Published online ahead of print March 18, 2013

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1685

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

Yingda Lu
Lally School of Management and Technology, Rensselaer Polytechnic Institute, Troy, New York 12180,
yingda.lu1@gmail.com

Kinshuk Jerath, Param Vir Singh
David A. Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213
{kinshuk@cmu.edu, psidhu@cmu.edu}

We study the drivers of the emergence of opinion leaders in a networked community where users establish links to others, indicating their "trust" for the link receiver's opinion. This leads to the formation of a network, with high in-degree individuals being the opinion leaders. We use a dyad-level proportional hazard model with time-varying covariates to model the growth of this network. To estimate our model, we use *Weighted Exogenous Sampling with Bayesian Inference*, a methodology that we develop for fast estimation of dyadic models on large network data sets. We find that, in the Epinions 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, we find that time is an important moderator of these effects—intrinsic node characteristics are a stronger short-term driver of additional inlinks, whereas the preferential attachment effect has a smaller impact but it persists for a longer time. Our novel insights have important managerial implications for the design of online review communities.

Key words: user-generated content; opinion leaders; social networks; network growth; proportional hazard model; weighted exogenous sampling
History: Received March 1, 2010; accepted August 29, 2012, by Sandra Slaughter, information systems. Published online in *Articles in Advance*.

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, primarily 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 (http://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 that 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 (http://www.fool.com) and Seeking Alpha (http://www.seekingalpha.com) for sharing opinions on topics related to financial markets, YouTube (http://www.youtube.com) for sharing videos, IMDb (http://www.imdb.com) and Rotten

Tomatoes (http://www.rottentomotoes.com) for sharing opinions on movies, yelp (http://www.yelp.com) for sharing information on local food and entertainment, and, last but not the least, social networks such as Facebook (http://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 data set 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 Leinhardt 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 with 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 understudied (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.

[1] This method of employing the number of incoming links as a proxy for measuring opinion leadership is called the sociometric method and has been used widely before in sociology and marketing (Burt 1999, Iyengar et al. 2011, 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.

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 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 whereas 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, whereas 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, our results show 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, whereas 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 data sets. 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 (MCMC) 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 accurate 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 the online technical appendix to this paper (available at http://ssrn.com/abstract=2206016), 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 this paper is organized as follows. In §2, we develop the theoretical foundations motivating our empirical work. In §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 §4, we describe the Epinions data set we constructed and our variable definitions. In §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 §6, we provide several extensions and robustness checks for our basic model. In §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 capital (Allison et al. 1982, Merton 1968). In a community of reviewers, high-status reviewers (ones with a high in-degree) would be considered more attractive for those seeking opinions (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 toward 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.

Although 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" (i.e., nonnetwork 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. Although 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 et al. (2012) 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 Ipeirotis 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 although 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] Because 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), that is, 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 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:

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

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 $\mathbf{z}_{ijt}$ be the set of sender-, receiver- and dyad-specific covariates for the dyad ij at time t.

[3] Some other papers that develop stochastic models for network phenomena include those by 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).

Then, the above can be written as $V_{ijt} = \mathbf{z}_{ijt}\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is the vector of coefficients for $\mathbf{z}_{ijt}$. We discuss in detail the different covariates included in $\mathbf{z}_{ijt}$ in §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.)

Whereas 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 $V_{ijt} = \mathbf{z}_{ijt}\boldsymbol{\beta} + a_i + b_j$, 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

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

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 $\mathbf{z}_{ijt}$, and the unobserved dyad-specific homophily is captured by using a dyad-specific unobserved random effect, d_{ij}, as $V_{ijt} = \mathbf{z}_{ijt}\boldsymbol{\beta} + a_i + b_j + d_{ij}$, where $d_{ij} \sim MVN(0, \sigma_d^2)$.[4] Furthermore, we assume that the dyad-specific unobserved effects are symmetric, that is, $d_{ij} = d_{ji}$.

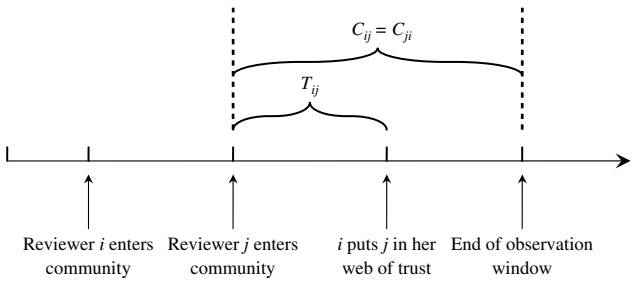
We can present V_{ijt} above in a simplified manner by aggregating the random effects together with the corresponding covariates as

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

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

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

Figure 1 Illustration of Link Formation Time and Censoring Time Used in the Model



Note. In this figure, $\mathbb{I}_{ij} = 1$ and $k_{ij} = \text{floor}(T_{ij})$, and $\mathbb{I}_{ji} = 0$ and $k_{ji} = C_{ji}$.

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 periods for which dyad ij has been observed, and let 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 $T_{ij} \le C_{ij}$ (i.e., if a tie formed within the observation time) and 0 otherwise, and $k_{ij} = \text{floor}(\min\{T_{ij}, C_{ij}\})$. We present this graphically in Figure 1.

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

$$\begin{aligned}\log L = \sum_{i,j\neq i} \Bigg\{ \mathbb{I}_{ij} \cdot \log[1 - \exp\{-\exp[\alpha(k_{ij}) + \mathbf{z}_{ij,k_{ij}}\boldsymbol{\beta} + a_i + b_j + d_{ij}]\}] \\ - \sum_{t=0}^{k_{ij}-1} \exp[\alpha(t) + \mathbf{z}_{ijt}\boldsymbol{\beta} + a_i + b_j + d_{ij}] \Bigg\}, \quad (1)\end{aligned}$$

where $\alpha(t) = \log\{\int_t^{t+1} \lambda_0(u)\,du\}$. (See Appendix A 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 Online Technical Appendix B 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, although our model above shares some commonalities with those of 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

4.1. 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 and 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 written at least one review on any movie to ensure that the individuals in our data set indeed have an expressed interest in movies. We relax these constraints on data collection later in §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 wrote 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 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 data set 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 data set 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).

[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.

4.2. 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.

4.2.1. 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, and 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:

$$\begin{aligned} \textit{Readability} = \text{GFI} = 0.4 * (&\text{average sentence length} \\ &+ \text{number of hard words for each 100 words}), \end{aligned}$$

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 pretrained for movies on a large movie data set, developed by 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 1,000" 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 1,000, 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).

4.2.2. Sender-Specific Covariates. This category consists of variables that provide information on the sender of a potential tie, and includes 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.

4.2.3. Dyad-Specific Covariates. This category consists of variables that provide information regarding the dyad in question and includes 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.

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 number of commonly trusted reviewers between the sender and the receiver. Note that whereas 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.

5. Estimation and Results

5.1. Estimation Methodology: WESBI

We have 5,180 individuals in our calibration data set, which generates 26,827,220 dyads. Because 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 data set 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 *Weighted Exogenous Sampling with Bayesian Inference*.

To employ this method, we collect all of the dyads that actually form ties within the observation time window and randomly sample from the dyads that do not form a tie within the observation time window. By aggregating these two sets of dyads, we construct a much smaller data set (we call this smaller data set as the *sampled data set*). 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 data set:

$$\begin{aligned}\log L &= w_1\Bigg(\sum_{(\mathbb{I}_{ij}=1)}\Bigg\{\log\big[1-\exp\{-\exp[\boldsymbol{\alpha}(k_{ij}) \\ &\qquad +\mathbf{z}_{ij,k_{ij}}\boldsymbol{\beta}+a_i+b_j+d_{ij}]\}\big] \\ &\qquad -\sum_{t=0}^{k_{ij}-1}\exp\big[\boldsymbol{\alpha}(t)+\mathbf{z}_{ijt}\boldsymbol{\beta}+a_i+b_j+d_{ij}\big]\Bigg\}\Bigg) \\ &\quad + w_0\Bigg(\sum_{(\mathbb{I}_{ij}=0)}\Bigg\{-\sum_{t=0}^{k_{ij}-1}\exp[\boldsymbol{\alpha}(t)+\mathbf{z}_{ijt}\boldsymbol{\beta}+a_i+b_j+d_{ij}]\Bigg\}\Bigg), \quad (2)\end{aligned}$$

where w_1 and w_0 are the weights of the log-conditional-likelihood functions for the ties that were

Table 1 Variable Definitions and Descriptive Statistics

Variables	Definition	Descriptive statistics[a]
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)

[a]Numbers outside parentheses are the means for the "Movies" data set, and those in parentheses are the corresponding standard deviations.

formed and the ties that were not formed, respectively. Here, $w_0=(1-Q_1)/(1-H_1)$ and $w_1=Q_1/H_1$, 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 data set.

We estimate the parameters of our model by using an MCMC hierarchical Bayes estimation procedure, using a Gibbs sampler and the Metropolis–Hastings algorithm. The full estimation procedure is provided in Appendix B. In Online Technical Appendix A, 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 data set, we sampled 10% of the dyads that did not form a link during our observation window. This final sampled data set has 1,906 established ties and 2,682,531 pairs that did not form a tie. Whereas the estimation for the full data set 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 data set. 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 and 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 that differ widely from those presented in our simulation results.

5.2. Estimation Results

We estimated our model in Matlab using the procedure in Appendix B. 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 overdispersed 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 for the data for the "Movies" community. For the estimation, we standardized 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 and 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				
$\text{Log}(a_0)$	−13.7542***	−17.4816***	−13.4542***	−12.5319***
$\text{Log}(a_1)$	−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***

*, **, *** The 90%, 95%, and 99% credible intervals, respectively, do not include zero.

5.2.1. 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, whereas 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 trust 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, whereas *past* review writing activity does not have a significant effect. This is likely because trust links are not explicitly dated and therefore get valued as endorsements even if a long time has passed, whereas 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, whereas reviews that are either too

long or too short are less preferred. The coefficient of the linear term of *Readability* is insignificant, whereas 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, whereas 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.

5.2.2. 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.

5.2.3. 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.

5.2.4. 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).

5.2.5. Unobserved Random Effects. The fact that σ_a, σ_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.

5.3. Model Performance

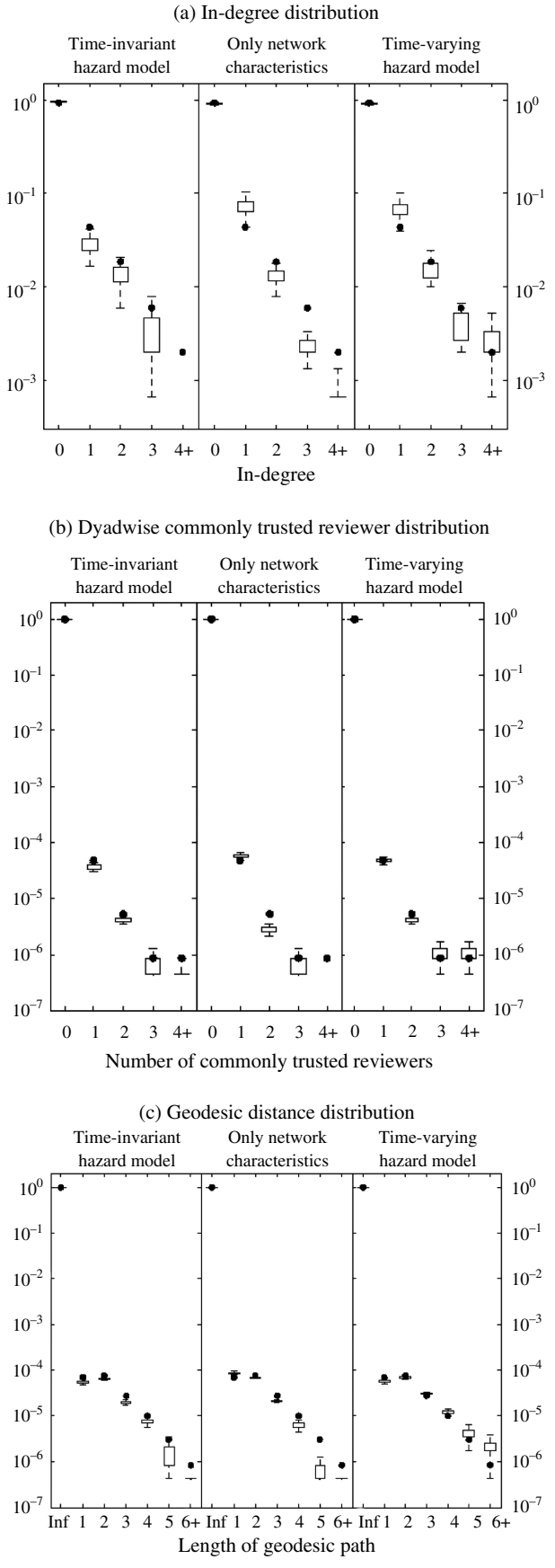
To test the performance of our model, we use two alternative models as benchmarks: (1) a time-invariant hazard model with all covariates (as in Narayan and Yang 2007) and (2) a time-varying hazard model with only network characteristics (and no node-level characteristics) as covariates (i.e., *Receiver's 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, dyadwise 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. Because we deal with a directed network, we use in-degree distribution, dyadwise 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, whereas 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 data set, and the boxes and whiskers represent the corresponding statistics across the simulated data sets. The whisker represents the upper and lower limits of the 200 corresponding simulated network statistics. The box represents the 25th and 75th percentiles. 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. 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

[6] Even a naïve model that 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.

Figure 2 Performance Tests



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 reviewers 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 that of our model as well as that of 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 performs better than the model with only network characteristics; however, its performance is 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.

6. Extensions and Robustness Checks

In this section, we extend our basic analysis in three different ways. First, we stratify our data set based on opinion leadership status and find that the strategies of forming links employed by individuals with high opinion leadership statuses (HOLSs) are very different from those employed by individuals with low opinion leadership statuses (LOLSs). 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]

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

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

	All links that are formed in data set are included		Only links that are formed first are included	
Variables	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*
$(Comprehensiveness)^2$	−0.5152***	−0.1022***	−0.5241***	−0.1216***
$(Objectivity)^2$	0.0436	−0.0896	0.0487	−0.0802
$(Readability)^2$	−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				
$\text{Log}(\alpha_0)$	−14.7342***	−11.4360***	−15.5124***	−12.4193***
$\text{Log}(\alpha\alpha_1)$	−4.7773	−5.3882	−4.2149	−5.3251
σ_a^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***

*, **, *** The 90%, 95%, and 99% credible intervals, respectively, do not include zero.

6.1. Analysis with Stratification Based on Opinion Leadership

We use the data set described in §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 behaviors of these two groups of senders differ from each other, we estimate our model for the two samples separately. We report the results in the first two columns of Table 3.

We uncover an interesting insight into the contrasting strategies for extending trust links employed by individuals with low and high opinion leadership statuses. Whereas 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, whereas 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, because 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 $A \to B$ and $B \to A$ both existing, and, say, $A \to B$ is formed before $B \to A$ is formed, then we remove the edge $B \to A$ 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.

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

In §5, we considered only the "Movies" community. In this section, we test our findings on a much larger, category-independent data set in which we also include individuals who only passively follow other reviewers without themselves writing any reviews. To collect this data set, in the first step, we use all individuals in the "Movies" community (as described in §4.1) 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 and 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 §4, and that 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—whereas the impact of previous opinion leadership carries over into future periods, previous reviews written have no significant impact on the rate of forming ties.

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

6.3. Analysis for Other Product Categories

To check the robustness of our estimation results, we replicated our analysis on the "Cars" and the "Home and Garden" categories. We construct the data sets 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 and Garden" community comprises 1,120 individuals with 457 ties formed within the community. We present the results for the "Cars" and the "Home and Garden" communities in the third and fourth columns of Table 2, respectively.

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, whereas both past and recent trust links have an impact—also hold for the "Cars" and "Home and Garden" categories. Note, however, that in both the "Cars" and "Home and 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 were 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 and 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 whereas 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

[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 and Garden" is extremely difficult to construct, so we work with this limited data set for this extension.

specific details about the opinion shared also available (such as the time of opinion sharing 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, whereas 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. (Although 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.) Because of 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, whereas 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 opinions, 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 objectivity (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 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 that could allow users to search reviews for a product based on these desirable characteristics.

Our study contributes not only to 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 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 on 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 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, because 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 review writing styles (which are, in fact, minimal in our data) and in the frequency of review writing 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 §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, because 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.

Acknowledgments
All authors contributed equally and have been listed in a random order.

Appendix A. Derivation of the Log-Conditional-Likelihood Function
For the basic proportional hazard model,

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

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

$$P(T_{ij}\geq t+1\mid T_t\geq t)=\exp\left(-\int_t^{t+1}\lambda_{ij}(u)\,du\right)$$
$$=\exp\left(-\exp(\mathbf{z}_{ijt}\boldsymbol{\beta})\int_t^{t+1}\lambda_0(u)\,du\right).$$

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

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

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 let 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 data set with N individuals in this basic model is

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

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 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 Online Technical Appendix B. For the procedures below, letters with superscript u represent the values of the updated corresponding parameters.

Step 1. Estimate $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma}^u\mid\boldsymbol{\beta},a_i,b_i,\alpha_0,\alpha_1,d_{ij},\text{data}$$
$$f(\boldsymbol{\gamma}^u\mid\boldsymbol{\beta},a_i,b_i,\alpha_0,\alpha_1,d_{ij},\text{data})$$
$$\propto|\Sigma_{\boldsymbol{\beta}\mathbf{0}}|^{-1/2}\exp\big[-\tfrac{1}{2}(\boldsymbol{\gamma}^u-\bar{\boldsymbol{\gamma}}_0)'\Sigma_{\boldsymbol{\gamma}\mathbf{0}}^{-1}(\boldsymbol{\gamma}^u-\bar{\boldsymbol{\gamma}}_0)\big]L(\boldsymbol{\Upsilon}),$$

where $\bar{\boldsymbol{\gamma}}_0$ and $\Sigma_{\boldsymbol{\gamma}\mathbf{0}}$ 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

$$\Pr(\text{acceptance})$$
$$=\min\left\{\frac{\exp[-(1/2)(\boldsymbol{\gamma}^u-\bar{\gamma}_0)'\Sigma_{\boldsymbol{\gamma}\mathbf{0}}^{-1}(\boldsymbol{\gamma}^u-\bar{\gamma}_0)]L(\boldsymbol{\Upsilon}\mid\boldsymbol{\gamma}^u)}{\exp[-(1/2)(\boldsymbol{\gamma}-\bar{\gamma}_0)'\Sigma_{\boldsymbol{\gamma}\mathbf{0}}^{-1}(\boldsymbol{\gamma}-\bar{\gamma}_0)]L(\boldsymbol{\Upsilon}\mid\boldsymbol{\beta})},1\right\}.$$

We define diffuse priors by setting $\bar{\boldsymbol{\gamma}}_0$ to be a vector of zeros and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0} = 30I$.

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

$$\begin{aligned} &f(a_i^u, b_i^u \mid \boldsymbol{\beta}^u, \alpha_0^u, \alpha_1^u, d_{ij}, \text{data}) \\ &\quad \propto N((a_i^u, b_i^u \mid \boldsymbol{\beta}^u, \alpha_0^u, \alpha_1^u, d_{ij}), \Sigma_{ab}) L(\mathbf{Y}) \\ &\quad \propto |\Sigma_{ab}|^{-1/2} \exp\left[-\tfrac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)'\right] L(\mathbf{Y}). \end{aligned}$$

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 $N(0, \Delta^2 \Lambda)$, 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

$$\begin{aligned} &\Pr(\text{acceptance}) \\ &= \min\left\{ \frac{[\exp(-\tfrac{1}{2}(a_i^u, b_i^u)\Sigma_{ab}^{-1}(a_i^u, b_i^u)')] L(\mathbf{Y} \mid a_i^u, b_i^u)}{[\exp(-\tfrac{1}{2}(a_i, b_i)\Sigma_{ab}^{-1}(a_i, b_i)')] L(\mathbf{Y} \mid a_i, b_i)}, 1 \right\}. \end{aligned}$$

Step 3. Generate Σ_{ab}^u: $\Sigma_{ab}^u \mid a_i^u, b_i^u$

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

where IW_2 denotes the inverse-Wishart distribution.

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

$$\begin{aligned} &f(d_{ij}^u, d_{ji}^u \mid \alpha_0^u, \boldsymbol{\beta}^u, a_i, b_i, \alpha_1^u, \sigma_d^2, \text{data}) \\ &\quad \propto N((d_{ij}^u, d_{ji}^u \mid \alpha_0^u, \boldsymbol{\beta}^u, a_i, b_i, \alpha_1^u), \sigma_d^2) L(\mathbf{Y}) \\ &\quad \propto \sigma_d^{-1} \exp\left[-\tfrac{1}{2}(d_{ij}^u + d_{ji}^u)^2 \sigma_d^{-2}\right] L(\mathbf{Y}). \end{aligned}$$

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 draws of the unobservable similarity effects from the previous iteration, and we draw d_{ij}^u and d_{ji}^u by $\begin{bmatrix} d_{ij}^u \\ d_{ji}^u \end{bmatrix} = \begin{bmatrix} d_{ij} \\ d_{ji} \end{bmatrix} + \Delta \mathbf{d}$, where $\Delta \mathbf{d}$ 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

$$\begin{aligned} &\Pr(\text{acceptance}) \\ &= \min\left\{ \frac{[\exp(-\tfrac{1}{2}(d_{ij}^u + d_{ji}^u)\sigma_d^{-2})] L(\mathbf{Y} \mid d_{ij}^u, d_{ji}^u)}{[\exp(-\tfrac{1}{2}(d_{ij} + d_{ji})\sigma_d^{-2})] L(\mathbf{Y} \mid d_{ij}, d_{ji})}, 1 \right\}. \end{aligned}$$

Step 5. Generate σ_d^u:

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

where IW_1 denotes the inverse-Wishart distribution.

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

References

Alias-I (2008) LingPipe home page. Accessed January 23, 2013, http://alias-i.com/lingpipe/index.html.

Allison P, Long S, Krauze T (1982) Cumulative advantage and inequality in science. *Amer. Sociol. Rev.* 47(5):615–625.

Ansari A, Koenigsberg O, Stahl F (2011) Modeling multiple relationships in social networks. *J. Marketing Res.* 48(4):713–728.

Atchade Y (2006) An adaptive version for the metropolis adjusted Langevin algorithm with a truncated drift. *Methodology Comput. Appl. Probab.* 8(2):235–254.

Barabasi A, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512.

Bonacich P (1987) Power and centrality: A family of measures. *Amer. J. Sociol.* 92(5):1170–1182.

Braun M, Bonfrer A (2011) Scalable inference of customer similarities from interactions data using Dirichlet processes. *Marketing Sci.* 30(3):513–531.

Burt R (1999) The social capital of opinion leaders. *Ann. Amer. Acad. Political Soc. Sci.* 566(1):37–54.

Chan K, Misra S (1990) Characteristics of the opinion leader: A new dimension. *J. Advertising* 19(3):53–60.

Dorogovtsev S, Mendes J (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, New York).

DuBay W (2004) The principles of readability. Accessed January 23, 2013, http://www.impact-information.com/.

Fader P, Hardie B, Lee K (2005) RFM and CLV: Using iso-value curves for customer base analysis. *J. Marketing Res.* 42(4): 415–430.

Fehr E, Gachter S (2000) Fairness and retaliation the economics of reciprocity. *J. Econom. Perspect.* 14(3):159–181.

Forman C, Ghose A, Wiesenfel B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. System Res.* 19(3):291–313.

Gelman A, Carlin J, Stern H, Rubin D (2003) *Bayesian Data Analysis* (Chapman and Hall/CRC, Boca Raton, FL).

Ghose A, Ipeirotis P (2011) Estimating helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans. Knowledge Data Engrg.* 23(10):1498–1512.

Ghose A, Ipeirotis P, Sundarajan A (2009) The dimensions of reputation in electronic markets. Working paper, New York University, New York.

Gladwell M (2000) *The Tipping Point* (Little, Brown and Company, New York).

Gould S (2002) *The Structure of Evolutionary Theory* (Harvard University Press, Boston).

Greene W (2003) *Econometric Analysis*, 5th ed. (Prentice Hall, Upper Saddle River, NJ).

Handcock M, Raftery A, Tantrum J (2007) Model-based clustering for social networks. *J. Royal Statist. Soc. A* 170(2):301–354.

Hill S, Provost F, Volinsky C (2006) Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.* 22(2):256–276.

Hoff P (2005) Bilinear mixed-effects models for dyadic data. *J. Amer. Statist. Assoc.* 100(469):286–295.

Hoff P, Raftery A, Handcock M (2002) Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* 97(460): 1090–1098.

Holland P, Leinhardt S (1972) Reply: Some evidence on the transitivity of positive interpersonal sentiment. *Amer. J. Sociol.* 77(6):1205–1209.

Hunter D, Goodreau S, Handcock M (2008) Goodness of fit of social network models. *J. Amer. Statist. Assoc.* 103(481):248–258.

Iacobucci D, Hopkins N (1992) Modeling dyadic interactions and networks in marketing. *J. Marketing Res.* 29(1):5–17.

Iyengar R, Van den Bulte C, Valente T (2011) Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* 30(2):195–212.

Jones J, Handcock M (2003) An assessment of preferential attachment as a mechanism for human sexual network formation. *Proc. Royal Soc.* 270(1520):1123–1128.

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

Katz L, Powell J (1955) Measurement of the tendency towards reciprocation of choice. *Sociometry* 18(4):403–409.

Kim S, Hovy E (2006) Automatic identification of pro and con reasons in online reviews. *Annual Meeting of the ACL, Proc. COLING/ACL on Main Conf. Poster Sessions, Sydney, Australia.*

King C, Summers J (1970) Overlap of opinion leadership across consumer product categories. *J. Marketing Res.* 7(1):43–50.

Kossinets G, Watts D (2006) Empirical analysis of an evolving social network. *Science* 311(5757):88–90.

Liu J, Cao Y, Lin C, Huang Y, Zhou M (2007) Low-quality product review detection in opinion summarization. *Joint Conf. Empirical Methods in NLP and Comput. NLP* (Association for Computational Linguistics, Stroudsburg, PA), 334–342.

Manski C, Lerman S (1977) The estimation of choice probabilities form choice based samples. *Econometrica* 45(8):1977–1988.

Mayzlin D, Yoganarasimhan H (2012) Link to success: How blogs build an audience by promoting rivals. *Management Sci.* 58(9):1651–1668.

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: Homophily in social networks. *Annual Rev. Sociol.* 27(August): 415–444.

Merton R (1968) The matthew effect in science. *Science* 159(3810): 56–63.

Mislove A, Marcon M, Gummadi K, Druschel P, Bhattacharjee B (2007) Measurement and analysis of online social networks. *Proc. 7th ACM SIGCOMM Conf. on Internet Measurement* (ACM, New York), 29–42.

Myers J, Robertson T (1972) Dimensions of opinion leadership. *J. Marketing Res.* 9(1):41–46.

Narayan V, Yang S (2007) Modeling the formation of dyadic relationships between consumers in online communities. Working paper, Cornell University, Ithaca, NY.

Otterbacher J (2009) "Helpfulness" in online communities: A measure of message quality. *Proc. SIGCHI Conf. Human Factor Comput. Systems* (ACM, New York), 955–964.

Pang B, Lee L (2004) A sentimental education. *Annual Meeting ACL. Proc. 42nd Annual Meeting Assoc. Comput. Linguistics, Barcelona, Spain.*

Robins G, Snijders T, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p^*) models for social networks. *Soc. Networks* 29(2007):192–215.

Rogers E (2003) *Diffusion of Innovation* (Free Press, New York).

Snijders T, Pattison P, Robins G, Handcock M (2006) New specifications for exponential random graph models. *Sociol. Methodology* 36(1):99–153.

Stephen A, Toubia O (2009) Explaining the power-law degree in a social commerce network. *Soc. Networks* 31(4):262–270.

Stephen A, Dover Y, Muchnik L, Goldenberg J (2012) The effects of transmitter activity and connectivity on information dissemination over online social networks. Working paper, University of Pittsburgh, Pittsburgh.

Tucker C, Zhang J (2010) Growing two-sided networks by advertising the user base: A field experiment. *Marketing Sci.* 29(5): 805–814.

Valente T, Hoffman B, Ritt-Olson A, Lichtman K, Johnson A (2003) Effects of a social-network method for group assignment strategies on peer-led tobacco prevention programs in schools. *Amer. J. Public Health* 93(11):1837–1843.

Van Alstyne M, Brynjolfsson E (2005) Global village or cyberbalkans? Modeling and measuring the integration of electronic communities. *Management Sci.* 51(6):851–868.

Van den Bulte C, Joshi Y (2007) New product diffusion with influentials and imitators. *Marketing Sci.* 26(3):400–421.

Vernette E (2004) Targeting women's clothing fashion opinion leaders in media planning: An application for magazine. *J. Advertising Res.* 44(1):90–107.

Watts D, Dodds P (2007) Influentials, networks, and public opinion formation. *J. Consumer Res.* 34(4):441–458.

Zhang Z, Varadarajan B (2006) Utility scoring of product reviews. *Proc. 15th ACM Internat. Conf. Inform. Knowledge Management, Arlington, VA.*