decision;
4) calculate individual state for next time period according to the updating rule;
5) repeat 1-4 for T periods.

This simulation procedure gives us the value function for any policy function. To estimate the parameters in the model, we first perform this simulation for the true policy functions that we derived from first stage because they are the solutions to the individual utility maximization problem. We also construct alternative policy functions by adding a random number $\varepsilon a_i \sim N(0,1)$ to the private shock and simulate the value function based on alternative policies. Next, we draw n_I profiles, $(\boldsymbol{i}, \boldsymbol{S_i}, \boldsymbol{\sigma}_{\boldsymbol{i}}^{\prime})$, from a specified distribution $\mathcal{H}$. Because the policy function from first stage is the equilibrium policy, the following inequality is satisfied at the true parameter values θ_0:

$$g\left(i, \boldsymbol{S_i}, \boldsymbol{\sigma}_{\boldsymbol{i}}^{\prime}; \boldsymbol{\theta_0}\right) = V_i(\boldsymbol{S_i}; \boldsymbol{\sigma}_{\boldsymbol{i}}^*, \boldsymbol{\sigma}^*_{-i}; \boldsymbol{\theta_0}) - V_i\left(\boldsymbol{S_i}; \boldsymbol{\sigma}_{\boldsymbol{i}}^{\prime}, \boldsymbol{\sigma}^*_{-i}; \boldsymbol{\theta_0}\right) \geq 0.$$

Our estimator $\widehat{\boldsymbol{\theta}}$ minimizes the objective function below (Bajari 2007):

$$\widehat{\boldsymbol{\theta}} = argmin\, \frac{1}{n_I} \sum_{i=1}^{n_I} (\min\{g(i, \boldsymbol{S_i}, \boldsymbol{\sigma}_{\boldsymbol{i}}^*;\ \boldsymbol{\theta}), 0\})^2.$$

By exploiting the linearity of the utility function in our text, we can easily retrieve the structural parameters that minimize the expression above. We follow Bajari et al. (2007) and use bootstrapping to compute the standard errors of the inequality estimators.

Social ties are likely to exist before the introduction of the online platform. For example, individuals within the same location (e.g., the same department within a company) are more likely to know each other. These off-line relationships can easily be transferred online. Thus, instead of simply filling the initial adjacency matrix with zeros, we incorporate individual characteristics for the initial values of the adjacency matrix A_0. More specifically, we initialize the (i, j) and (j, i) elements of the adjacency matrix ($A_{0,ij}$ and $A_{0,ji}$) to be α_0 (which is to be estimated) if individuals i and j work for the same department, and zero otherwise. As a result, individuals in departments that are actively participating in the online forum have a higher initial social status because they have more connections to start with.

Similarly, the longer an individual stays with the company, the more experience she has accumulated; hence, she is more knowledgeable. As a result, we would expect employees with longer tenures to have higher knowledge levels to start with. Accordingly, we set the initial individual knowledge level, $K_{i,0}$, to one if her tenure is more than the mean tenure of the members of the community and to zero otherwise.

Appendix II.C. The Measure of the Degree of Core/Periphery Structure of a Network.

A network has a Core/Periphery structure if individuals within the network can be divided into two parts: the Core, and the Periphery. Individuals who belong to the Core form links with others in the Core and those who are in the Periphery. However, those in the periphery do not form links with each other. For example, the network below is an ideal core/periphery network.

Figure A2. An Example of a Core/Periphery Structure[a]

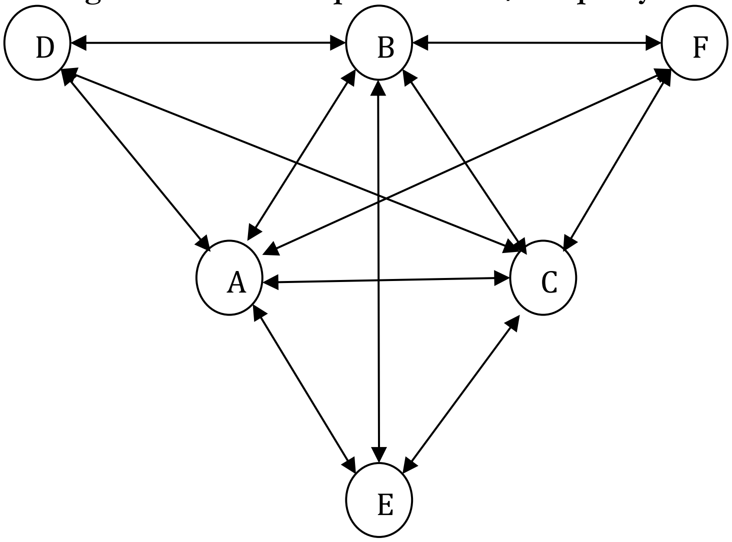


Figure A2 shows an example of a core/periphery structure with six users. The arrows represent answering questions. We can see from this figure that A, B and C belong to the core group, in which they have connections with everyone. D, E and F belong to the periphery group, in which they only have connections with core individuals. We use the following adjacency matrix to represent this ideal core/periphery network:

$$P = \left[\begin{array}{ccc|ccc} & 1 & 1 & 1 & 1 & 1 \\ 1 & & 1 & 1 & 1 & 1 \\ 1 & 1 & & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & & 0 & 0 \\ 1 & 1 & 1 & 0 & & 0 \\ 1 & 1 & 1 & 0 & 0 & \end{array}\right]$$

Network A has six individuals; the first three individuals form the core of this network, and the remaining six individuals form the periphery of this network. This is an ideal core/periphery network because peripheral individuals do not have connections with each other, whereas the only connections within this network are between core people or between core people and periphery people. The ideal matrix is called a *pattern matrix*. In this example, we show a symmetric network where elements in the matrix are either zero or one. However, the method can also be applied to our network, where the network is asymmetric, and the elements in the matrix can be larger than one.

Borgatti and Evrett (2000) propose a method for measuring the degree of core/periphery structure in a network by computing the Pearson correlation coefficient between the pattern matrix with the actual matrix: $Corr(A, P)$, where *A* is the real matrix, and *P* is the pattern matrix. The higher the correlation, the more the network has a Core/Periphery structure. For an ideal core/periphery network, the coefficient is 1. For a matrix such as

$$P = \begin{bmatrix} & 1 & 1 & 1 & 1 & 0 \\ 1 & & 1 & 1 & 0 & 1 \\ 1 & 1 & & 0 & 1 & 1 \\ 1 & 1 & 0 & & 0 & 0 \\ 1 & 0 & 1 & 0 & & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix},$$

the correlation coefficient is 0.7071. Below is a matrix that has no core group at all:

$$P = \begin{bmatrix} & 1 & 0 & 0 & 0 & 0 \\ 0 & & 1 & 0 & 0 & 0 \\ 0 & 0 & & 1 & 0 & 0 \\ 0 & 0 & 0 & & 1 & 0 \\ 0 & 0 & 0 & 0 & & 1 \\ 1 & 0 & 0 & 0 & 0 & \end{bmatrix}.$$

In our example, shown in Figure A2, each member asks and answers only one question from her peers. Then, the correlation coefficient is 0, which is exactly what a "flat" network should be. In this case, there is no group of people who are the core of the network, and everyone's social status is equally weighted. In Figure 3, we report the degree of core/periphery network structure for each time period. As described in the text, we select 30 individuals who rank in the top 30 in terms of social status and calculate the coefficient described above accordingly.

Technical Appendix: For Chapter 2

Technical Appendix I

Performance of the Weighted Exogenous Sampling with Bayesian Inference (WESBI) Method

In this technical appendix, we examine the efficacy of the Weighted Exogenous Sampling with Bayesian Inference (WESBI) method for estimating a proportional hazard network growth model. We conduct a comprehensive simulation study covering a large variety of possible network structures characterized by different parameter values. For each network structure, we show that by sampling a small proportion of the total observations, we can recover the true network generating parameters with very high accuracy.

The basic simulation process for each of the sets of parameter values we use is the following: First, we simulate a network according to the parameter values in the set. Second, we consider different sampling proportions of this simulated network; for each sampling proportion, we sample the simulated network 25 times and estimate the model using the WESBI method. For each of the different sampling proportions, we report the average posterior means and average posterior standard deviations of the parameter estimates across the 25 estimations.

We consider a total of 56 different sets of parameter values. To investigate the performance of the WESBI method on different network structures, we conduct experiments in two distinct categories of networks: networks with long tails and networks without long tails, as determined by the in-degree distribution. Because the long tail is a characteristic found in most online social networks, we use the first 32 experiments to show the performance of the WESBI method under various parameter combinations that lead to networks with long tails. For the following 24 experiments, we focus on the performance of the WESBI method for networks without long tails.

The skewness of the in-degree distribution in our Epinions.com dataset lies within the range of skewness levels of the simulated networks we consider. This suggests that the WESBI method is appropriate to use for our research context.

Network Generation Process

We simulate networks by using a variation of the classic Barabasi and Albert (1999) model. There are initially m_0 isolated nodes in the network at time $t = 0$, and m nodes are added into the network in each time period for T time periods. Subsequently, we allow the network to evolve further by allowing the tie-formation process to continue for *K* additional time periods.

The expressions below specify the proportional hazard process governing the formation of a directed link from node i to node j:

$$\lambda_{ij} = \lambda_0 \exp\left(\beta_{1,i} z_{1,j} + \beta_{2,i} z_{2,j}\right),\ \lambda_0 > 0, \tag{1}$$

$$\boldsymbol{\beta}_i = \begin{bmatrix} \beta_{1,i} \\ \beta_{2,i} \end{bmatrix} = \boldsymbol{\delta} + \boldsymbol{\varepsilon_i} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \boldsymbol{\varepsilon_i},\ \boldsymbol{\varepsilon_i} \sim \mathrm{MVN}(0, \boldsymbol{\Sigma}_\beta).$$

In the above, λ_0 is the baseline hazard rate which describes the inherent propensity of individual i forming a link with j without considering other factors and is independent of time. $z_{1,j}$ and $z_{2,j}$ are two different time-constant characteristics for individual j. Individual specific coefficients $\boldsymbol{\beta}_i$ capture how covariates have different impacts on individual tie-formation decisions across people. The quantity $\exp\left(\beta_{1,i} z_{1,j} + \beta_{2,i} z_{2,j}\right)$ increases or decreases the baseline hazard rate of tie formation between i and j.

Simulation Design for Long-Tailed Networks

It has been observed that many complex networks, especially online social networks, have long-tailed degree distributions (Barabasi and Albert 1999; Mislove et al. 2007). As a result, it is especially important to study the performance of the WESBI model for networks with long tails.

To generate the networks, we set $m_0 = 1, m = 1$. To compare how our model adapts to networks of different sizes, we set $T \in \{2000,5000\}$, resulting in networks of size 2001 or 5001, and set *K*=200. We set λ_0 to be a very small number so that the rate at which ties are formed is slow and the simulated network are relatively sparse. Specifically, we set $\lambda_0 \in \{e^{-50}, e^{-55}\}$ (i.e., log $\lambda_0 \in \{-50, -55\}$). For the parameters, $\beta_{1,i}$ and $\beta_{2,i}$, we set $\delta_1 \in \{-2,-3\}, \delta_2 \in \{2,3\}$. The variance-covariance matrix of the coefficients are set to be the same across all simulated networks: $\boldsymbol{\Sigma_\beta} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \times 0.5.$

We employ two scenarios for the distributions of individual characteristics. In the first scenario, individual characteristics, $z_{1,j}$ and $z_{2,j}$ are drawn from two independent distributions: $z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$. In the second scenario, they are drawn from two independent distributions: $z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$. Note that, in the first scenario the distribution for one covariate is skewed and the distribution for the other is symmetric, while in the second scenario both distributions are symmetric. While the exponential and the normal distributions themselves are not long-tail distributions, the hazard model we employ here allows us to construct networks in which the distributions of node degree are long-tailed. Intuitively, this is because the hazard rate of extending links is proportional to the exponent of the covariate values. Thus, the impact on tie formation of covariates with large values will be greatly magnified, implying that there will be individuals who have a large number of incoming links and will therefore contribute to a long tail. To support this argument, in the following sections, we report the mean and the max of the degrees of the various networks we generate, and compare these statistics across networks with and without long-tailed distributions. The parameter σ_z governs the variance of each distribution, and its value is set so that the generated networks are sparse.

We can vary the values of T, λ_0, δ_1, δ_2, and the distributions of the two covariates to generate networks with different characteristics. In this first simulation, we have $2^5 = 32$ different parameter combinations, i.e., we generate 32 different types of networks. As we can see from Table TA2, the simulated networks cover a large variety of network structures. The size of the network is either 2001 or 5001, the maximum in-degree ranges from 441 to 1296, and the network density ranges from 0.007% to 0.078%. The low densities are representative of common online social networks such as the ones in our Epinions study. By setting the two factors $\delta_1$and δ_2 to have opposite mean impact ($\delta_1\delta_2 < 0$), we can test the robustness of WESBI on estimating parameters with different signs. Furthermore, we assume that different factors can have different average impact on the formation of ties ($|\delta_1| \neq |\delta_2|$). Thus we can also investigate how well WESBI estimates model parameters when one factor dominates the other by varying the values of δ_1 and δ_2. While we fix the variance-covariance matrix of individual heterogeneous parameters, by changing the values of δ_1 and δ_2, we can also illustrate how the relative values in the variance-covariance matrix, compared with the mean values of the parameters, will influence the estimation results.

The conditional-log-likelihood function for the data in our simulations simplifies to the following (the notation is described in the paper):

$$\log L$$
$$= w_1\left(\sum_{(\mathbb{I}_{ij}=1)}\left\{\log\left[1-\exp\left\{-\lambda_0\exp\left(\beta_{1,i}z_{1,j}+\beta_{2,i}z_{2,j}\right)\right\}\right]-\lambda_0(k_{ij}-1)\cdot\exp\left(\beta_{1,i}z_{1,j}+\beta_{2,i}z_{2,j}\right)\right\}\right)$$

$$+w_0\left(\sum_{(\mathbb{I}_{ij}=0)}\left\{-\lambda_0(k_{ij}-\mathit{1})\cdot\exp\bigl(\beta_{1,i}z_{1,j}+\beta_{2,i}z_{2,j}\bigr)\right\}\right)$$

Here, $w_0=\frac{1-Q_1}{1-H_1}$ and $w_1=\frac{Q_1}{H_1}$, where Q_1 is the fraction of the ties formed in the whole population, and H_1 is the fraction of the ties formed in the sampled dataset. For example, in a directed network with 2000 individuals, there are in total 3,998,000 possible pairs that can form a tie. If 2,000 ties are formed, $Q_1=\frac{2000}{3998000}=0.0005$. All observations where ties are formed are included in the sampled dataset, thus if we sample 100,000 pairs out of the 3,998,000 possible pairs, $H_1=\frac{2000}{100{,}000}=0.02$. This gives $w_0=1.0199$, and $w_1=0.025$. By varying the fraction of dyads with ties formed in the sampled dataset we can explore how the effectiveness of the WESBI method depends on the number of sampled observations. We pick three possible values of the sampling proportion: 5%, 10%, 15%, which means that we sample, respectively, 5%, 10% and 15% of the total number of dyad pairs that do not form ties. Note that we always sample all the ties that are formed. For each sampling proportion of each network, we repeatedly sample the network 25 times, each time to obtain the target sampling proportion. We then estimate the model 25 times on the 25 samples using the WESBI method. We report the average posterior means and the average posterior standard deviations of the parameter value estimates recovered from the 25 runs, and compare these results with the true parameter values that we used to generate the networks.

In summary, we are estimating the model on $32\times 3=96$ different datasets, each 25 times. These 96 datasets show that the WESBI method recovers parameter values accurately and, therefore, works very well in a wide range of conditions.

Table TA1: Parameter Values Used to Simulate Long-Tailed Networks

	T	$\log(\lambda_0)$	δ_1	δ_2	Distribution of Characteristics
Network 1	2000	-50	-2	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 2	2000	-50	-2	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 3	2000	-50	-3	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 4	2000	-50	-3	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 5	2000	-55	-2	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 6	2000	-55	-2	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 7	2000	-55	-3	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 8	2000	-55	-3	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 9	5000	-50	-2	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 10	5000	-50	-2	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 11	5000	-50	-3	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 12	5000	-50	-3	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 13	5000	-55	-2	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 14	5000	-55	-2	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 15	5000	-55	-3	2	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 16	5000	-55	-3	3	$z_{1,j} \sim \text{Exponential}(1) * \sigma_z$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 17	2000	-50	-2	2	$z_{1,j} \sim \text{N}(0, \sigma_z^2)$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 18	2000	-50	-2	3	$z_{1,j} \sim \text{N}(0, \sigma_z^2)$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$
Network 19	2000	-50	-3	2	$z_{1,j} \sim \text{N}(0, \sigma_z^2)$, $z_{2,j} \sim \text{N}(0, \sigma_z^2)$

Network 20	2000	-50	-3	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 21	2000	-55	-2	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 22	2000	-55	-2	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 23	2000	-55	-3	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 24	2000	-55	-3	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 25	5000	-50	-2	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 26	5000	-50	-2	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 27	5000	-50	-3	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 28	5000	-50	-3	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 29	5000	-55	-2	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 30	5000	-55	-2	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 31	5000	-55	-3	2	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$
Network 32	5000	-55	-3	3	$z_{1,j} \sim \mathrm{N}(0, \sigma_z^2)$, $z_{2,j} \sim \mathrm{N}(0, \sigma_z^2)$

Table TA2: Statistics for the Long-Tailed Networks Simulated

	Number of ties formed in the simulated network	Mean of In-Degree	Max of In-Degree	Max/Mean Ratio of In-Degree	Network density
Network 1	2670	1.335	604	452.434	0.067%
Network 2	3079	1.540	682	443.001	0.077%
Network 3	2436	1.218	550	451.560	0.061%
Network 4	2472	1.236	619	500.809	0.062%
Network 5	1805	0.903	575	637.119	0.045%
Network 6	2611	1.306	535	409.805	0.065%
Network 7	1603	0.802	527	657.517	0.040%
Network 8	2840	1.420	729	513.380	0.071%
Network 9	2029	0.406	531	1308.526	0.008%
Network 10	2597	0.519	679	1307.278	0.010%
Network 11	2073	0.415	632	1524.361	0.008%
Network 12	2668	0.534	665	1246.252	0.011%
Network 13	1801	0.360	592	1643.531	0.007%
Network 14	3386	0.677	934	1379.209	0.014%
Network 15	2579	0.516	595	1153.548	0.010%
Network 16	2646	0.529	636	1201.814	0.011%
Network 17	3126	1.563	594	380.038	0.078%
Network 18	2693	1.347	441	327.516	0.067%
Network 19	2392	1.196	709	592.809	0.060%
Network 20	2767	1.384	960	693.892	0.069%
Network 21	2728	1.364	608	445.748	0.068%
Network 22	2448	1.224	659	538.399	0.061%
Network 23	1783	0.892	464	520.471	0.045%
Network 24	2393	1.197	563	470.539	0.060%
Network 25	2768	0.554	836	1510.116	0.011%
Network 26	3191	0.638	825	1292.698	0.013%
Network 27	2215	0.443	888	2004.515	0.009%
Network 28	2527	0.505	771	1525.524	0.010%
Network 29	2126	0.425	619	1456.471	0.009%
Network 30	3895	0.779	1296	1663.671	0.016%
Network 31	2621	0.524	881	1680.656	0.010%
Network 32	3887	0.777	1118	1438.127	0.016%

As we see from Table TA2 above, while the mean in-degree for each of the 32 networks is less than two, the maximum in-degree for each network is two or three orders of magnitude larger. This is evident from the ratio between the max and the mean in-degree. For comparison, we simulate 500 scale-free networks with long-tailed in-degree distributions following the procedure in Barabasi and Albert (1999).[25] The mean of in-degrees across the 500 networks is 1.00, and the node with maximum in-degree across the 500 networks has an in-degree of 325. By comparing statistics of the scale-free networks with those of our 32 simulated networks, it is clear that the long tail property is more salient in the 32 networks we simulated.

The 32 tables that follow show the estimation results from our simulation study. Each table reports parameter estimates for one set of parameter values. In each table, the first column reports the true parameter values used to generate the 25 sample networks. The second column reports the parameter estimates when 5% of the dyads that did not form a tie were sampled and used for parameter estimation; for each parameter, we report the average posterior mean and, in parentheses, the average standard deviation across the 25 instances. We put a check mark adjacent to these numbers if the true value of the parameter falls within the 95% credible interval that is constructed by using the average posterior mean and the average posterior standard deviation. Similarly, we report the parameter estimates when 10% and 15% of the dyads that did not form a tie were sampled for parameter estimation in the fourth and sixth columns, respectively.

The results show that sampling 10% or 15% of the dyads that do not form a tie gives very accurate estimation results with the true parameter value *always* falling in the credible interval. Even

[25] We start with one node in the network at time $\mathrm{t}=0$. Then, for T=5000 time periods, at each time period, we add one node with one tie that links the new node to one node already present in the network. The probability that a new node will link to node i depends on the degree of node i: $\mathrm{Pr(link\ to\ node\ }i) = \mathrm{Degree}_i / \sum_j \mathrm{Degree}_j$. Note that the degree distribution of a scale-free network follows a power law, which implies that it has a long tail.

when we sample only 5% of the total dyads that do not form a tie, the true network parameter value falls in the corresponding 95% credible interval approximately 75% of the time. These results show that we can estimate the parameters with high accuracy by sampling a relatively small fraction of the total network and using the WESBI method for estimation.

Network 1:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0420 ✓ (0.0248)	-50.0158 ✓ (0.0226)	-50.0067 ✓ (0.0208)
δ_1	-2	-2.0428 (0.0191)	-2.0305 ✓ (0.0171)	-2.0132 ✓ (0.0169)
δ_2	2	2.0329 ✓ (0.0181)	2.0244 ✓ (0.0174)	2.0076 ✓ (0.0169)
$\Sigma_{\beta,11}$	0.5	0.5268 ✓ (0.0214)	0.5153 ✓ (0.0195)	0.5108 ✓ (0.0185)
$\Sigma_{\beta,12}$	0.25	0.2564 ✓ (0.0164)	0.2532 ✓ (0.0161)	0.2526 ✓ (0.0159)
$\Sigma_{\beta,22}$	0.5	0.5257 ✓ (0.0192)	0.5138 ✓ (0.0162)	0.5117 ✓ (0.0157)

Network 2

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0391 ✓ (0.0224)	-50.0239 ✓ (0.0209)	-49.9983 ✓ (0.0195)
δ_1	-2	-1.9836 ✓ (0.0182)	-1.9850 ✓ (0.0180)	-1.9885 ✓ (0.0176)
δ_2	3	3.0183 ✓ (0.0174)	2. 9925 ✓ (0.0171)	3.0038 ✓ (0.0167)
$\Sigma_{\beta,11}$	0.5	0.5357 ✓ (0.0194)	0.5202 ✓ (0.0189)	0.5147 ✓ (0.0184)
$\Sigma_{\beta,12}$	0.25	0.2534 ✓ (0.0150)	0.2510 ✓ (0.0138)	0.2508 ✓ (0.0136)
$\Sigma_{\beta,22}$	0.5	0.5058 ✓ (0.0176)	0.5030 ✓ (0.0173)	0.5016 ✓ (0.0171)

Network 3:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9681 ✓ (0.0226)	-49.9746 ✓ (0.0217)	-49.9804 ✓ (0.0203)
δ_1	-3	-3.0312 ✓ (0.0184)	-3.0268 ✓ (0.0183)	-3.0192 ✓ (0.0180)
δ_2	2	1.9602 (0.0186)	1.9738 ✓ (0.0184)	1.9832 ✓ (0.0179)
$\Sigma_{\beta,11}$	0.5	0.5279 ✓ (0.0187)	0.5145 ✓ (0.0185)	0.5073 ✓ (0.0183)
$\Sigma_{\beta,12}$	0.25	0.2552 ✓ (0.0163)	0.2549 ✓ (0.0158)	0.2545 ✓ (0.0149)
$\Sigma_{\beta,22}$	0.5	0.5089 ✓ (0.0178)	0.5069 ✓ (0.0171)	0.5056 ✓ (0.0164)

Network 4:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0507 (0.0194)	-50.0305 ✓ (0.0190)	-50.0177 ✓ (0.0182)
δ_1	-3	-3.0479 (0.0185)	-3.0276 ✓ (0.0177)	-3.0176 ✓ (0.0173)
δ_2	3	2.9668 ✓ (0.0183)	2.9727 ✓ (0.0174)	2.9819 ✓ (0.0170)
$\Sigma_{\beta,11}$	0.5	0.5304 ✓ (0.0196)	0.5231 ✓ (0.0191)	0.5165 ✓ (0.0187)
$\Sigma_{\beta,12}$	0.25	0.2681 ✓ (0.0172)	0.2658 ✓ (0.0168)	0.2598 ✓ (0.0159)
$\Sigma_{\beta,22}$	0.5	0.4813 ✓ (0.0176)	0.4876 ✓ (0.0173)	0.4915 ✓ (0.0167)

Network 5:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-54.8736 ✓ (0.0233)	-54.8966 ✓ (0.0219)	-54.9361 ✓ (0.0192)
δ_1	-2	-2.0140 ✓ (0.0194)	-2.0115 ✓ (0.0184)	-2.0080 ✓ (0.0179)
δ_2	2	2.0395 (0.0182)	2.0288 ✓ (0.0180)	2.0152 ✓ (0.0172)
$\Sigma_{\beta,11}$	0.5	0.5376 (0.0187)	0.5216 ✓ (0.0183)	0.5148 ✓ (0.0182)
$\Sigma_{\beta,12}$	0.25	0.2612 ✓ (0.0171)	0.2601 ✓ (0.0167)	0.2585 ✓ (0.0156)
$\Sigma_{\beta,22}$	0.5	0.5126 ✓ (0.0184)	0.5095 ✓ (0.0172)	0.5086 ✓ (0.0164)

Network 6:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-55.0291 ✓ (0.0204)	-54.9833 ✓ (0.0199)	-55.0084 ✓ (0.0199)
δ_1	-2	-1.9850 ✓ (0.0185)	-1.9877 ✓ (0.0181)	-1.9928 ✓ (0.0165)
δ_2	3	3.0261 ✓ (0.0177)	3.0255 ✓ (0.0176)	3.0173 ✓ (0.0169)
$\Sigma_{\beta,11}$	0.5	0. 4807 ✓ (0.0192)	0. 4815 ✓ (0.0180)	0. 4862 ✓ (0.0174)
$\Sigma_{\beta,12}$	0.25	0.2426 ✓ (0.0151)	0.2442 ✓ (0.0146)	0.2447 ✓ (0.0140)
$\Sigma_{\beta,22}$	0.5	0.5162 ✓ (0.0176)	0.5115 ✓ (0.0168)	0.5107 ✓ (0.0165)

Network 7:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0329 (0.0216)	✓	-55.0279 (0.0203)	✓	-55.0138 (0.0183)	✓
δ_1	-3	-2.9613 (0.0184)		-2.9796 (0.0177)	✓	-2.9853 (0.0173)	✓
δ_2	2	1.9710 (0.0184)	✓	1.9763 (0.0180)	✓	1.9810 (0.0176)	✓
$\Sigma_{\beta,11}$	0.5	0.4857 (0.0183)	✓	0.4872 (0.0180)	✓	0.4937 (0.0173)	✓
$\Sigma_{\beta,12}$	0.25	0.2566 (0.0169)	✓	0.2460 (0.0164)	✓	0.2532 (0.0156)	✓
$\Sigma_{\beta,22}$	0.5	0.5291 (0.0176)	✓	0.5230 (0.0170)	✓	0.5184 (0.0168)	✓

Network 8:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0385 (0.0286)	✓	-55.0295 (0.0264)	✓	-55.0207 (0.0253)	✓
δ_1	-3	-3.0310 (0.0187)	✓	-3.0193 (0.0182)	✓	-3.0102 (0.0175)	✓
δ_2	3	3.0312 (0.0175)	✓	3.0236 (0.0172)	✓	3.0120 (0.0168)	✓
$\Sigma_{\beta,11}$	0.5	0.4877 (0.0185)	✓	0.4893 (0.0174)	✓	0.4918 (0.0169)	✓
$\Sigma_{\beta,12}$	0.25	0.2578 (0.0182)	✓	0.2563 (0.0174)	✓	0.2558 (0.0169)	✓
$\Sigma_{\beta,22}$	0.5	0.4863 (0.0167)	✓	0.5036 (0.0163)	✓	0.4973 (0.0157)	✓

Network 9:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9711 ✓ (0.0184)	-49.9778 ✓ (0.0174)	-49.9820 ✓ (0.0168)
δ_1	-2	-2.0245 (0.0117)	-2.0191 ✓ (0.0112)	-2.0159 ✓ (0.0109)
δ_2	2	2.0283 (0.0122)	2.0183 ✓ (0.0114)	2.0146 ✓ (0.0112)
$\Sigma_{\beta,11}$	0.5	0.5148 ✓ (0.0116)	0.5103 ✓ (0.0115)	0.5086 ✓ (0.0109)
$\Sigma_{\beta,12}$	0.25	0.2583 ✓ (0.0103)	0.2567 ✓ (0.0097)	0.2539 ✓ (0.0097)
$\Sigma_{\beta,22}$	0.5	0.5134 ✓ (0.0115)	0.5084 ✓ (0.0112)	0.5079 ✓ (0.0109)

Network 10:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0279 ✓ (0.0178)	-50.0250 ✓ (0.0173)	-50.0167 ✓ (0.0169)
δ_1	-2	-2.0275 (0.0116)	-2.0193 ✓ (0.0115)	-2.0174 ✓ (0.0114)
δ_2	3	3.0178 ✓ (0.0115)	3.0157 ✓ (0.0112)	3.0122 ✓ (0.0111)
$\Sigma_{\beta,11}$	0.5	0.5135 ✓ (0.0129)	0.5078 ✓ (0.0124)	0.5064 ✓ (0.0119)
$\Sigma_{\beta,12}$	0.25	0.2611 ✓ (0.0086)	0.2594 ✓ (0.0086)	0.2583 ✓ (0.0082)
$\Sigma_{\beta,22}$	0.5	0.5137 ✓ (0.0114)	0.5124 ✓ (0.0106)	0.5084 ✓ (0.0106)

Network 11:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0395 (0.0185)	-50.0216 ✓ (0.0183)	-50.0141 ✓ (0.0178)
δ_1	-3	-3.0218 ✓ (0.0123)	-3.0135 ✓ (0.0119)	-3.0063 ✓ (0.0118)
δ_2	2	2.0256 (0.0119)	2.0158 ✓ (0.0115)	2.0126 ✓ (0.0109)
$\Sigma_{\beta,11}$	0.5	0.5122 ✓ (0.0116)	0.5109 ✓ (0.0114)	0.5089 ✓ (0.0114)
$\Sigma_{\beta,12}$	0.25	0.2561 ✓ (0.0106)	0.2558 ✓ (0.0104)	0.2556 ✓ (0.0098)
$\Sigma_{\beta,22}$	0.5	0.5160 ✓ (0.0112)	0.5138 ✓ (0.0110)	0.5112 ✓ (0.0110)

Network 12:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9653 ✓ (0.0190)	-49.9775 ✓ (0.0186)	-49.9831 ✓ (0.0183)
δ_1	-3	-3.0260 (0.0118)	-3.0150 ✓ (0.0115)	-3.0126 ✓ (0.0115)
δ_2	3	3.0140 ✓ (0.0117)	3.0102 ✓ (0.0115)	3.0088 ✓ (0.0114)
$\Sigma_{\beta,11}$	0.5	0.5069 ✓ (0.0114)	0.5061 ✓ (0.0113)	0.5057 ✓ (0.0110)
$\Sigma_{\beta,12}$	0.25	0.2549 ✓ (0.0097)	0.2535 ✓ (0.0097)	0.2496 ✓ (0.0097)
$\Sigma_{\beta,22}$	0.5	0.5197 ✓ (0.0113)	0.5123 ✓ (0.0112)	0.5086 ✓ (0.0110)

Network 13:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-54.9694 (0.0192)	✓	-54.9821 (0.0183)	✓	-54.9878 (0.0177)	✓
δ_1	-2	-2.0291 (0.0116)		-2.0184 (0.0115)	✓	-2.0123 (0.0113)	✓
δ_2	2	2.0215 (0.0121)	✓	2.0165 (0.0116)	✓	2.0108 (0.0114)	✓
$\Sigma_{\beta,11}$	0.5	0.5194 (0.0118)	✓	0.5089 (0.0117)	✓	0.5062 (0.0115)	✓
$\Sigma_{\beta,12}$	0.25	0.2529 (0.0096)	✓	0.2516 (0.0095)	✓	0.2497 (0.0095)	✓
$\Sigma_{\beta,22}$	0.5	0.5121 (0.0117)	✓	0.5093 (0.0116)	✓	0.5064 (0.0113)	✓

Network 14:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0399 (0.0186)		-55.0218 (0.0181)	✓	-55.0169 (0.0176)	✓
δ_1	-2	-1.9731 (0.0120)		-1.9812 (0.0116)	✓	-1.9884 (0.0113)	✓
δ_2	3	3.0127 (0.0114)	✓	3.0073 (0.0114)	✓	3.0054 (0.0112)	✓
$\Sigma_{\beta,11}$	0.5	0. 5165 (0.0116)	✓	0. 5114 (0.0112)	✓	0. 5065 (0.0110)	✓
$\Sigma_{\beta,12}$	0.25	0.2417 (0.0103)	✓	0.2441 (0.0100)	✓	0.2449 (0.0097)	✓
$\Sigma_{\beta,22}$	0.5	0.5089 (0.0118)	✓	0.5079 (0.0114)	✓	0.5065 (0.0112)	✓

Network 15:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-55.0314 ✓ (0.0194)	-55.0176 ✓ (0.0187)	-55.0086 ✓ (0.0180)
δ_1	-3	-3.0145 ✓ (0.0120)	-3.0116 ✓ (0.0117)	-3.0088 ✓ (0.0116)
δ_2	2	2.0176 ✓ (0.0113)	2.0128 ✓ (0.0113)	2.0076 ✓ (0.0111)
$\Sigma_{\beta,11}$	0.5	0.5092 ✓ (0.0118)	0.5089 ✓ (0.0116)	0.5033 ✓ (0.0114)
$\Sigma_{\beta,12}$	0.25	0.2590 ✓ (0.0103)	0.2584 ✓ (0.0100)	0.2563 ✓ (0.0097)
$\Sigma_{\beta,22}$	0.5	0.5128 ✓ (0.0110)	0.5102 ✓ (0.0107)	0.5072 ✓ (0.0104)

Network 16:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-54.9583 (0.0193)	-54.9812 ✓ (0.0184)	-54.9843 ✓ (0.0181)
δ_1	-3	-3.0314 (0.0119)	-3.0184 ✓ (0.0117)	-3.0138 ✓ (0.0115)
δ_2	3	2.9743 (0.0116)	2.9824 ✓ (0.0114)	2.9875 ✓ (0.0114)
$\Sigma_{\beta,11}$	0.5	0.5180 ✓ (0.0114)	0.5124 ✓ (0.0113)	0.5089 ✓ (0.0110)
$\Sigma_{\beta,12}$	0.25	0.2579 ✓ (0.0097)	0.2573 ✓ (0.0096)	0.2553 ✓ (0.0093)
$\Sigma_{\beta,22}$	0.5	0.5156 ✓ (0.0117)	0.5083 ✓ (0.0114)	0.5068 ✓ (0.0113)

Network 17:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0468 (0.0259)	✓	-50.0285 (0.0232)	✓	-50.0208 (0.0207)	✓
δ_1	-2	-2.0362 (0.0195)	✓	-2.0278 (0.0186)	✓	-2.0132 (0.0182)	✓
δ_2	2	2.0321 (0.0190)	✓	2.0208 (0.0185)	✓	2.0132 (0.0181)	✓
$\Sigma_{\beta,11}$	0.5	0.5348 (0.0207)	✓	0.5235 (0.0191)	✓	0.5176 (0.0183)	✓
$\Sigma_{\beta,12}$	0.25	0.2423 (0.0159)	✓	0.2452 (0.0156)	✓	0.2466 (0.0151)	✓
$\Sigma_{\beta,22}$	0.5	0.5195 (0.0205)	✓	0.5164 (0.0196)	✓	0.5103 (0.0173)	✓

Network 18:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0573 (0.0236)		-50.0287 (0.0221)	✓	-50.0185 (0.0209)	✓
δ_1	-2	-2.0277 (0.0192)	✓	-2.0098 (0.0187)	✓	-2.0034 (0.0183)	✓
δ_2	3	3.0397 (0.0182)		3.0169 (0.0179)	✓	3.0144 (0.0171)	✓
$\Sigma_{\beta,11}$	0.5	0.5295 (0.0185)	✓	0.5208 (0.0181)	✓	0.5124 (0.0175)	✓
$\Sigma_{\beta,12}$	0.25	0.2622 (0.0159)	✓	0.2576 (0.0147)	✓	0.2541 (0.0140)	✓
$\Sigma_{\beta,22}$	0.5	0.5167 (0.0175)	✓	0.5134 (0.0172)	✓	0.5065 (0.0171)	✓

Network 19:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0271 (0.0218)	✓	-50.0176 (0.0204)	✓	-50.0117 (0.0196)	✓
δ_1	-3	-3.0456 (0.0192)		-3.0187 (0.0187)	✓	-3.0145 (0.0183)	✓
δ_2	2	2.0207 (0.0198)	✓	2.0164 (0.0193)	✓	2.0117 (0.0185)	✓
$\Sigma_{\beta,11}$	0.5	0.4734 (0.0186)	✓	0.4895 (0.0181)	✓	0.4944 (0.0174)	✓
$\Sigma_{\beta,12}$	0.25	0.2611 (0.0156)	✓	0.2570 (0.0153)	✓	0.2561 (0.0150)	✓
$\Sigma_{\beta,22}$	0.5	0.4761 (0.0182)	✓	0.4820 (0.0181)	✓	0.4788 (0.0173)	✓

Network 20:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0502 (0.0213)		-50.0278 (0.0197)	✓	-49.9950 (0.0191)	✓
δ_1	-3	-3.0529 (0.0172)		-3.0306 (0.0168)	✓	-3.0208 (0.0164)	✓
δ_2	3	3.0502 (0.0182)		3.0314 (0.0180)	✓	3.0239 (0.0167)	✓
$\Sigma_{\beta,11}$	0.5	0.4965 (0.0177)	✓	0.4972 (0.0175)	✓	0.4998 (0.0165)	✓
$\Sigma_{\beta,12}$	0.25	0.2682 (0.0171)	✓	0.2637 (0.0168)	✓	0.2572 (0.0157)	✓
$\Sigma_{\beta,22}$	0.5	0.5174 (0.0188)	✓	0.5155 (0.0184)	✓	0.5084 (0.0181)	✓

Network 21:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-54.4894 (0.0227)	-54.8065 ✓ (0.0221)	-54.8543 ✓ (0.0202)
δ_1	-2	-2.0308 ✓ (0.0191)	-2.0207 ✓ (0.0187)	-2.0143 ✓ (0.0182)
δ_2	2	2.0298 ✓ (0.0187)	2.0230 ✓ (0.0179)	2.0186 ✓ (0.0172)
$\Sigma_{\beta,11}$	0.5	0.5138 ✓ (0.0191)	0.5117 ✓ (0.0187)	0.5063 ✓ (0.0181)
$\Sigma_{\beta,12}$	0.25	0.2543 ✓ (0.0168)	0.2524 ✓ (0.0162)	0.2485 ✓ (0.0159)
$\Sigma_{\beta,22}$	0.5	0.5163 ✓ (0.0185)	0.5112 ✓ (0.0176)	0.5076 ✓ (0.0169)

Network 22:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-55.0581 (0.0214)	-55.0407 ✓ (0.0212)	-55.0157 ✓ (0.0198)
δ_1	-2	-2.0308 ✓ (0.0175)	-2.0249 ✓ (0.0173)	-3.0206 ✓ (0.0167)
δ_2	3	3.0390 (0.0175)	3.0303 ✓ (0.0172)	3.0262 ✓ (0.0165)
$\Sigma_{\beta,11}$	0.5	0.5170 ✓ (0.0179)	0.5136 ✓ (0.0172)	0.5117 ✓ (0.0167)
$\Sigma_{\beta,12}$	0.25	0.2708 ✓ (0.0142)	0.2693 ✓ (0.0132)	0.2668 ✓ (0.0130)
$\Sigma_{\beta,22}$	0.5	0.5216 ✓ (0.0181)	0.5203 ✓ (0.0174)	0.5173 ✓ (0.0167)

Network 23:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-54.9529 (0.0225)	-54.9727 ✓ (0.0208)	-54.9858 ✓ (0.0196)
δ_1	-3	-3.0094 ✓ (0.0184)	-3.0049 ✓ (0.0173)	-2.9994 ✓ (0.0169)
δ_2	2	2.0178 ✓ (0.0177)	2.0088 ✓ (0.0176)	2.0032 ✓ (0.0172)
$\Sigma_{\beta,11}$	0.5	0.5189 ✓ (0.0187)	0.5148 ✓ (0.0182)	0.5128 ✓ (0.0177)
$\Sigma_{\beta,12}$	0.25	0.2418 ✓ (0.0141)	0.2429 ✓ (0.0139)	0.2441 ✓ (0.0136)
$\Sigma_{\beta,22}$	0.5	0.5072 ✓ (0.0183)	0.5022 ✓ (0.0178)	0.5011 ✓ (0.0171)

Network 24:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-55	-55.0519 (0.0243)	-55.0291 ✓ (0.0215)	-55.0159 ✓ (0.0202)
δ_1	-3	-3.0372 ✓ (0.0193)	-3.0244 ✓ (0.0182)	-3.0145 ✓ (0.0177)
δ_2	3	3.0138 ✓ (0.0182)	3.0105 ✓ (0.0175)	3.0074 ✓ (0.0171)
$\Sigma_{\beta,11}$	0.5	0.5125 ✓ (0.0183)	0.5098 ✓ (0.0179)	0.5053 ✓ (0.0174)
$\Sigma_{\beta,12}$	0.25	0.2407 ✓ (0.0178)	0.2446 ✓ (0.0176)	0.2472 ✓ (0.0171)
$\Sigma_{\beta,22}$	0.5	0.5426 (0.0175)	0.5214 ✓ (0.0171)	0.5133 ✓ (0.0168)

Network 25:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9575 (0.0192)	-49.9754 ✓ (0.0186)	-49.9840 ✓ (0.0174)
δ_1	-2	-2.0281 (0.0120)	-2.0204 ✓ (0.0117)	-2.0187 ✓ (0.0114)
δ_2	2	2.0299 (0.0119)	2.0187 ✓ (0.0115)	2.0140 ✓ (0.0111)
$\Sigma_{\beta,11}$	0.5	0.5186 ✓ (0.0121)	0.5154 ✓ (0.0117)	0.5145 ✓ (0.0117)
$\Sigma_{\beta,12}$	0.25	0.2619 ✓ (0.0105)	0.2586 ✓ (0.0101)	0.2572 ✓ (0.0097)
$\Sigma_{\beta,22}$	0.5	0.5255 ✓ (0.0114)	0.5184 ✓ (0.0111)	0.5127 ✓ (0.0109)

Network 26:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9550 (0.0191)	-49.9703 ✓ (0.0163)	-49.9982 ✓ (0.0144)
δ_1	-2	-2.0157 ✓ (0.0115)	-2.0133 ✓ (0.0113)	-2.0114 ✓ (0.0103)
δ_2	3	3.0288 (0.0118)	3.0201 ✓ (0.0112)	3.0169 ✓ (0.0111)
$\Sigma_{\beta,11}$	0.5	0.5159 ✓ (0.0121)	0.5142 ✓ (0.0115)	0.5126 ✓ (0.0113)
$\Sigma_{\beta,12}$	0.25	0.2464 ✓ (0.0099)	0.2477 ✓ (0.0091)	0.2482 ✓ (0.0090)
$\Sigma_{\beta,22}$	0.5	0.5025 ✓ (0.0119)	0.4975 ✓ (0.0116)	0.4987 ✓ (0.0109)

Network 27:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9730 ✓ (0.0236)	-49.9828 ✓ (0.0207)	-49.9893 ✓ (0.0183)
δ_1	-3	-3.0251 (0.0119)	-3.0207 ✓ (0.0111)	-3.0196 ✓ (0.0109)
δ_2	2	2.0078 ✓ (0.0115)	2.0077 ✓ (0.0111)	2.0002 ✓ (0.0106)
$\Sigma_{\beta,11}$	0.5	0.4908 ✓ (0.0119)	0.4942 ✓ (0.0117)	0.4969 ✓ (0.0111)
$\Sigma_{\beta,12}$	0.25	0.2425 ✓ (0.0099)	0.2433 ✓ (0.0094)	0.2442 ✓ (0.0088)
$\Sigma_{\beta,22}$	0.5	0.4939 ✓ (0.0116)	0.4974 ✓ (0.0113)	0.4985 ✓ (0.0107)

Network 28:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0160 ✓ (0.0194)	-50.0126 ✓ (0.0190)	-50.0087 ✓ (0.0184)
δ_1	-3	-3.0235 (0.0117)	-3.0194 ✓ (0.0112)	-3.0126 ✓ (0.0107)
δ_2	3	3.0271 (0.0115)	3.0197 ✓ (0.0111)	3.0137 ✓ (0.0108)
$\Sigma_{\beta,11}$	0.5	0.5284 (0.0124)	0.5142 ✓ (0.0120)	0.5095 ✓ (0.0115)
$\Sigma_{\beta,12}$	0.25	0.2592 ✓ (0.0104)	0.2558 ✓ (0.0101)	0.2476 ✓ (0.0093)
$\Sigma_{\beta,22}$	0.5	0.5201 ✓ (0.0115)	0.5193 ✓ (0.0111)	0.5135 ✓ (0.0109)

Network 29:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-54.9721 (0.0200)	✓	-54.9794 (0.0192)	✓	-54.9853 (0.0185)	✓
δ_1	-2	-2.0294 (0.0121)		-2.0204 (0.0117)	✓	-2.0158 (0.0112)	✓
δ_2	2	2.0322 (0.0116)		2.0184 (0.0113)	✓	2.0116 (0.0111)	✓
$\Sigma_{\beta,11}$	0.5	0.5183 (0.0120)	✓	0.5120 (0.0115)	✓	0.5062 (0.0109)	✓
$\Sigma_{\beta,12}$	0.25	0.2423 (0.0093)	✓	0.2458 (0.0092)	✓	0.2477 (0.0092)	✓
$\Sigma_{\beta,22}$	0.5	0.5159 (0.0119)	✓	0.5137 (0.0117)	✓	0.5087 (0.0112)	✓

Network 30:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0402 (0.0193)		-55.0311 (0.0191)	✓	-55.0248 (0.0186)	✓
δ_1	-2	-1.9718 (0.0118)		-1.9862 (0.0113)	✓	-1.9913 (0.0110)	✓
δ_2	3	2.9769 (0.0114)		2.9820 (0.0111)	✓	2.9897 (0.0110)	✓
$\Sigma_{\beta,11}$	0.5	0. 5072 (0.0120)	✓	0. 5026 (0.0113)	✓	0. 4992 (0.0107)	✓
$\Sigma_{\beta,12}$	0.25	0.2588 (0.0103)	✓	0.2523 (0.0101)	✓	0.2497 (0.0097)	✓
$\Sigma_{\beta,22}$	0.5	0.5102 (0.0114)	✓	0.5093 (0.0112)	✓	0.5064 (0.0111)	✓

Network 31:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0165 (0.0202)	✓	-55.0133 (0.0190)	✓	-55.0083 (0.0180)	✓
δ_1	-3	-3.0221 (0.0121)	✓	-3.0168 (0.0115)	✓	-3.0107 (0.0113)	✓
δ_2	2	2.0269 (0.0123)		2.0205 (0.0116)	✓	2.0154 (0.0108)	✓
$\Sigma_{\beta,11}$	0.5	0.5161 (0.0123)	✓	0.5116 (0.0115)	✓	0.5074 (0.0114)	✓
$\Sigma_{\beta,12}$	0.25	0.2428 (0.0104)	✓	0.2469 (0.0101)	✓	0.2478 (0.0095)	✓
$\Sigma_{\beta,22}$	0.5	0.5080 (0.0106)	✓	0.5051 (0.0102)	✓	0.5021 (0.0098)	✓

Network 32:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-55	-55.0441 (0.0202)		-55.0259 (0.0196)	✓	-55.0187 (0.0182)	✓
δ_1	-3	-2.9730 (0.0124)		-2.9879 (0.0118)	✓	-2.9930 (0.0110)	✓
δ_2	3	2.9739 (0.0119)		2.9819 (0.0112)	✓	2.9883 (0.0111)	✓
$\Sigma_{\beta,11}$	0.5	0.5107 (0.0110)	✓	0.5077 (0.0107)	✓	0.5020 (0.0102)	✓
$\Sigma_{\beta,12}$	0.25	0.2425 (0.0093)	✓	0.2458 (0.0092)	✓	0.2484 (0.0092)	✓
$\Sigma_{\beta,22}$	0.5	0.5182 (0.0115)	✓	0.5082 (0.0113)	✓	0.5037 (0.0111)	✓

Simulation Design for Non-Long-Tailed Networks

All generated networks in the preceding simulation exercise have a long-tailed degree distribution. While this characteristic is present in most online social networks, it is nonetheless important to demonstrate the performance of WESBI on networks that are of "short" tail. We conduct this exercise now. Following the same simulation scheme described above for networks with long tails, we vary the values of $T,$ δ_1, δ_2, and the distributions of the two covariates to generate networks with different characteristics. Notably, we choose distributions for the individual characteristics, $z_{1,j}$ and $z_{2,j}$, such that the generated networks do not have long-tailed degree distributions.

Specifically, we consider three scenarios. In the first scenario, $z_{1,j}$ and $z_{2,j}$ are drawn from the following independent distributions: $z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$. In the second scenario, $z_{1,j}$ and $z_{2,j}$ are drawn from the following independent distributions: $z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$ $+\mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$ $+ \mu_{z2}$, where $\mu_{z1} < 0$ and $\mu_{z2} > 0$. In the third scenario, $z_{1,j}$ and $z_{2,j}$ are drawn from the following independent distributions: $z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j}$~$\text{Log-normal}(0,1) * \sigma_{z_2}$. Note that in the first scenario the value of individual characteristics are bounded from above and the distribution is uniform, and thus we will not observe individuals with an exceptionally large number of incoming links. Correspondingly, networks generated according to the first scenario will not be long-tailed. In the second scenario, we assign a small value to σ_z and set $\mu_{z1} < 0$ and $\mu_{z2} > 0$, which simulates networks which are significantly less heterogeneous and less skewed, i.e., with shorter tails, compared with those in the first scenario. In the third scenario, we consider "hybrid" cases in which the distribution of one individual characteristic is uniform and bounded from above, while the distribution of the other individual characteristic is skewed and not bounded from above (by virtue of being log-normally distributed). In a later part of this section, we use some statistics to show this "short tail" property in networks generated in these three scenarios.

In the simulation in this section, we have $2^3 \times 3 = 24$ different parameter combinations, thus we generate 24 different types of networks. As we can see from Table TA4, the simulated networks cover a large variety of network structures. The size of the network can be either 2001 or 5001, the maximum in-degree ranges from 9 to 347, and the network density ranges from 0.007% to 0.090%. Table TA4 shows some statistics of the networks that are generated in these scenarios. As we can see, the maximum in-degrees of the 24 networks generated in these scenarios are significantly smaller than the maximum in-degrees of the 32 long-tail networks in Table TA2. This suggests that the networks studied in this section have relatively "short" tails.

For further comparison, we plot the in-degree distributions of representative short- and long-tailed networks in Figure TA1. We use data from Network 45 as an example of a short-tail network, and Network 25 as an example of a long-tail network. The *x*-axis shows the in-degree (exact in-degree when this value is ≤3, and a range for larger in-degree, with the range progressively increasing), and the *y*-axis shows the frequency on a log scale. As we can see from Figure TA1, the tail of the in-degree distribution for the short-tail network disappears even for small in-degree (the maximum in-degree is 9 in this case), while the tail of the in-degree distribution for the long-tail network extends to much larger numbers (the maximum in-degree is 836 in this case). By combining generated networks from all scenarios with both long tail and short tail, our simulation study covers a wide range of network structures in terms of their thickness of the tails of the in-degree distributions.

Table TA3: Parameter Values Used to Simulate Non-Long-Tailed Networks

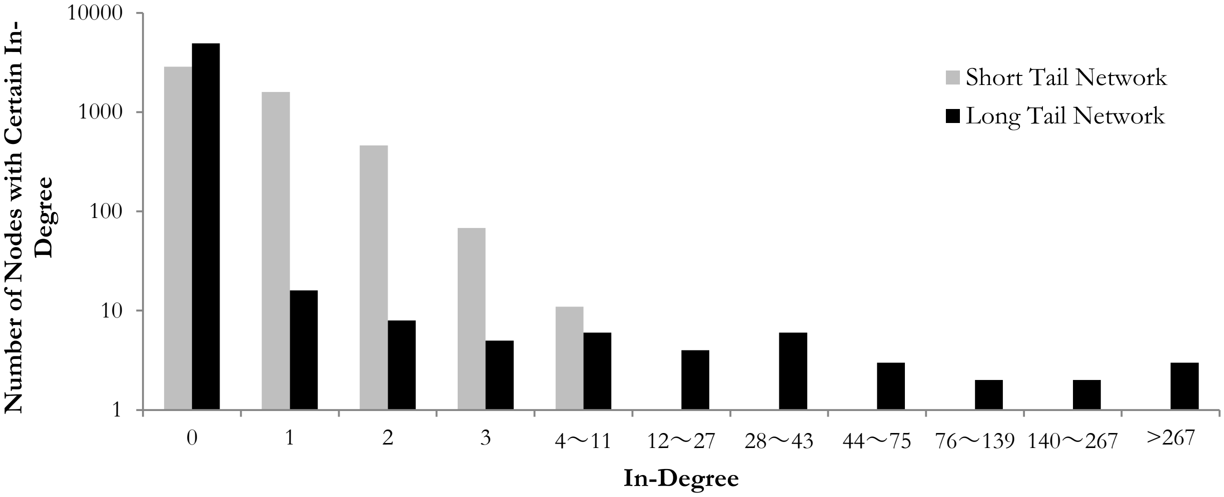
	T	$\log(\lambda_0)$	δ_1	δ_2	Distribution of Characteristics
Network 33	2000	-50	-2	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 34	2000	-50	-2	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 35	2000	-50	-3	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 36	2000	-50	-3	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 37	5000	-50	-2	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 38	5000	-50	-2	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 39	5000	-50	-3	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 40	5000	-50	-3	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z$
Network 41	2000	-50	-2	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 42	2000	-50	-2	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 43	2000	-50	-3	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 44	2000	-50	-3	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 45	5000	-50	-2	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 46	5000	-50	-2	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 47	5000	-50	-3	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 48	5000	-50	-3	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z1}$, $z_{2,j}$~$\text{Uniform}(-1,1) * \sigma_z + \mu_{z2}$
Network 49	2000	-50	-2	2	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j}$~$\text{Log-normal}(0,1) * \sigma_{z_2}$
Network 50	2000	-50	-2	3	$z_{1,j}$~$\text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j}$~$\text{Log-normal}(0,1) * \sigma_{z_2}$

Network 51	2000	-50	-3	2	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$
Network 52	2000	-50	-3	3	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$
Network 53	5000	-50	-2	2	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$
Network 54	5000	-50	-2	3	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$
Network 55	5000	-50	-3	2	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$
Network 56	5000	-50	-3	3	$z_{1,j} \sim \text{Uniform}(-1,1) * \sigma_{z_1}$, $z_{2,j} \sim \text{Log-normal}(0,1) * \sigma_{z_2}$

Table TA4: Statistics for the Non-Long-Tailed Networks Simulated

	Number of ties formed in the simulated network	Mean of In-Degree	Max of In-Degree	Max/Mean Ratio of In-Degree	Network density
Network 33	2087	1.0430	63	60.404	0.052%
Network 34	1774	0.8866	54	60.910	0.044%
Network 35	3259	1.6287	102	62.627	0.081%
Network 36	2839	1.4188	86	60.615	0.071%
Network 37	2254	0.4507	37	82.093	0.009%
Network 38	2021	0.4041	39	96.506	0.008%
Network 39	2490	0.4979	32	64.270	0.010%
Network 40	2530	0.5059	35	69.184	0.010%
Network 41	2176	1.087	13	11.955	0.054%
Network 42	2523	1.261	11	8.724	0.063%
Network 43	2861	1.430	15	10.491	0.071%
Network 44	3604	1.801	16	8.883	0.090%
Network 45	2786	0.557	9	16.155	0.011%
Network 46	3173	0.634	15	23.642	0.013%
Network 47	2991	0.598	18	30.096	0.012%
Network 48	2604	0.521	11	21.126	0.010%
Network 49	2932	1.465	128	87.356	0.073%
Network 50	2269	1.134	70	61.732	0.057%
Network 51	2036	1.017	58	57.030	0.051%
Network 52	3241	1.620	163	100.637	0.081%
Network 53	2491	0.498	245	491.869	0.010%
Network 54	2604	0.521	175	336.089	0.010%
Network 55	2862	0.572	221	386.171	0.011%
Network 56	3040	0.608	242	398.106	0.012%

Figure TA1: In-Degree Distributions for Representative Short- and Long-Tail Networks



For each target value of the sampling proportion (5%, 10%, 15% of the total dyads that do not form a tie) of the 24 networks, we repeatedly sample the network 25 times, and estimate the model on the 25 samples using the WESBI method. We report the average posterior means and the average posterior standard deviations of the parameter values recovered from the 25 runs, and check whether the true network generating parameters fall within the 95% credible intervals that are constructed using these values.

The 24 tables that follow show the estimation results, presented as before, from our simulation study. A check mark indicates that the true value of the parameter falls within the 95% credible interval that is constructed by using the average posterior mean and the average posterior standard deviation using estimates from the 25 runs. As before, the results show that sampling 10% or 15% of the dyads that do not form a tie gives very accurate estimation results with the true parameter value *always* falling in the credible interval. This shows that we can estimate the parameters with high accuracy by sampling a relatively small fraction of the total network and using the WESBI method for estimation.

Network 33:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0463 ✓ (0.0251)	-50.0293 ✓ (0.0221)	-50.0163 ✓ (0.0196)
δ_1	-2	-1.9669 ✓ (0.0173)	-1.9765 ✓ (0.0161)	-1.9811 ✓ (0.0154)
δ_2	2	1.9517 (0.0176)	1.9718 ✓ (0.0171)	1.9853 ✓ (0.0160)
$\Sigma_{\beta,11}$	0.5	0.5194 ✓ (0.0213)	0.5132 ✓ (0.0201)	0.5081 ✓ (0.0185)
$\Sigma_{\beta,12}$	0.25	0.2581 ✓ (0.0161)	0.2568 ✓ (0.0157)	0.2540 ✓ (0.0146)
$\Sigma_{\beta,22}$	0.5	0.5126 ✓ (0.0194)	0.5111 ✓ (0.0185)	0.5093 ✓ (0.0180)

Network 34:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9849 ✓ (0.0216)	-49.9903 ✓ (0.0204)	-49.9941 ✓ (0.0196)
δ_1	-2	-2.0133 ✓ (0.0183)	-2.0102 ✓ (0.0177)	-2.0075 ✓ (0.0164)
δ_2	3	3.0477 (0.0193)	3.0179 ✓ (0.0181)	3.0081 ✓ (0.0172)
$\Sigma_{\beta,11}$	0.5	0.5179 ✓ (0.0187)	0.5139 ✓ (0.0174)	0.5091 ✓ (0.0165)
$\Sigma_{\beta,12}$	0.25	0.2457 ✓ (0.0156)	0.2471 ✓ (0.0143)	0.2489 ✓ (0.0140)
$\Sigma_{\beta,22}$	0.5	0.5237 ✓ (0.0176)	0.5153 ✓ (0.0163)	0.5101 ✓ (0.0160)

Network 35:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9739 ✓ (0.0226)	-49.9801 ✓ (0.0215)	-49.9864 ✓ (0.0185)
δ_1	-3	-3.0089 ✓ (0.0165)	-3.0050 ✓ (0.0182)	-3.0020 ✓ (0.0169)
δ_2	2	2.0134 ✓ (0.0181)	2.0100 ✓ (0.0178)	2.0071 ✓ (0.0171)
$\Sigma_{\beta,11}$	0.5	0.4863 ✓ (0.0179)	0.4870 ✓ (0.0178)	0.4880 ✓ (0.0172)
$\Sigma_{\beta,12}$	0.25	0.2321 ✓ (0.0143)	0.2394 ✓ (0.0143)	0.2429 ✓ (0.0140)
$\Sigma_{\beta,22}$	0.5	0.4871 ✓ (0.0178)	0.4884 ✓ (0.0177)	0.4931 ✓ (0.0171)

Network 36:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9564 (0.0213)	-49.9799 ✓ (0.0202)	-49.9905 ✓ (0.0199)
δ_1	-3	-3.0465 (0.0179)	-3.0269 ✓ (0.0170)	-3.0067 ✓ (0.0169)
δ_2	3	3.0311 ✓ (0.0175)	3.0174 ✓ (0.0165)	3.0077 ✓ (0.0162)
$\Sigma_{\beta,11}$	0.5	0.5276 ✓ (0.0186)	0.5199 ✓ (0.0179)	0.5066 ✓ (0.0175)
$\Sigma_{\beta,12}$	0.25	0.2559 ✓ (0.0147)	0.2547 ✓ (0.0142)	0.2532 ✓ (0.0140)
$\Sigma_{\beta,22}$	0.5	0.5183 ✓ (0.0184)	0.5153 ✓ (0.0173)	0.5081 ✓ (0.0168)

Network 37:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0578 (0.0191)		-50.0279 (0.0191)	✓	-50.0105 (0.0185)	✓
δ_1	-2	-2.0369 (0.0121)		-2.0196 (0.0120)	✓	-2.0144 (0.0117)	✓
δ_2	2	2.0374 (0.0117)		2.0157 (0.0117)	✓	2.0038 (0.0112)	✓
$\Sigma_{\beta,11}$	0.5	0.5087 (0.0120)	✓	0.5052 (0.0118)	✓	0.4970 (0.0112)	✓
$\Sigma_{\beta,12}$	0.25	0.2589 (0.0107)	✓	0.2564 (0.0103)	✓	0.2544 (0.0098)	✓
$\Sigma_{\beta,22}$	0.5	0.5137 (0.0119)	✓	0.5082 (0.0114)	✓	0.5026 (0.0114)	✓

Network 38:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-49.9635 (0.0214)	✓	-49.9729 (0.0194)	✓	-49.9870 (0.0168)	✓
δ_1	-2	-2.0302 (0.0119)		-2.0165 (0.0115)	✓	-2.0076 (0.0110)	✓
δ_2	3	3.0291 (0.0121)		3.0167 (0.0116)	✓	3.0114 (0.0116)	✓
$\Sigma_{\beta,11}$	0.5	0.4873 (0.0123)	✓	0.4914 (0.0120)	✓	0.4966 (0.0118)	✓
$\Sigma_{\beta,12}$	0.25	0.2593 (0.0093)	✓	0.2565 (0.0093)	✓	0.2497 (0.0091)	✓
$\Sigma_{\beta,22}$	0.5	0.4895 (0.0117)	✓	0.4925 (0.0112)	✓	0.5008 (0.0112)	✓

Network 39:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9556 (0.0218)	-49.9761 ✓ (0.0210)	-50.0127 ✓ (0.0185)
δ_1	-3	-2.9719 (0.0124)	-2.9844 ✓ (0.0122)	-2.9860 ✓ (0.0116)
δ_2	2	2.0101 ✓ (0.0121)	2.0076 ✓ (0.0115)	1.9973 ✓ (0.0113)
$\Sigma_{\beta,11}$	0.5	0.5062 ✓ (0.0120)	0.5056 ✓ (0.0115)	0.5048 ✓ (0.0112)
$\Sigma_{\beta,12}$	0.25	0.2539 ✓ (0.0100)	0.2537 ✓ (0.0096)	0.2536 ✓ (0.0091)
$\Sigma_{\beta,22}$	0.5	0.5089 ✓ (0.0130)	0.5083 ✓ (0.0120)	0.5081 ✓ (0.0116)

Network 40:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0387 ✓ (0.0214)	-50.0164 ✓ (0.0189)	-50.0065 ✓ (0.0176)
δ_1	-3	-3.0068 ✓ (0.0119)	-3.0039 ✓ (0.0116)	-2.9984 ✓ (0.0115)
δ_2	3	3.0181 ✓ (0.0127)	3.0133 ✓ (0.0121)	3.0052 ✓ (0.0112)
$\Sigma_{\beta,11}$	0.5	0.5077 ✓ (0.0117)	0.5029 ✓ (0.0113)	0.4995 ✓ (0.0114)
$\Sigma_{\beta,12}$	0.25	0.2455 ✓ (0.0096)	0.2467 ✓ (0.093)	0.2474 ✓ (0.0093)
$\Sigma_{\beta,22}$	0.5	0.5210 ✓ (0.0117)	0.5171 ✓ (0.0112)	0.5096 ✓ (0.0108)

Network 41:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0202 ✓ (0.0240)	-50.0077 ✓ (0.0212)	-49.9949 ✓ (0.0188)
δ_1	-2	-2.0553 (0.0185)	-2.0158 ✓ (0.0169)	-2.0072 ✓ (0.0152)
δ_2	2	2.0312 ✓ (0.0175)	2.0196 ✓ (0.0171)	2.0122 ✓ (0.0170)
$\Sigma_{\beta,11}$	0.5	0.5178 ✓ (0.0202)	0.5098 ✓ (0.0197)	0.5015 ✓ (0.0159)
$\Sigma_{\beta,12}$	0.25	0.2416 ✓ (0.0158)	0.2422 ✓ (0.0158)	0.2461 ✓ (0.0152)
$\Sigma_{\beta,22}$	0.5	0.5093 ✓ (0.0192)	0.5039 ✓ (0.0191)	0.5010 ✓ (0.0181)

Network 42:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9548 (0.0213)	-49.9692 ✓ (0.0209)	-49.9956 ✓ (0.0204)
δ_1	-2	-2.0331 ✓ (0.0185)	-2.0277 ✓ (0.0163)	-2.0040 ✓ (0.0151)
δ_2	3	3.0519 (0.0200)	3.0310 ✓ (0.0192)	3.0146 ✓ (0.0183)
$\Sigma_{\beta,11}$	0.5	0.4829 ✓ (0.0182)	0.4882 ✓ (0.0180)	0.4928 ✓ (0.0174)
$\Sigma_{\beta,12}$	0.25	0.2534 ✓ (0.0161)	0.2513 ✓ (0.0157)	0.2489 ✓ (0.0151)
$\Sigma_{\beta,22}$	0.5	0.4929 ✓ (0.0172)	0.4942 ✓ (0.0159)	0.4962 ✓ (0.0152)

Network 43:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0537 (0.0237)	-50.0314 ✓ (0.0222)	-50.0140 ✓ (0.0199)
δ_1	-3	-3.0149 ✓ (0.0182)	-3.0098 ✓ (0.0182)	-3.0049 ✓ (0.0173)
δ_2	2	2.0470 (0.0173)	2.0031 ✓ (0.0171)	2.0020 ✓ (0.0162)
$\Sigma_{\beta,11}$	0.5	0.5136 ✓ (0.0182)	0.5084 ✓ (0.0181)	0.5017 ✓ (0.0176)
$\Sigma_{\beta,12}$	0.25	0.2539 ✓ (0.0138)	0.2537 ✓ (0.0134)	0.2521 ✓ (0.0131)
$\Sigma_{\beta,22}$	0.5	0.5024 ✓ (0.0176)	0.5012 ✓ (0.0175)	0.4997 ✓ (0.0168)

Network 44:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9439 (0.0225)	-49.9622 ✓ (0.0218)	-49.9877 ✓ (0.0193)
δ_1	-3	-2.9634 (0.0184)	-2.9764 ✓ (0.0179)	-2.9930 ✓ (0.0162)
δ_2	3	3.0214 ✓ (0.0180)	3.0112 ✓ (0.0168)	3.0040 ✓ (0.0167)
$\Sigma_{\beta,11}$	0.5	0.5176 ✓ (0.0175)	0.5072 ✓ (0.0171)	0.5048 ✓ (0.0170)
$\Sigma_{\beta,12}$	0.25	0.2580 ✓ (0.0144)	0.2514 ✓ (0.0137)	0.2497 ✓ (0.0135)
$\Sigma_{\beta,22}$	0.5	0.5075 ✓ (0.0182)	0.5044 ✓ (0.0179)	0.5031 ✓ (0.0178)

Network 45:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0437 (0.0188)	-50.0226 ✓ (0.0181)	-50.0077 ✓ (0.0173)
δ_1	-2	-2.0242 (0.0118)	-2.0149 ✓ (0.0113)	-2.0072 ✓ (0.0106)
δ_2	2	1.9714 (0.0120)	1.9824 ✓ (0.0114)	1.9940 ✓ (0.0114)
$\Sigma_{\beta,11}$	0.5	0.5156 ✓ (0.0120)	0.5072 ✓ (0.0117)	0.5030 ✓ (0.0114)
$\Sigma_{\beta,12}$	0.25	0.2610 ✓ (0.0102)	0.2603 ✓ (0.0102)	0.2578 ✓ (0.0098)
$\Sigma_{\beta,22}$	0.5	0.5047 ✓ (0.0121)	0.5029 ✓ (0.0114)	0.5022 ✓ (0.0114)

Network 46:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9703 ✓ (0.0210)	-49.9841 ✓ (0.0192)	-49.9931 ✓ (0.0190)
δ_1	-2	-1.9740 (0.0117)	-1.9833 ✓ (0.0110)	-1.9863 ✓ (0.0110)
δ_2	3	3.0214 ✓ (0.0115)	3.0068 ✓ (0.0109)	3.0037 ✓ (0.0107)
$\Sigma_{\beta,11}$	0.5	0.4911 ✓ (0.0118)	0.4932 ✓ (0.0115)	0.5006 ✓ (0.0115)
$\Sigma_{\beta,12}$	0.25	0.2685 ✓ (0.0101)	0.2623 ✓ (0.0096)	0.2535 ✓ (0.0095)
$\Sigma_{\beta,22}$	0.5	0.4874 ✓ (0.0119)	0.4886 ✓ (0.0117)	0.4940 ✓ (0.0114)

Network 47:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-50.0393 (0.0185)	✓	-50.0309 (0.0176)	✓	-50.0271 (0.0156)	✓
δ_1	-3	-2.9951 (0.0112)	✓	-2.9954 (0.0110)	✓	-2.9975 (0.0108)	✓
δ_2	2	2.0156 (0.0112)	✓	2.0144 (0.0110)	✓	2.0144 (0.0109)	✓
$\Sigma_{\beta,11}$	0.5	0.4800 (0.0121)	✓	0.4801 (0.0114)	✓	0.4823 (0.0114)	✓
$\Sigma_{\beta,12}$	0.25	0.2293 (0.0098)		0.2383 (0.0094)	✓	0.2411 (0.0087)	✓
$\Sigma_{\beta,22}$	0.5	0.4889 (0.0120)	✓	0.4891 (0.0115)	✓	0.4893 (0.0112)	✓

Network 48:

	True value	5% of non-formed dyads sampled		10% of non-formed dyads sampled		15% of non-formed dyads sampled	
$log\lambda_0$	-50	-49.8488 (0.0207)	✓	-49.8961 (0.0193)	✓	-49.9916 (0.0178)	✓
δ_1	-3	-3.0358 (0.0118)		-3.0183 (0.0114)	✓	-3.0057 (0.0111)	✓
δ_2	3	3.0049 (0.0122)	✓	3.0037 (0.0118)	✓	3.0030 (0.0114)	✓
$\Sigma_{\beta,11}$	0.5	0.4894 (0.0114)	✓	0.4988 (0.0111)	✓	0.5005 (0.0110)	✓
$\Sigma_{\beta,12}$	0.25	0.2603 (0.0100)	✓	0.2573 (0.097)	✓	0.2549 (0.0094)	✓
$\Sigma_{\beta,22}$	0.5	0.4883 (0.0113)	✓	0.4908 (0.0110)	✓	0.4924 (0.0109)	✓

Network 49:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0377 ✓ (0.0233)	-50.0258 ✓ (0.0222)	-50.0088 ✓ (0.0207)
δ_1	-2	-2.0436 (0.0184)	-2.0139 ✓ (0.0173)	-2.0054 ✓ (0.0163)
δ_2	2	2.0205 ✓ (0.0181)	2.0148 ✓ (0.0170)	2.0041 ✓ (0.0167)
$\Sigma_{\beta,11}$	0.5	0.5207 ✓ (0.0197)	0.5163 ✓ (0.0188)	0.5034 ✓ (0.0172)
$\Sigma_{\beta,12}$	0.25	0.2630 ✓ (0.0173)	0.2588 ✓ (0.0164)	0.2510 ✓ (0.0159)
$\Sigma_{\beta,22}$	0.5	0.5129 ✓ (0.0188)	0.5053 ✓ (0.0175)	0.5037 ✓ (0.0171)

Network 50:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0422 ✓ (0.0241)	-50.0319 ✓ (0.0212)	-50.0186 ✓ (0.0203)
δ_1	-2	-1.9640 ✓ (0.0193)	-1.9811 ✓ (0.0186)	-1.9866 ✓ (0.0171)
δ_2	3	3.0311 ✓ (0.0186)	3.0097 ✓ (0.0171)	3.0056 ✓ (0.0167)
$\Sigma_{\beta,11}$	0.5	0.5108 ✓ (0.0163)	0.5055 ✓ (0.161)	0.5042 ✓ (0.158)
$\Sigma_{\beta,12}$	0.25	0.2396 ✓ (0.0157)	0.2416 ✓ (0.0152)	0.2424 ✓ (0.0140)
$\Sigma_{\beta,22}$	0.5	0.5088 ✓ (0.0176)	0.5040 ✓ (0.0164)	0.5024 ✓ (0.0149)

Network 51:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0119 ✓ (0.0214)	-50.0096 ✓ (0.0209)	-50.0057 ✓ (0.0200)
δ_1	-3	-3.0269 ✓ (0.0185)	-3.0124 ✓ (0.0172)	-3.0116 ✓ (0.0168)
δ_2	2	2.0304 ✓ (0.0182)	2.0085 ✓ (0.0169)	2.0038 ✓ (0.0162)
$\Sigma_{\beta,11}$	0.5	0.4933 ✓ (0.0190)	0.4947 ✓ (0.0184)	0.4988 ✓ (0.0181)
$\Sigma_{\beta,12}$	0.25	0.2335 ✓ (0.0136)	0.2349 ✓ (0.0134)	0.2379 ✓ (0.0129)
$\Sigma_{\beta,22}$	0.5	0.5213 ✓ (0.0183)	0.5135 ✓ (0.0172)	0.5096 ✓ (0.0171)

Network 52:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9509 (0.0219)	-49.9781 ✓ (0.0209)	-49.9825 ✓ (0.0197)
δ_1	-3	-3.0582 (0.0186)	-3.0272 ✓ (0.0184)	-3.0150 ✓ (0.0175)
δ_2	3	2.9538 (0.0182)	2.9771 ✓ (0.0175)	2.9935 ✓ (0.0171)
$\Sigma_{\beta,11}$	0.5	0.5229 ✓ (0.0182)	0.5127 ✓ (0.0172)	0.5083 ✓ (0.0169)
$\Sigma_{\beta,12}$	0.25	0.2311 ✓ (0.0153)	0.2426 ✓ (0.0132)	0.2452 ✓ (0.0120)
$\Sigma_{\beta,22}$	0.5	0.5079 ✓ (0.0183)	0.5071 ✓ (0.0181)	0.5022 ✓ (0.0178)

Network 53:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0293 ✓ (0.0194)	-50.0112 ✓ (0.0185)	-50.0030 ✓ (0.0175)
δ_1	-2	-1.9786 (0.0117)	-1.9879 ✓ (0.0113)	-1.9923 ✓ (0.0109)
δ_2	2	2.0351 (0.0113)	2.0183 ✓ (0.0112)	2.0086 ✓ (0.0107)
$\Sigma_{\beta,11}$	0.5	0.4823 ✓ (0.0118)	0.4905 ✓ (0.0118)	0.4926 ✓ (0.0116)
$\Sigma_{\beta,12}$	0.25	0.2607 ✓ (0.0097)	0.2566 ✓ (0.0095)	0.2521 ✓ (0.0093)
$\Sigma_{\beta,22}$	0.5	0.4861 ✓ (0.0115)	0.4946 ✓ (0.0112)	0.4962 ✓ (0.0111)

Network 54:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-49.9657 ✓ (0.0206)	-49.9872 ✓ (0.0197)	-49.9940 ✓ (0.0191)
δ_1	-2	-2.0325 (0.0116)	-2.0184 ✓ (0.0115)	-2.0065 ✓ (0.0111)
δ_2	3	2.9833 ✓ (0.0111)	2.9844 ✓ (0.0108)	2.9953 ✓ (0.0106)
$\Sigma_{\beta,11}$	0.5	0.5112 ✓ (0.0115)	0.5076 ✓ (0.0109)	0.5025 ✓ (0.0106)
$\Sigma_{\beta,12}$	0.25	0.2318 ✓ (0.0098)	0.2428 ✓ (0.0093)	0.2454 ✓ (0.0092)
$\Sigma_{\beta,22}$	0.5	0.5046 ✓ (0.0116)	0.5028 ✓ (0.0114)	0.5013 ✓ (0.0114)

Network 55:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0417 (0.0201)	-50.0283 ✓ (0.0178)	-50.0171 ✓ (0.0163)
δ_1	-3	-2.9899 ✓ (0.0115)	-3.0047 ✓ (0.0111)	-2.9978 ✓ (0.0107)
δ_2	2	1.9742 (0.0113)	1.9943 ✓ (0.0109)	2.0016 ✓ (0.0108)
$\Sigma_{\beta,11}$	0.5	0.5167 ✓ (0.0118)	0.5069 ✓ (0.0117)	0.5011 ✓ (0.0114)
$\Sigma_{\beta,12}$	0.25	0.2404 ✓ (0.0101)	0.2434 ✓ (0.0094)	0.2441 ✓ (0.0091)
$\Sigma_{\beta,22}$	0.5	0.4905 ✓ (0.0117)	0.4926 ✓ (0.0113)	0.4949 ✓ (0.0108)

Network 56:

	True value	5% of non-formed dyads sampled	10% of non-formed dyads sampled	15% of non-formed dyads sampled
$log\lambda_0$	-50	-50.0397 ✓ (0.0206)	-50.0254 ✓ (0.0196)	-50.0047 ✓ (0.0181)
δ_1	-3	-3.0083 ✓ (0.0114)	-2.9973 ✓ (0.0112)	-2.9997 ✓ (0.0109)
δ_2	3	2.9628 (0.0119)	2.9813 ✓ (0.0114)	2.9910 ✓ (0.0109)
$\Sigma_{\beta,11}$	0.5	0.5065 ✓ (0.0116)	0.4971 ✓ (0.0114)	0.4992 ✓ (0.0112)
$\Sigma_{\beta,12}$	0.25	0.2425 ✓ (0.0103)	0.2482 ✓ (0.098)	0.2514 ✓ (0.0093)
$\Sigma_{\beta,22}$	0.5	0.5124 ✓ (0.0112)	0.5049 ✓ (0.0107)	0.5019 ✓ (0.0104)

Overall Conclusions

We can make the following conclusions from the simulations:

- By sampling all the dyads that form ties and 10% (or 15%) of the dyads that do not form ties, we can accurately estimate parameter values using the WESBI method.
- The average posterior standard deviation in parameter estimates grows smaller as we sample more data.
- For both long- and short-tailed networks, parameter estimation using the WESBI method is equally good, and we observe the same patterns as we sample more data.

Estimation Time Advantage of the WESBI method

To assess the estimation time advantage of the WESBI method, we compare the average time taken by one iteration of the Bayesian inference procedure when the full dataset is used and when smaller sampled datasets are used. In the algorithm, the difference between the time taken by one iteration of the Bayesian inference procedure is very small regardless of whether the iteration belongs to burn-in phase or after the chains have converged. (The time taken is always within ±9.3% of the average time taken by one iteration.) Thus we take the average across the first 1,000 iterations in the estimation procedure. In Table TA5, we report the average time taken in seconds for completing the calculations of one iteration in the MCMC procedure of estimation from the full dataset and three different sampled datasets (sampling proportions are 5%, 10%, 15%, respectively) for each of the 56 networks we generated above. As we stated, the variance in time taken across iterations is small within each sampled dataset, so the reported numbers are suitable for comparison. The numbers in parentheses in the last three columns denote the percentage time taken as compared to using the full dataset.

Table TA5: Iteration Times for WESBI for Different Sampling Proportions

	Time per iteration with full dataset	Time per iteration when 5% of non-tie dyads are sampled	Time per iteration when 10% of non-tie dyads are sampled	Time per iteration when 15% of non-tie dyads are sampled
Network 1	5.33	0.65 (12.3%)	0.88 (16.4%)	1.08 (20.2%)
Network 2	5.38	0.65 (12.1%)	0.88 (16.4%)	1.07 (20.0%)
Network 3	5.32	0.66 (12.4%)	0.88 (16.4%)	1.07 (20.1%)
Network 4	5.22	0.65 (12.4%)	0.86 (16.6%)	1.05 (20.0%)
Network 5	5.19	0.64 (12.3%)	0.84 (16.2%)	1.03 (19.8%)
Network 6	5.37	0.66 (12.3%)	0.88 (16.4%)	1.08 (20.1%)
Network 7	5.14	0.64 (12.4%)	0.84 (16.3%)	1.01 (19.7%)
Network 8	5.36	0.66 (12.3%)	0.88 (16.4%)	1.08 (20.1%)
Network 9	30.03	4.01 (13.4%)	5.28 (17.6%)	6.33 (21.1%)
Network 10	29.71	4.06 (13.7%)	5.10 (17.1%)	6.02 (20.3%)
Network 11	29.79	3.91 (13.1%)	5.16 (17.3%)	6.21 (20.8%)
Network 12	29.67	3.93 (13.3%)	5.03 (16.9%)	6.04 (20.4%)
Network 13	29.91	3.89 (13.0%)	5.25 (17.6%)	6.26 (20.9%)
Network 14	29.58	3.93 (13.3%)	4.94 (16.7%)	5.86 (19.8%)
Network 15	30.02	3.78 (12.6%)	5.30 (17.6%)	6.41 (21.3%)
Network 16	29.90	3.98 (13.3%)	5.19 (17.4%)	6.17 (20.6%)
Network 17	5.31	0.65 (12.3%)	0.88 (16.6%)	1.07 (20.2%)
Network 18	5.26	0.64 (12.2%)	0.87 (16.5%)	1.05 (19.9%)
Network 19	5.26	0.65 (12.3%)	0.86 (16.4%)	1.06 (20.2%)
Network 20	5.30	0.65 (12.3%)	0.87 (16.4%)	1.06 (19.9%)
Network 21	5.34	0.65 (12.2%)	0.87 (16.4%)	1.06 (19.9%)
Network 22	5.25	0.65 (12.4%)	0.87 (16.5%)	1.05 (20.1%)
Network 23	5.20	0.64 (12.3%)	0.85 (16.4%)	1.03 (19.8%)
Network 24	5.26	0.64 (12.2%)	0.85 (16.2%)	1.04 (19.9%)
Network 25	29.84	4.05 (13.6%)	5.14 (17.2%)	6.15 (20.6%)
Network 26	29.92	3.94 (13.2%)	5.18 (17.3%)	6.25 (20.9%)
Network 27	30.09	3.93 (13.1%)	5.33 (17.7%)	6.46 (21.5%)
Network 28	29.69	3.93 (13.3%)	5.02 (16.9%)	5.95 (20.1%)
Network 29	29.84	3.95 (13.2%)	5.15 (17.2%)	6.16 (20.6%)
Network 30	29.62	3.89 (13.1%)	5.05 (17.0%)	5.96 (20.1%)
Network 31	30.23	3.81 (12.6%)	5.45 (18.0%)	6.56 (21.7%)
Network 32	29.87	3.91 (13.1%)	5.14 (17.2%)	6.13 (20.5%)
Network 33	5.27	0.65 (12.3%)	0.86 (16.3%)	1.04 (19.8%)

Network 34	5.16	0.63 (12.3%)	0.84 (16.3%)	1.01 (19.6%)
Network 35	5.41	0.67 (12.3%)	0.90 (16.6%)	1.10 (20.3%)
Network 36	5.34	0.66 (12.3%)	0.89 (16.6%)	1.08 (20.2%)
Network 37	29.90	3.88 (13.0%)	5.21 (17.4%)	6.23 (20.8%)
Network 38	29.65	4.05 (13.7%)	5.03 (17.0%)	5.94 (20.0%)
Network 39	29.56	3.92 (13.3%)	4.99 (16.9%)	5.92 (20.0%)
Network 40	29.79	4.06 (13.6%)	5.12 (17.2%)	6.10 (20.5%)
Network 41	5.21	0.64 (12.3%)	0.85 (16.3%)	1.04 (20.0%)
Network 42	5.34	0.66 (12.3%)	0.88 (16.4%)	1.08 (20.2%)
Network 43	5.36	0.66 (12.2%)	0.89 (16.5%)	1.07 (20.0%)
Network 44	5.47	0.66 (12.2%)	0.90 (16.5%)	1.11 (20.3%)
Network 45	30.20	3.87 (12.8%)	5.35 (17.7%)	6.51 (21.6%)
Network 46	29.82	3.84 (12.9%)	5.15 (17.3%)	6.23 (20.9%)
Network 47	29.99	3.91 (13.0%)	5.26 (17.5%)	6.36 (21.2%)
Network 48	29.97	3.91 (13.0%)	5.25 (17.5%)	6.27 (20.9%)
Network 49	5.42	0.66 (12.3%)	0.89 (16.4%)	1.08 (20.0%)
Network 50	5.24	0.65 (12.4%)	0.87 (16.5%)	1.04 (19.8%)
Network 51	5.13	0.63 (12.3%)	0.84 (16.3%)	1.02 (19.9%)
Network 52	5.34	0.66 (12.3%)	0.88 (16.5%)	1.09 (20.5%)
Network 53	30.12	4.01 (13.3%)	5.37 (17.8%)	6.45 (21.4%)
Network 54	29.73	3.97 (13.3%)	5.09 (17.1%)	6.04 (20.3%)
Network 55	29.78	3.94 (13.2%)	5.11 (17.2%)	6.11 (20.5%)
Network 56	29.91	3.94 (13.2%)	5.25 (17.6%)	6.29 (21.0%)

References

Barabasi, Albert-Laszlo and Reka Albert. 1999. "Emergence of Scaling in Random Networks." *Science*. 286(5439) 509-512.

Mislove, Alan, Massimiliano Marcon, Krishna Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. "Measurement and Analysis of Online Social Networks." *IMC'07*, San Diego, CA, USA.

Technical Appendix II

Random Coefficients Model

Model Estimation

We extend our basic model to include individual-level heterogeneity. First, to capture unobserved heterogeneity in the baseline hazard rates across reviewers, we allow the parameter α_1 of the baseline hazard function to vary across senders using a log-normal distribution in the following way: $\lambda_{0,i}(t) = \alpha_0 \alpha_{1i} t^{\alpha_{1i}-1}$ and $\log(\alpha_{1,i}) \sim \mathrm{N}(\overline{\alpha_1}, \sigma_\alpha^2)$, where *i* is the index over senders. (Note that the heterogeneity in α_0 is absorbed by a_i, the sender-specific random effect.) Second, heterogeneity may exist because the same covariates may have different impacts on different reviewers' propensities to form trust relationships. To control for this, we allow for heterogeneity in the coefficients as follows: $\begin{bmatrix} \boldsymbol{\beta}_i^i \\ \boldsymbol{\beta}_i^j \\ \boldsymbol{\beta}_i^{ij} \end{bmatrix} = \boldsymbol{\beta}_i = \boldsymbol{\delta} + \boldsymbol{\varepsilon_i},\ \boldsymbol{\varepsilon_i} \sim MVN(0, \boldsymbol{\Sigma}_\beta)$. The notation used here is similar to that used in the homogenous model in Section 3 of the paper. Let $\mathrm{C_{ij}}$ be the number of time periods for which dyad ij has been observed, and $\mathrm{T_{ij}}$ be the length of time from the starting point to the time period when i extends a tie to j. We define $\mathbb{I}_{\mathrm{ij}} = 1$ if $\mathrm{T_{ij}} \leq \mathrm{C_{ij}}$ (i.e., if a tie formed within the observation time) and 0 otherwise, and $\mathrm{k_{ij}} = \mathrm{floor}\left(\min\{\mathrm{T_{ij}}, \mathrm{C_{ij}}\}\right)$. The log-conditional-likelihood function for this formulation is given by:

$$\log L = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \Bigg\{ \mathbb{I}_{ij} \log\left[1 - \exp\left\{-\exp\left[\alpha_i(k_{ij}) + \boldsymbol{z}_{ij,k_{ij}}\boldsymbol{\beta_i} + \boldsymbol{a_i} + \boldsymbol{b_j} + \boldsymbol{d_{ij}}\right]\right\}\right]$$

$$- \sum_{t=0}^{k_{ij}-1} \exp[\alpha_i(t) + \boldsymbol{z}_{ijt}\boldsymbol{\beta_i} + a_i + b_j + d_{ij}] \Bigg\},$$

where $\alpha_i(t) = \ln\left(\int_t^{t+1} \lambda_{0,i}(u)du\right)$. The results for the model with heterogeneity are provided in the table below. We find that the impact of preferential attachment and recency are qualitatively the same as in the model with homogenous individuals.

Parameter Estimates for the "Movies" Category with Heterogeneous Coefficients

Variables	Posterior Mean	Posterior Std Deviation across Individuals
Receiver Characteristics		
Receiver's PrevAggReview	0.0774	0.5961***
Receiver's CurReview	0.4193***	0.2916***
Receiver's AggOpnLeadership	0.1789***	0.2587***
Receiver's CurOpnLeadership	0.3045***	0.4966***
Comprehensiveness	0.2351***	0.2437***
Objectivity	0.1157	0.2658***
Readability	0.1480	0.3049***
(Comprehensiveness)^2	-0.2223***	0.5855***
(Objectivity)^2	-0.0873***	0.1825***
(Readability)^2	-0.3301***	0.4390***
Top Reviewer Label	0.1968***	0.3546***
Sender Characteristics		
Sender's AggReview	0.1477***	0.4048***
Sender's AggOutgoingLink	0.0888***	0.2001***
Dyad Characteristics		
Dissimilarity in Comprehnsiveness	-0.1445***	0.2035***
Dissimilarity in Objectivity	-0.1003**	0.1900***
Dissimilarity in Readability	-0.0739	0.6119***
Reciprocity	0.1941***	0.2138***
Number of Commonly Trusted Reviewers	0.1672***	0.4610***
Hazard Rate Parameters		
Log(α_0)	-14.7143***	
$\overline{\alpha_1}$	-6.0919***	
σ_α^2	0.4977***	
σ_d^2	0.2078***	
σ_a^2	0.6080***	
σ_b^2	0.5282***	
$\boldsymbol{\sigma_{ab}}$	0.1379***	

***, ** and * denote that the 99% credible interval, the 95% credible interval, and the 90% credible interval, respectively, does not include zero.

Estimation Procedure

For the procedure described below, letters with superscript u represent the values of the corresponding updated parameters.

Step 1: $\boldsymbol{\beta}_i^u|\boldsymbol{\delta}, \Sigma_{\boldsymbol{\beta}}, a_i, b_i, \alpha_0, \alpha_{1,i}, d_{ij}, \text{data}$

$$f(\boldsymbol{\beta}_i^u|\boldsymbol{\delta}, \Sigma_{\boldsymbol{\beta}}, a_i, b_i, \alpha_0, \alpha_{1,i}, d_{ij}, \text{data})$$

$$\propto \mathrm{N}\left(\left(\boldsymbol{\beta}_i^u|\boldsymbol{\delta}, a_i, b_i, \alpha_0, \alpha_{1,i}, d_{ij}\right), \Sigma_{\boldsymbol{\beta}}\right) L(\boldsymbol{Y})$$

$$\propto \left|\Sigma_{\boldsymbol{\beta}}\right|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(\boldsymbol{\beta}_i^u - \boldsymbol{\delta})'\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_i^u - \boldsymbol{\delta})\right] L(\boldsymbol{Y})$$

where $L(\boldsymbol{Y})$ is the likelihood function. Since this distribution does not have a closed form, we use the Metropolis-Hastings algorithm to draw from the conditional distribution of $\boldsymbol{\beta_i}$. $\boldsymbol{\beta_i}$ is the draw of coefficients from the previous iteration, and we draw $\boldsymbol{\beta}_i^u$ by $\boldsymbol{\beta}_i^u = \boldsymbol{\beta_i} + \Delta\boldsymbol{\beta}$, where $\Delta\boldsymbol{\beta}$ is a draw from $\mathrm{N}(\mathbf{0}, \Delta^2\Lambda)$, and Δ and Λ are chosen adaptively to reduce the autocorrelation among the MCMC draws following Atchade (2006). The probability of accepting this $\boldsymbol{\beta}_i^u$, the updated value for $\boldsymbol{\beta_i}$ is:

$$\text{Pr(acceptance)} = \min\{\frac{[\exp(-\frac{1}{2}\left((\boldsymbol{\beta}_i^u - \boldsymbol{\delta})'\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_i^u - \boldsymbol{\delta})\right)]L(\boldsymbol{Y}|\boldsymbol{\beta}_i^u)}{[\exp(-\frac{1}{2}\left((\boldsymbol{\beta}_i \ - \boldsymbol{\delta})'\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}_i \ - \boldsymbol{\delta})\right)]L(\boldsymbol{Y}|\boldsymbol{\beta}_i \ )}, 1\}$$

Step 2: $\boldsymbol{\delta}^u|\Sigma_{\boldsymbol{\beta}}, \boldsymbol{\beta}_i^u$

$\boldsymbol{\delta}^u$ is generated from the distribution $\mathrm{MVN}(\boldsymbol{\mu_\delta}, \boldsymbol{V_\delta})$, where $\boldsymbol{\mu_\delta} = \boldsymbol{V_\delta}\left[\boldsymbol{\Sigma}_\beta^{-1}\sum_{i=1}^N \boldsymbol{\beta}_i^u + V_0^{-1}U_0\right]$, $\boldsymbol{V_\delta} = (\boldsymbol{N}\boldsymbol{\Sigma}_\beta^{-1} + V_0^{-1})^{-\mathbf{1}}$. We define diffuse priors by setting $V_0 = 100I$ and $U_0 = 0$.

Step 3: $\Sigma_{\boldsymbol{\beta}}^u|\boldsymbol{\beta}_i^u, \boldsymbol{\delta}^u$

$$(\Sigma_{\boldsymbol{\beta}}^u|\boldsymbol{\beta}_i^u, \boldsymbol{\delta}^u) \sim \mathrm{IW}_{n(\beta)}(f_0 + N, \boldsymbol{G}_0^{-1} + \sum_{i=1}^N (\boldsymbol{\beta}_i^u - \boldsymbol{\delta}^u)(\boldsymbol{\beta}_i^u - \boldsymbol{\delta}^u)')$$

where we set $f_0 = n(\beta) + 5$ and $\boldsymbol{G_0} = I_{n(\beta)}$ to be diffuse hyperpriors. f_0 is the degrees of freedom, $\boldsymbol{G_0}$is the scale matrix of the inverse-Wishart distribution, and $n(\beta)$ is the number of δ parameters, the ones before observed covariates that we are interested in.

Step 4: $\alpha_{1,i}^{u}|\boldsymbol{\beta}_i^u, a_i, b_i, \alpha_0, \overline{\alpha_1}, \sigma_\alpha^2, d_{ij}$, data

We can define the distribution of $\alpha_{1,i}^{u}$ as:

$$f\left(\alpha_{1,i}^{u}\middle|\boldsymbol{\beta}_i^u, a_i, b_i, \alpha_0, \overline{\alpha_1}, \sigma_\alpha^2, d_{ij}, \text{data}\right)$$

$$\propto \mathrm{N}((\alpha_{1,i}^{u}|\boldsymbol{\beta}_i^u, a_i, b_i, \alpha_0, \overline{\alpha_1}, d_{ij}), \sigma_\alpha^2)L(\boldsymbol{Y})$$

$$\propto \sigma_\alpha \exp\left[-\tfrac{1}{2}\left(\alpha_{1,i}^{u} - \overline{\alpha_1}\right)^2 \sigma_\alpha^{-2}\right] L(\boldsymbol{Y})$$

We use the Metropolis-Hastings algorithm to draw from the conditional distribution of a_i, a_i is the draw of coefficients from the previous iteration, and we draw $\alpha_{1,i}^{u}$ according to $\alpha_{1,i}^{u} = \alpha_{1,i} + \Delta\alpha$, where $\Delta\alpha$ is a draw from N(0,$\Delta^2\Lambda$), and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006),. The acceptance probability is:

$$\Pr(\text{acceptance}) = \min\{\frac{\left[\exp\left(-\frac{1}{2}\left(\alpha_{1,i}^{u} - \overline{\alpha_1}\right)^2 \sigma_\alpha^{-2}\right)\right] L\left(\boldsymbol{Y}\middle|\alpha_{1,i}^{u}\right)}{\left[\exp\left(-\frac{1}{2}\left(\alpha_{1,i} - \overline{\alpha_1}\right)^2 \sigma_\alpha^{-2}\right)\right] L\left(\boldsymbol{Y}\middle|\alpha_{1,i}\right)}, 1\}$$

Step 5: $\overline{\alpha_1}^u|\alpha_{1,i}^{u}, \sigma_\alpha^2$, data

$\overline{\alpha_1}^u$ is generated from a distribution $\mathrm{N}(\mu_\alpha, \nu_\alpha)$, where $\mu_\alpha = \nu_\alpha\left[\sigma_\alpha^{-2}\sum_{i=1}^{N}\alpha_{1,i}^{u} + \nu_{\alpha_0}^{-1}U_0\right]$, $\nu_\alpha = (N\sigma_\alpha^{-2} + \nu_{\alpha_0}^{-1})^{-1}$. We define diffuse priors by setting $\nu_{\alpha 0} = 100$ and $U_0 = 0$.

Step 6: $(\sigma_\alpha^2)^u|\overline{\alpha_1}^u, \alpha_{1,i}^{u}$

$$\left((\sigma_\alpha^2)^u\middle|\overline{\alpha_1}^u, \alpha_{1,i}^{u}\right) \sim \text{InverseGamma}(f_0 + N, g_0^{-1} + \sum_{i=1}^{N}(\alpha_{1,i}^{u} - \overline{\alpha_1}^u)^2)$$

where we set $f_0 = 6$ and $g_0 = 1$ to be diffuse hyperprior. f_0 is the degrees of freedom, g_0 is the scale matrix of the inverse Gamma distribution.

Step 7: $\alpha_0^u|\boldsymbol{\beta}_i^u, a_i, b_i, \alpha_{1,i}^{u}, d_{ij}$, data

$$f(\alpha_0^u|\boldsymbol{\beta}_i^u, a_i, b_i, \alpha_{1,i}^{u}, d_{ij}, \text{data}) \propto \sigma_{\alpha_0}^{-1}\exp[-\frac{1}{2}(\alpha_0^u - \overline{\alpha_0})^2\sigma_{\alpha_0}^{-2}]L(\boldsymbol{Y})$$

where $\overline{\alpha_0}$ and $\sigma_{\alpha 0}^2$ are diffuse priors. Because there is no closed form for this, we use the Metropolis-Hastings algorithm to draw from this conditional distribution of α_0^u . The probability of accepting α_0^u is:

$$\Pr(\text{acceptance}) = \min\{\frac{\left[\exp\left(-\frac{1}{2}(\alpha_0^u - \overline{\alpha_0})^2 \sigma_{\alpha_0}^{-2}\right)\right] L(\boldsymbol{Y}|\alpha_0^u)}{\left[\exp\left(-\frac{1}{2}\left(\alpha_0 \ - \overline{\alpha_0}\right)^2 \sigma_{\alpha_0}^{-2}\right)\right] L(\boldsymbol{Y}|\alpha_0 \ )}, 1\}$$

We define diffuse priors by setting $\overline{\alpha_0} = 0$ and $\sigma_{\alpha_0}^2$=30.

Step 8: Generate a_i^u, b_i^u:

$$f(a_i^u, b_i^u | \boldsymbol{\beta}_i^u, \Sigma_{ab}, \alpha_0^u, \alpha_{1,i}^u, d_{ij}, \text{data})$$

$$\propto \mathrm{N}\left(\left(a_i^u, b_i^u | \boldsymbol{\beta}_i^u, \alpha_0^u, \alpha_{1,i}^u, d_{ij}\right), \Sigma_{ab}\right) L(\boldsymbol{Y})$$

$$\propto |\Sigma_{ab}|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right] L(\boldsymbol{Y})$$

Because this distribution does not have a closed form, we use the Metropolis-Hastings algorithm to draw from the conditional distribution of a_i, b_i: a_i, b_i is the draw of the random effect from the previous iteration, and we draw a_i^u, b_i^u by $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix} = \begin{bmatrix} a_i \\ b_i \end{bmatrix} + \Delta \begin{bmatrix} a \\ b \end{bmatrix}$, where $\Delta \begin{bmatrix} a \\ b \end{bmatrix}$ is a draw from $\mathrm{N}(0, \Delta^2 \Lambda)$, and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006). The probability of accepting this $\begin{bmatrix} a_i^u \\ b_i^u \end{bmatrix}$, the updated value for $\begin{bmatrix} a_i \\ b_i \end{bmatrix}$ is:

$$\Pr(\text{acceptance}) = \min\{\frac{\left[\exp\left(-\frac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right)\right] L(\boldsymbol{Y}|a_i^u, b_i^u)}{\left[\exp\left(-\frac{1}{2}\left(a_i \ , b_i \ \right)\Sigma_{ab}^{-1}\left(a_i \ , b_i \ \right)'\right)\right] L(\boldsymbol{Y}|a_i \ , b_i \ )}, 1\}$$

Step 9: $\Sigma_{\mathrm{ab}}^{\mathrm{u}} | a_i^u, b_i^u$

$$(\Sigma_{\mathrm{ab}}^{\mathrm{u}} | a_i^u, b_i^u) \sim IW_2(7 + N, G_0^{-1} + \sum_{i=1}^{N} (a_i^u, b_i^u)(a_i^u, b_i^u)')$$

Step 10: $d_{ij}^u, d_{ji}^u | \alpha_0^u, \boldsymbol{\beta}_i^u, a_i, b_i, \alpha_{1,i}^u, \sigma_d^2$, data

$$f(d_{ij}^u, d_{ji}^u | \alpha_0^u, \boldsymbol{\beta}_i^u, a_i, b_i, \alpha_{1,i}^u, \sigma_d^2, \text{data})$$

$$\propto \mathrm{N}\left(\left(d_{ij}^u, d_{ji}^u | \alpha_0^u, \boldsymbol{\beta}_i^u, a_i, b_i, \alpha_{1,i}^u\right), \sigma_d^2\right) L(\boldsymbol{Y})$$

$$\propto \sigma_d^{-1} \exp\left[-\tfrac{1}{2}\left(d_{ij}^u + d_{ji}^u\right)^2 \sigma_d^{-2}\right] L(\boldsymbol{Y})$$

We use the Metropolis-Hastings algorithm to draw from this conditional distribution of d_{ij}^{u} and d_{ji}^{u}: d_{ij} and d_{ji} are the draw of the unobservable similarity effects from the previous iteration, and we draw d_{ij}^{u}, d_{ji}^{u} by $\begin{bmatrix} d_{ij}^{u} \\ d_{ji}^{u} \end{bmatrix} = \begin{bmatrix} d_{ij} \\ d_{ji} \end{bmatrix} + \Delta\boldsymbol{d}$, where $\Delta\boldsymbol{d}$ is a draw from N(0,$\Delta^2\Lambda$), and Δ and Λ are chosen adaptively to reduce autocorrelation among MCMC draws following Atchade (2006). The probability of accepting $\begin{bmatrix} d_{ij}^{u} \\ d_{ji}^{u} \end{bmatrix}$ is:

$$\mathrm{Pr(acceptance)} = \min\{\frac{\left[\exp\left(-\frac{1}{2}\left(d_{ij}^{u} + d_{ji}^{u}\right)\sigma_{d}^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}^{u}, d_{ji}^{u})}{\left[\exp\left(-\frac{1}{2}\left(d_{ij} + d_{ji}\right)\sigma_{d}^{-2}\right)\right] L(\boldsymbol{Y}|d_{ij}, d_{ji})}, 1\}$$

Step 11: Generating σ_{d}^{u}

$$(\sigma_{d}^{u} |\, d_{ij}^{u}, d_{ji}^{u}) \sim \mathrm{IW}_1(1 + N(N-1), 1 + \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} (d_{ij}^{u} + d_{ji}^{u})^2)$$

Step 12: If convergence is not yet reached, go to Step 1.