Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Utility-Based Link Recommendation for Online Social Networks

Zhepeng Li, Xiao Fang, Xue Bai, Olivia R. Liu Sheng

To cite this article:
Zhepeng Li, Xiao Fang, Xue Bai, Olivia R. Liu Sheng (2017) Utility-Based Link Recommendation for Online Social Networks. Management Science 63(6):1938-1952. https://doi.org/10.1287/mnsc.2016.2446

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 63, No. 6, June 2017, pp. 1938–1952
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Utility-Based Link Recommendation for Online Social Networks

Zhepeng Li,[a] Xiao Fang,[b,*] Xue Bai,[c] Olivia R. Liu Sheng[d]

[a] Schulich School of Business, York University, Toronto, Ontario M3J 1P3, Canada; [b] Lerner College of Business and Economics, University of Delaware, Newark, Delaware 19716; [c] School of Business, University of Connecticut, Storrs, Connecticut 06268; [d] David Eccles School of Business, University of Utah, Salt Lake City, Utah 84112

[*] Corresponding author

Contact: zli@schulich.yorku.ca (ZL); xfang@udel.edu (XF); xue.bai@uconn.edu (XB); olivia.sheng@eccles.utah.edu (ORLS)

Received: February 15, 2015
Revised: September 15, 2015; November 26, 2015
Accepted: December 17, 2015
Published Online in Articles in Advance: May 24, 2016

https://doi.org/10.1287/mnsc.2016.2446

Abstract. Link recommendation, which suggests links to connect currently unlinked users, is a key functionality offered by major online social networks. Salient examples of link recommendation include "People You May Know" on Facebook and LinkedIn as well as "You May Know" on Google+. The main stakeholders of an online social network include users (e.g., Facebook users) who use the network to socialize with other users and an operator (e.g., Facebook Inc.) that establishes and operates the network for its own benefit (e.g., revenue). Existing link recommendation methods recommend links that are likely to be established by users but overlook the benefit a recommended link could bring to an operator. To address this gap, we define the utility of recommending a link and formulate a new research problem—the utility-based link recommendation problem. We then propose a novel utility-based link recommendation method that recommends links based on the value, cost, and linkage likelihood of a link, in contrast to existing link recommendation methods that focus solely on linkage likelihood. Specifically, our method models the dependency relationship between the value, cost, linkage likelihood, and utility-based link recommendation decision using a Bayesian network; predicts the probability of recommending a link with the Bayesian network; and recommends links with the highest probabilities. Using data obtained from a major U.S. online social network, we demonstrate significant performance improvement achieved by our method compared with prevalent link recommendation methods from representative prior research.

History: Accepted by Anandhi Bharadwaj, information systems.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2016.2446.

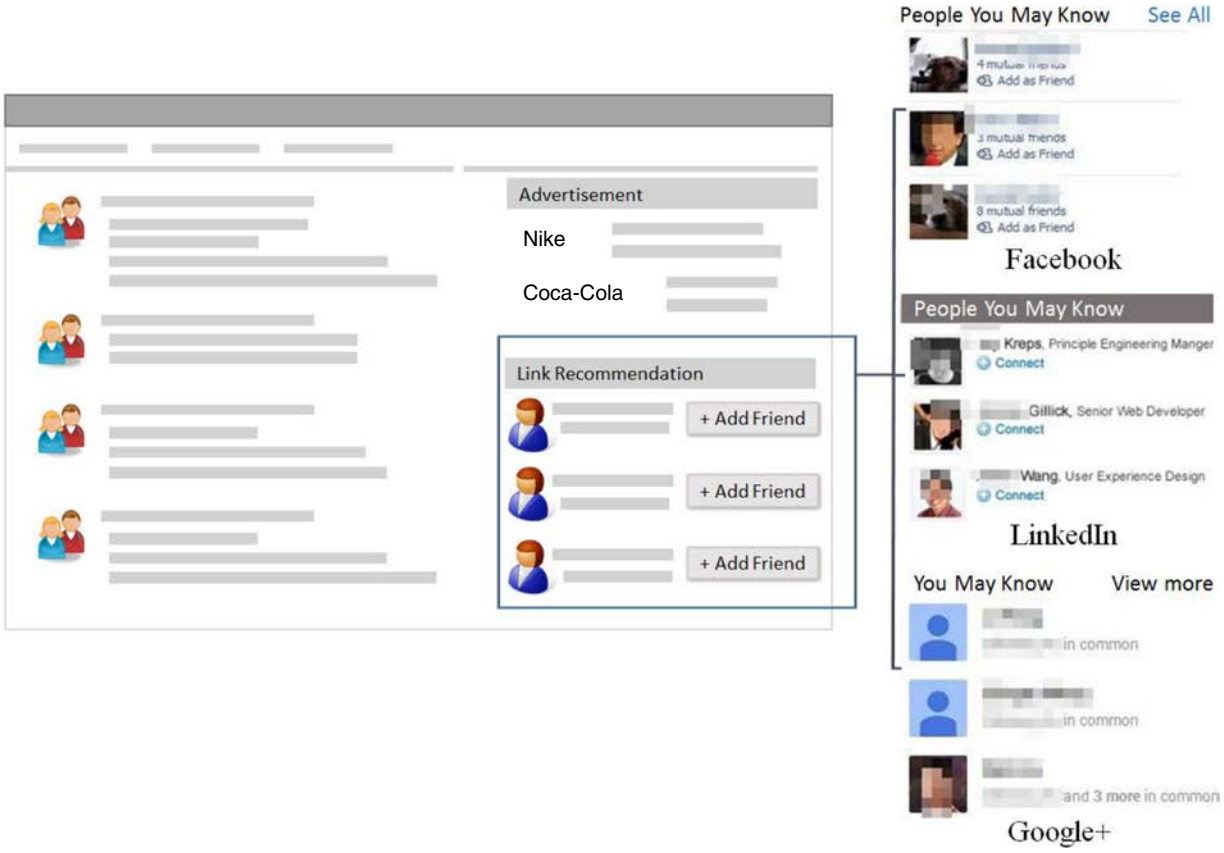
Keywords: utility-based link recommendation • link prediction • Bayesian network learning • continuous latent factor • online social network • machine learning • network formation

1. Introduction

Online social networks such as Facebook, LinkedIn, and Google+ have gained unprecedented numbers of users in a short time, attracting massive attention from both industry and academia to study and utilize these networks for economic and societal benefits (Jackson 2008, Backstrom and Leskovec 2011, Fang et al. 2013b). It is common for online social networks to implement a link recommendation mechanism, which suggests links to connect currently unlinked users. As shown in Figure 1, salient examples of link recommendation include "People You May Know" on Facebook and LinkedIn as well as "You May Know" on Google+. Since its early success on LinkedIn, link recommendation has become a standard feature of online social networks (Davenport and Patil 2012).

The main stakeholders of an online social network include users (e.g., Facebook users) who use the network to connect and communicate with other users (i.e., friends) and an operator (e.g., Facebook Inc.) that establishes and operates the network for its own benefit (e.g., revenue) (Ellison et al. 2007, Huberman et al. 2009, Kaplan and Haenlein 2010). Therefore, the advantages of link recommendation are twofold. First, link recommendation could cater to users' needs of socializing and networking with others in an online social network. By helping users connect with new friends, link recommendation allows new users to quickly become engaged in a community and facilitates existing users to enlarge their circles of friends. Second, link recommendation could benefit the operator of an online social network as well. According to *eMarketer* (2012), operators of online social networks reaped an estimated $12 billion from advertisements on these networks in 2014, up from $10 billion in 2013. Understandably, link recommendation potentially leads to a more connected network of users, which drives advertisements to reach more users and ultimately brings more revenue to the network's operator. Existing research develops link recommendation methods from the perspective of link prediction (Al Hasan et al. 2006, Liben-Nowell and Kleinberg 2007, Lichtenwalter et al. 2010, Gong et al. 2012). In general, these methods predict the likelihood that a

Figure 1. (Color online) Link Recommendation in Online Social Networks

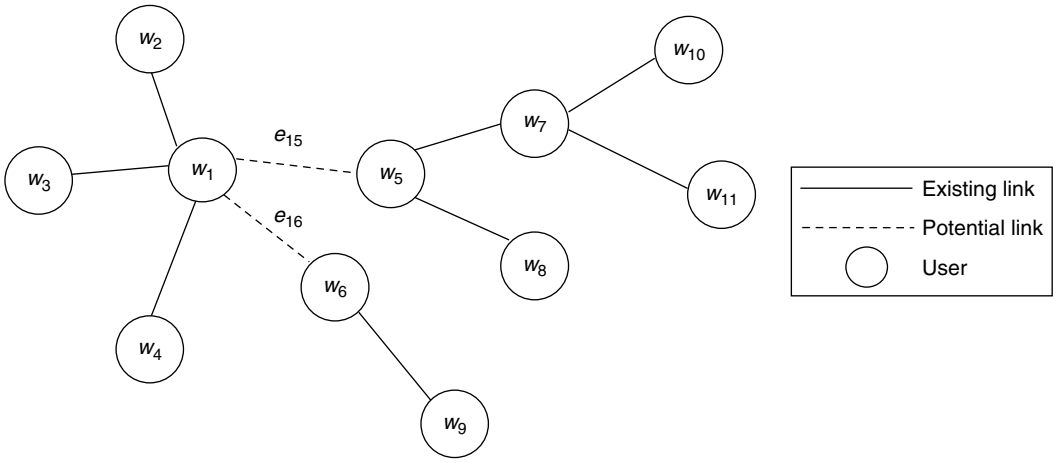


potential link[1] will be established by users, namely linkage likelihood, and recommend potential links with the highest linkage likelihoods (Al Hasan et al. 2006, Liben-Nowell and Kleinberg 2007, Lichtenwalter et al. 2010, Gong et al. 2012). While existing link recommendation methods cater well to users' social needs by recommending links that are likely to be established, they largely overlook the other advantage of link recommendation, i.e., benefiting the operator of an online social network; this is a fundamental gap that motivates our research.

We illustrate the gap using the example of Facebook, whose operator harvests the majority of its $7.9 billion revenue from advertisements on the network (Facebook Inc. 2013). Facebook allows an advertisement to be placed on the Facebook page of selected users. A user can interact with the advertisement through actions including click, comment, like, and share. Such interaction propagates the advertisement to the user's friends, who can also interact with the advertisement and further propagate it to their friends. As this propagation process continues, the advertisement can reach a much larger number of users than the initially selected users. Facebook Inc. obtains revenue each time the advertisement reaches a user. In this context, let us consider recommending one link out of the two potential links e_{15} and e_{16}, shown in Figure 2. Assuming that e_{15} and e_{16} have the same linkage likelihood, existing methods that recommend links purely based on linkage likelihood are indifferent about them and randomly pick one to recommend (Al Hasan et al. 2006, Liben-Nowell and Kleinberg 2007, Lichtenwalter et al. 2010, Gong et al. 2012). However, e_{15} could bring much more advertisement revenue to Facebook Inc. than e_{16} because of the following considerations. First, advertisements initially placed on the Facebook page of user w_1 could reach more users through e_{15} than through e_{16}. Second, advertisements propagated from users w_2, w_3, and w_4 to user w_1 could reach more users through e_{15} than through e_{16}. Third, more users could propagate advertisements to users w_1, w_2, w_3, and w_4 through e_{15} than through e_{16}. Therefore, for the benefit of Facebook Inc. it is much more desirable to recommend e_{15} than e_{16}, whereas existing link recommendation methods are indifferent about them.

To address this gap, this study defines a new link recommendation problem and proposes a novel link recommendation method. The key difference between our link recommendation problem and link recommendation problems defined in prior studies is the consideration of the operator's benefit from link recommendation in our problem formulation. Since different potential links occupy different structural positions in

Figure 2. An Illustrating Example



an online social network, they could each bring different values to the network's operator. Furthermore, the value of a potential link can only be realized if it is established by users; on the other hand, a cost to the network's operator is incurred if a recommended link is not established. Therefore, we define the utility of recommending a potential link by considering its value, cost, and whether it will be established, and we formulate the utility-based link recommendation problem. To solve the problem, we propose a novel utility-based link recommendation method. Unlike existing link recommendation methods that recommend potential links solely based on their linkage likelihoods, our method considers their values, costs, and linkage likelihoods when making link recommendation decisions. Specifically, we propose a Bayesian network learning method that models the dependency relationship between the value, cost, linkage likelihood, and link recommendation decision; predicts the probability of recommending a potential link with the learned Bayesian network; and recommends potential links with the highest probabilities. We note that linkage likelihood is latent (unobserved) and continuous. Hence, the principal methodological obstacle overcome by our proposed method is how to learn a Bayesian network with a continuous latent factor.

The rest of the paper is organized as follows. We begin in Section 2 by reviewing prior works related to our study. We then define the utility-based link recommendation problem in Section 3 and propose a utility-based link recommendation method in Section 4. The effectiveness of our proposed method is evaluated using data collected from a major online social network in Section 5. The paper concludes with implications and future research directions in Section 6.

2. Related Work

Link recommendation methods proposed in prior studies predict the linkage likelihood that a potential link will be established and recommend potential links with the highest linkage likelihoods. According to different prediction approaches used, prior link recommendation methods can be broadly categorized into learning-based link recommendation methods and proximity-based link recommendation methods. In the following, we review representative methods in each category.

Learning-based link recommendation methods learn a model from observed link establishments and predict linkage likelihood using the learned model. Given a social network, one can construct training data from observed link establishments in the network. In general, each record of the training data has the format $\langle f_1, f_2, \ldots, f_m, c\rangle$, where $f_1, f_2, \ldots, f_m$ represent features that affect link establishment and c is the class label. The class label c is 1 for an existing link and 0 for a potential link. Commonly used features include topological features that are derived from the structure of a social network and nodal features that are computed from the characteristics of individual users in a social network. Topological features such as the number of common neighbors and the shortest distance between users have frequently been used by learning-based link recommendation methods (O'Madadhain et al. 2005, Al Hasan et al. 2006, Wang et al. 2007, Benchettara et al. 2010, Lichtenwalter et al. 2010). Nodal features that are calculated from users' demographical and geographical characteristics have also been widely employed (O'Madadhain et al. 2005, Zheleva et al. 2010, Scellato et al. 2011, Wang et al. 2011).

Once training data are constructed, supervised machine learning methods can be applied to the data to predict linkage likelihood. O'Madadhain et al. (2005) employ logistic regression to predict the likelihood of interaction between users using data collected from CiteSeer and Enron emails. Wang et al. (2007) combine a local Markov random field model and logistic regression to predict the likelihood of coauthorship using

DBLP and PubMed article data sets. Benchettara et al. (2010) also target predicting the likelihood of coauthorship but employ a decision tree classifier enhanced with Adaboost for this prediction. Hopcroft et al. (2011) adopt a factor-based graphical model to predict reciprocal relationships on Twitter. Gong et al. (2012) predict linkage likelihood in Google+ using a support vector machine (SVM). Besides classification methods, other supervised learning methods such as supervised random walk (Backstrom and Leskovec 2011), matrix factorization-based methods (Kunegis et al. 2010, Yang et al. 2011), and relational learning (Popescul and Ungar 2003) have been employed to predict linkage likelihood in unipartite or bipartite social networks.

Proximity-based link recommendation methods surrogate the linkage likelihood of a potential link using the proximity between users that would be connected by the link. According to McPherson et al. (2001), similar users are more likely to interact and connect with each other. Therefore, higher proximity indicates a higher chance of linkage. Proximity metrics employed by link recommendation methods consist of nodal proximity metrics and structural proximity metrics. Nodal proximity metrics measure the similarity between users using their characteristics; this includes demographical characteristics such as age, education, and occupation (Zheleva et al. 2010); geographical characteristics such as colocation and distance (Quercia and Capra 2009, Crandall et al. 2010, Wang et al. 2011); and semantic characteristics such as keywords and annotation tags (Shen et al. 2006, Chen et al. 2009, Schifanella et al. 2010, Kuo et al. 2013). For the computation of nodal proximity between users, typical similarity functions such as the Manhattan distance, the cosine similarity, the Kullback–Leibler divergence, and the Jaccard coefficient have been applied to users' characteristics (Shen et al. 2006, Chen et al. 2009, Scellato et al. 2011, Wang et al. 2011, Adali et al. 2012).

Users' structural features have been widely employed to study their behaviors in social networks. For example, Doreian (1989) and Zhang et al. (2013) propose autocorrelation models to examine the impact of users' structural features such as cohesion and structural equivalence on their actions and choices in social networks. In link recommendation, structural proximity metrics measure the proximity between users using their structural features in a social network (Liben-Nowell and Kleinberg 2007). One type of structural proximity metrics targets users' neighborhoods. For example, the common neighbor between users is defined as the number of mutual neighbors in a social network (Newman 2001, Liben-Nowell and Kleinberg 2007). Extended from the common neighbor metric, the Adamic/Adar metric assigns less weight for more connected common neighbors (Adamic and Adar 2003, Liben-Nowell and Kleinberg 2007). Motivated by the finding that the likelihood of linking two users is correlated with their neighborhood sizes, the preferential attachment between users is defined as the product of their neighborhood sizes (Barabási and Albert 1999, Newman 2001, Barabási et al. 2002, Liben-Nowell and Kleinberg 2007). Observing that two users are similar if their neighbors are similar, Jeh and Widom (2002) define the SimRank score between users as the average of their neighbors' SimRank scores. Going beyond neighborhoods, another type of structural proximity metrics focuses on the paths connecting users. The Katz index measures the structural proximity between users using the number of paths connecting them, weighted by the lengths of these paths (Katz 1953). Originally developed for measuring the social status of a social entity, the Katz index has been shown to be effective in forecasting linkage likelihood (Liben-Nowell and Kleinberg 2007). Considering link establishment between users as a random walk from one to the other, Tong et al. (2006) adapts the PageRank algorithm (Brin and Page 1998) to compute the structural proximity between users as the summation of stationary probabilities that one user reaches the other. Using a similar idea of treating link establishment as a random walk, Fouss et al. (2007) define the hitting time between users as the expected number of steps to reach one user from the other for the first time and measure the structural proximity between them as the negation of their hitting time.

Our literature review suggests that existing link recommendation methods focus on predicting linkage likelihood but overlook the benefit of link recommendation to an operator. To overcome this limitation, we define a new link recommendation problem that takes into account operator's benefit from link recommendation. We then propose a novel link recommendation method to solve the problem. The essential novelty of our proposed method lies in its consideration of the value, cost, and linkage likelihood of a potential link when making a link recommendation decision, in contrast to existing methods that recommend links based solely on linkage likelihood.

3. Utility-Based Link Recommendation Problem

In this section, we define the utility of recommending a potential link and formulate the utility-based link recommendation problem. Let $W = \{w_j\}$, $j = 1, 2, \ldots, n$, be a set of users in an online social network. User w_j makes value contribution v_j to the operator of the online social network (Granovetter 2005, Jackson 2008). Value v_j consists of two parts: intrinsic value and network value, and

$$v_j = v_j^I + v_j^N, \tag{1}$$

where v_j^I and v_j^N represent the intrinsic value and network value of w_j, respectively (Jackson and Wolinsky 1996, Domingos and Richardson 2001, Watts 2001). Intrinsic value refers to a user's value that is independent of the link structure of an online social network. One example of a user's intrinsic value is his or her membership fee paid to LinkedIn (LinkedIn Corporation 2014).[2] Network value, on the other hand, denotes a user's value that is dependent on the link structure of an online social network. Let us consider advertisements, the major revenue source for operators of online social networks (Facebook Inc. 2013, LinkedIn Corporation 2014), as an example. In this example, a user's network value is the advertisement revenue contributed by the user, which depends on the number of other users that advertisements initiated by the user (e.g., advertisements initially placed on the user's Facebook page) can reach via the link structure of an online social network (Facebook Inc. 2013, LinkedIn Corporation 2014).

In general, a user's network value depends on the size of the user's direct and indirect neighborhood in an online social network (Ballester et al. 2006, Jackson 2008). Intuitively, a user's network value increases as the user has more direct and indirect neighbors (Ballester et al. 2006, Jackson 2008). Moreover, a user's impact on his or her neighbor decays as the distance between them increases (Granovetter 1973, Newman et al. 2002, Jackson 2008). Therefore, a user's network value can be defined as

$$v_j^N = m_j \cdot \sum_{x=1}^{X} \alpha^x |N_{j,x}|. \tag{2}$$

In Equation (2), $N_{j,x}$ represents the set of xth-degree neighbors of w_j, and $|\cdot|$ denotes the cardinality of a set. For example, $N_{j,1}$ refers to first-degree neighbors of w_j or direct neighbors of w_j. To model a user's diminishing impact on his or her farther neighbors, we introduce decay factor $\alpha \in (0,1)$ in Equation (2). Locality parameter X specifies the farthest neighbors considered when calculating a user's network value. The value of X is set such that $\alpha^X |N_{j,X}|$ becomes trivial (Jackson and Rogers 2005, Jackson 2008). Having defined the network impact of w_j as $\sum_{x=1}^{X} \alpha^x |N_{j,x}|$, we use m_j to model the value contribution by one unit of this impact. For example, m_j can be estimated as the revenue generated if an advertisement initiated by w_j reaches a (direct or indirect) neighbor of w_j, multiplied by the number of advertisements initiated by w_j.

Combining Equations (1) and (2), we can compute a user's value v_j as

$$v_j = v_j^I + m_j \cdot \sum_{x=1}^{X} \alpha^x |N_{j,x}|. \tag{3}$$

For an online social network with n users, the total value TV of these users can be obtained by summing the value of each user in the network (Jackson 2008). We therefore have

$$\mathrm{TV} = \sum_{j=1}^{n} v_j. \tag{4}$$

We are now ready to define the value of a potential link. Let E be the set of links currently existing in an online social network. We denote e_{jh} as a potential link that would connect currently unlinked users w_j and w_h. Value V_{jh} of potential link e_{jh} can be calculated as

$$V_{jh} = \mathrm{TV}_{E \cup \{e_{jh}\}} - \mathrm{TV}_E, \tag{5}$$

where $\mathrm{TV}_{E\cup\{e_{jh}\}}$ and TV_E denote the total user value in the online social network with and without link e_{jh} respectively, which can be obtained using Equations (4) and (3). By applying Equation (5) to calculate V_{jh}, users' intrinsic values before and after adding e_{jh} cancel each other out. This is reasonable because adding a link to an online social network affects only the structure of the network, and intrinsic value is independent of network structure.

By recommending potential link e_{jh}, value V_{jh} is realized if the recommended link is accepted by users w_j and w_h and hence established. On the other hand, if the recommended link is considered irrelevant by w_j or w_h and thus not established, cost C_{jh} is incurred. One example of C_{jh} is the opportunity cost of not being able to recommend another potential link because of recommending e_{jh}, considering that the number of links recommended to a user is limited. Therefore, utility U_{jh} of recommending potential link e_{jh} can be computed as

$$U_{jh} = I_{jh} \cdot V_{jh} - (1 - I_{jh}) \cdot C_{jh}, \tag{6}$$

where $I_{jh} = 1$ if e_{jh} is established and $I_{jh} = 0$ otherwise.

Having defined the utility of recommending a potential link, we can formulate the utility-based link recommendation problem as follows.

Problem. Given an online social network, its users W, its existing links E, and K, recommend top K potential links[3] with the highest utilities among all potential links, where the utility of recommending a potential link is defined in Equation (6).

4. Utility-Based Link Recommendation Method

To solve the utility-based link recommendation problem, we first identify key factors determining utility-based link recommendation decision and construct a Bayesian network to capture dependency relationships among the identified factors and utility-based link recommendation decision. We then propose how to learn the distribution of each identified factor in the Bayesian network and how to predict the probability of recommending a potential link with the learned Bayesian network. Finally, potential links with the highest recommendation probabilities are recommended.

4.1. A Bayesian Network for Utility-Based Link Recommendation

Utility-based link recommendation decision depends on three factors: value (V), cost (C), and latent linkage likelihood (L). The value factor (V) refers to the value of a potential link, which can be calculated using Equation (5). The cost factor (C) stands for the cost incurred if a potential link is recommended but not established. The latent linkage likelihood factor (L) represents the likelihood that a potential link will be established. Link recommendation brings value to an operator, if a recommended link is established, or incurs cost otherwise. Thus, linkage likelihood is an essential factor for utility-based link recommendation decision. Further, linkage likelihood is unobserved; thus it is latent in our proposed Bayesian network. The utility of a potential link depends on its value (V), cost (C), and linkage likelihood (L), and the objective of utility-based link recommendation is to recommend links with the highest utilities. Therefore, factors value (V), cost (C), and latent linkage likelihood (L) jointly determine utility-based link recommendation decision R, where $R = 1$ if a potential link is recommended and $R = 0$ otherwise.

It has been shown theoretically and empirically that nodal proximity and structural proximity are two effective predictors for linkage likelihood (L) (Heider 1958, McPherson et al. 2001, Liben-Nowell and Kleinberg 2007, Crandall et al. 2010). Nodal proximity denotes the similarity between users in terms of their individual characteristics such as age, gender, and education (Chen et al. 2009, Crandall et al. 2010, Schifanella et al. 2010). For two users w_j and w_h, nodal proximity $N(w_j, w_h)$ between them is calculated as

$$N(w_j, w_h) = \text{sim}(\mathbf{r}_j, \mathbf{r}_h), \tag{7}$$

where sim( ) is a similarity function, and $\mathbf{r}_j$ and $\mathbf{r}_h$ denote characteristics of w_j and w_h, respectively. Choice of the similarity function depends on data types of user characteristics (Tan et al. 2005); the specific similarity function used in this study is described in Section 5.2. The effectiveness of nodal proximity in predicting linkage likelihood can be explained by homophily theory, which states that "similarity breeds connection" (McPherson et al. 2001, p. 415). Therefore, the higher the nodal proximity between users, the more likely a link connecting them will be established.

Structural proximity measures the proximity between users using their structural features in a social network (Liben-Nowell and Kleinberg 2007). Prior studies have empirically shown the power of structural proximity metrics in predicting linkage likelihood (e.g., Liben-Nowell and Kleinberg 2007). Among structural proximity metrics, the Katz index consistently performs well in predicting linkage likelihood (Liben-Nowell and Kleinberg 2007). We thus adopt the Katz index to measure structural proximity between users. Accordingly, the structural proximity $S(w_j, w_h)$ between users w_j and w_h is given by

$$S(w_j, w_h) = \sum_k \beta^k |path_{jh}^{\langle k \rangle}|, \tag{8}$$

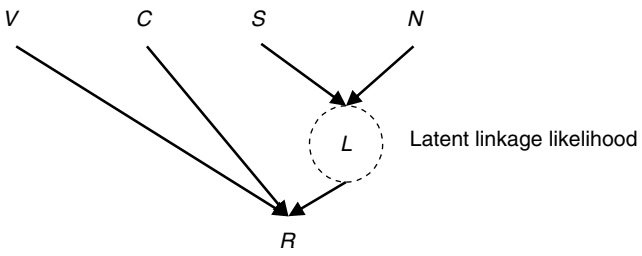
where $path_{jh}^{\langle k \rangle}$ represents the set of length k paths connecting users w_j and w_h, $|\cdot|$ is the cardinality of a set, and weight β is between 0 and 1 (Katz 1953, Liben-Nowell and Kleinberg 2007). The predictive power of the Katz index is rooted in cognitive balance theory (Heider 1958). According to this theory, sentiments (or attitudes) of indirectly connected users could become consistent gradually, which in turn could drive them to link to each other (Heider 1958). In this light, the more paths connecting two users and the shorter the distances of these paths, the more likely it is that a link connecting them will be established.

Having identified key factors underlying utility-based link recommendation decision, we model dependency relationships among them using a Bayesian network. Our choice of Bayesian network is driven by the following considerations. First, the Bayesian network is a natural choice for probability prediction, and thus we choose it to predict the probability of recommending a potential link. Second, a Bayesian network is a powerful but easy-to-understand model for capturing dependencies among variables (Heckerman 2008, Zheng and Pavlou 2010). As shown in Figure 3, the proposed Bayesian network consists of five factors: value (V), cost (C), structural proximity (S), nodal proximity (N), and latent linkage likelihood (L), as well as utility-based link recommendation decision (R). The network assumes mutual independences among factors value (V), cost (C), structural proximity (S), and nodal proximity (N); it also captures two dependency relationships: (i) value (V), cost (C), and latent linkage likelihood (L) jointly determine utility-based link recommendation decision (R); and (ii) structural proximity (S) and nodal proximity (N) together predict latent linkage likelihood (L).

4.2. Learning the Bayesian Network

To employ the Bayesian network to predict recommendation probability, we need to learn the distributions of

Figure 3. A Bayesian Network for Utility-Based Link Recommendation



its factors V, C, S, N, and L. As a first step of this learning task, we construct training data from observed link establishments in an online social network. Let t be the current time. To construct training data, we focus on potential links at time $t-1$, i.e., those that had not been established by time $t-1$. For a potential link i at time $t-1$, we can calculate its V_i, N_i, and S_i according to Equations (5), (7), and (8), respectively, and we estimate its C_i based on the status[4] of the social network at time $t-1$. It is noted that we can observe whether link i is established or not at current time t. Therefore, we set $R_i = 1$ if link i is ranked the top Kamong all potential links in terms of the utility of recommending a potential link defined in Equation (6), and we set $R_i = 0$ otherwise. We now have one training record, $O_i = \langle V_i, C_i, S_i, N_i, R_i \rangle$. Continuing the procedure for other potential links, we can construct training data $O = \{O_i\}$, where $i = 1, 2, \ldots, M$ and M is the number of records in the training data. We note that linkage likelihood L is unobserved. Thus, we do not have training data on L, but we need to learn the distribution of L, which is the key methodological challenge for learning the Bayesian network.

To learn the distributions of factors V, C, S, N, and L from training data O, we assume exponential family distributions for these factors by following a common Bayesian network learning procedure (Heckerman 2008). Thus, we need to learn parameters of these assumed distributions and denote the vector of these parameters as $\boldsymbol{\theta}$. Specifics about the assumed distributions and their parameters will be discussed later in this subsection. Parameters in $\boldsymbol{\theta}$ can be estimated as those that maximize the log-likelihood $H(O \mid \boldsymbol{\theta})$ of training data O given $\boldsymbol{\theta}$ (Friedman 1998, Mitchell 1997). Formally, the optimal estimation $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} H(O \mid \boldsymbol{\theta}), \tag{9}$$

where

$$H(O \mid \boldsymbol{\theta}) = \sum_{i=1}^{M} \ln[P(O_i \mid \boldsymbol{\theta})] \tag{10}$$

and $\ln[P(O_i \mid \boldsymbol{\theta})]$ is the log-likelihood of training record O_i given $\boldsymbol{\theta}$.

However, we cannot obtain $\boldsymbol{\theta}^*$ using Equation (9), because we do not have training data on linkage likelihood L but we need to estimate parameters for L. To address this challenge, we propose a Bayesian network learning algorithm based on the framework of expectation-maximization (EM), a framework for learning from incomplete data (Dempster et al. 1977). Following the EM framework (Dempster et al. 1977), our algorithm estimates parameters in $\boldsymbol{\theta}$ through an iterative process. In each iteration, our algorithm takes previous parameter estimates as input and produces updated parameter estimates by maximizing the objective function $Q(\cdot)$. Stated concretely, we have

$$\boldsymbol{\theta}_{k+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k), \tag{11}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta}_{k+1}$ denote the vector of parameter estimates in iterations k and $k+1$, respectively, and $k = 0, 1, 2, \ldots$. The objective function $Q(\cdot)$ is defined as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k) = \sum_{i=1}^{M} \int \ln[P(O_i, L_i \mid \boldsymbol{\theta})] P(L_i \mid O_i, \boldsymbol{\theta}_k) d_{L_i}, \tag{12}$$

where $\ln[P(O_i, L_i \mid \boldsymbol{\theta})]$ is the log-likelihood of complete data (including observed training data O_i and unobserved linkage likelihood L_i) given $\boldsymbol{\theta}$, and $P(L_i \mid O_i, \boldsymbol{\theta}_k)$ is the probability of L_i given O_i and previous parameter estimates $\boldsymbol{\theta}_k$. In Equation (12), $\int \ln[P(O_i, L_i \mid \boldsymbol{\theta})] P(L_i \mid O_i, \boldsymbol{\theta}_k) d_{L_i}$ represents the expected log-likelihood of complete data, expected on L_i. The iterative process stops if the absolute difference between $H(O \mid \boldsymbol{\theta}_{k+1})$ and $H(O \mid \boldsymbol{\theta}_k)$ is sufficiently small. According to Bishop (2006), the iterative process is guaranteed to converge, and the converged parameter estimation by our algorithm is a local optimal estimation of $\boldsymbol{\theta}$ for function $H(\cdot)$, defined in Equation (10). For parameter estimation from incomplete data, a local optimum is the best result possible (Bishop 2006).

While the above-discussed iterative process of learning $\boldsymbol{\theta}$ follows the standard EM framework (Dempster et al. 1977), the core of the process, i.e., how to compute $\boldsymbol{\theta}_{k+1}$ from $\boldsymbol{\theta}_k$ according to Equations (11) and (12), is specific to our study and represents the key methodological contribution of our proposed algorithm for learning the Bayesian network. In the following, we identify the parameters in $\boldsymbol{\theta}$ by properly decomposing the objective function $Q(\cdot)$, and then we show how to compute $\boldsymbol{\theta}_{k+1}$ from $\boldsymbol{\theta}_k$. To decide the specific parameters in $\boldsymbol{\theta}$, we rewrite the objective function $Q(\cdot)$ as

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}_k) = {} & \sum_{i=1}^{M} \ln[P(R_i \mid \boldsymbol{\theta})] + \sum_{i=1}^{M} \ln[P(V_i \mid R_i, \boldsymbol{\theta})] \\ & + \sum_{i=1}^{M} \ln[P(C_i \mid R_i, \boldsymbol{\theta})] \\ & + \sum_{i=1}^{M} \int \{\ln[P(S_i, L_i \mid R_i, \boldsymbol{\theta})] \\ & + \ln[P(N_i, L_i \mid R_i, \boldsymbol{\theta})] \\ & - \ln[P(L_i \mid R_i, \boldsymbol{\theta})]\} P(L_i \mid O_i, \boldsymbol{\theta}_k) d_{L_i} \end{aligned} \tag{13}$$

The derivation of Equation (13) is given in Online Appendix A. According to Equation (13), we need to estimate $P(R_i)$, $P(V_i \mid R_i)$, $P(C_i \mid R_i)$, $P(S_i, L_i \mid R_i)$, $P(N_i, L_i \mid R_i)$, and $P(L_i \mid R_i)$ for $R_i = 0, 1$. To estimate $P(R_i)$, we denote parameters $p_0 = P(R_i = 0)$ and $p_1 = P(R_i = 1)$. It is common to assume an exponential family distribution (e.g., exponential or normal) for

a continuous factor in a Bayesian network (Friedman 1998, Heckerman 2008). Since factor V is continuous and positive, we assume an exponential distribution for factor V given R. Accordingly, we can estimate $P(V_i \mid R_i = 0)$ with an exponential density $\lambda_V^0 e^{-\lambda_V^0 \times V_i}$ and $P(V_i \mid R_i = 1)$ with an exponential density $\lambda_V^1 e^{-\lambda_V^1 \times V_i}$. Hence, to estimate $P(V_i \mid R_i)$, we need to estimate parameter $\lambda_V^{R_i}$ for $R_i = 0, 1$. In a similar way, we assume factor C given R following an exponential distribution, and we estimate $P(C_i \mid R_i)$ with its density. Thus, to estimate $P(C_i \mid R_i)$, we need to estimate parameter $\lambda_C^{R_i}$ for $R_i = 0, 1$.

Similarly, we assume that the joint distribution of factors S and L given R follows a bivariate exponential distribution, and we estimate $P(S_i, L_i \mid R_i)$ using its density. In particular, Freund (1961) defines the density function of a bivariate exponential distribution as

$$f(x,y) = \begin{cases} \lambda_x \lambda'_y e^{-\lambda'_y \cdot y - (\lambda_x + \lambda_y - \lambda'_y)\cdot x} & \text{for } 0 < x < y, \\ \lambda_y \lambda'_x e^{-\lambda'_x \cdot x - (\lambda_x + \lambda_y - \lambda'_x)\cdot y} & \text{for } 0 < y < x. \end{cases} \quad (14)$$

Using the bivariate exponential density defined in Equation (14) to estimate $P(S_i, L_i \mid R_i)$, we have

$$P(S_i, L_i \mid R_i) = \begin{cases} \lambda_S^{R_i} \lambda_L'^{R_i} e^{-\lambda_L'^{R_i} \cdot L_i - (\lambda_S^{R_i} + \lambda_L^{R_i} - \lambda_L'^{R_i})\cdot S_i} \\ \quad \text{for } 0 < S_i < L_i, \\ \lambda_L^{R_i} \lambda_S'^{R_i} e^{-\lambda_S'^{R_i} \cdot S_i - (\lambda_S^{R_i} + \lambda_L^{R_i} - \lambda_S'^{R_i})\cdot L_i} \\ \quad \text{for } 0 < L_i < S_i. \end{cases} \quad (15)$$

We thus need to estimate parameters $\lambda_S^{R_i}$, $\lambda_S'^{R_i}$, $\lambda_L^{R_i}$, and $\lambda_L'^{R_i}$ for $R_i = 0, 1$. In a similar way, we assume the joint distribution of factors N and L given R following a bivariate exponential distribution, and we estimate $P(N_i, L_i \mid R_i)$ as

$$P(N_i, L_i \mid R_i) = \begin{cases} \lambda_N^{R_i} \lambda_L'^{R_i} e^{-\lambda_L'^{R_i} \cdot L_i - (\lambda_N^{R_i} + \lambda_L^{R_i} - \lambda_L'^{R_i})\cdot N_i} \\ \quad \text{for } 0 < N_i < L_i, \\ \lambda_L^{R_i} \lambda_N'^{R_i} e^{-\lambda_N'^{R_i} \cdot N_i - (\lambda_N^{R_i} + \lambda_L^{R_i} - \lambda_N'^{R_i})\cdot L_i} \\ \quad \text{for } 0 < L_i < S_i. \end{cases} \quad (16)$$

Hence, we need to estimate parameters $\lambda_N^{R_i}$, $\lambda_N'^{R_i}$, $\lambda_L^{R_i}$, and $\lambda_L'^{R_i}$ for $R_i = 0, 1$. Finally, we need to estimate $P(L_i \mid R_i)$. Factor L participates in the joint distribution with factor S and the joint distribution with factor N, whose densities are defined in Equations (15) and (16), respectively. Given R, factor L follows an exponential distribution with parameter λ_L^R if $L < \min(S, N)$ and with parameter $\lambda_L'^R$ otherwise (Freund 1961), where $\min(x, y)$ returns the minimum between x and y. We thus have

$$P(L_i \mid R_i) = \begin{cases} \lambda_L^{R_i} e^{-\lambda_L^{R_i} \cdot L_i} & \text{for } L_i < \min(S_i, N_i), \\ \lambda_L'^{R_i} e^{-\lambda_L'^{R_i} \cdot L_i} & \text{for } L_i \geq \min(S_i, N_i). \end{cases} \quad (17)$$

In sum, the parameter vector $\boldsymbol{\theta}$ to be estimated is $\boldsymbol{\theta} = \langle p_0, p_1, \lambda_V^0, \lambda_V^1, \lambda_C^0, \lambda_C^1, \lambda_S^0, \lambda_S^1, \lambda_S'^0, \lambda_S'^1, \lambda_N^0, \lambda_N^1, \lambda_N'^0, \lambda_N'^1, \lambda_L^0, \lambda_L^1, \lambda_L'^0, \lambda_L'^1 \rangle$. We next show how to compute $\boldsymbol{\theta}$ that maximizes the objective function $Q(\cdot)$ defined in Equation (12).

Theorem 1. *Given the previous parameter estimation* $\boldsymbol{\theta}_k = \langle \bar{p}_0, \bar{p}_1, \bar{\lambda}_V^0, \bar{\lambda}_V^1, \bar{\lambda}_C^0, \bar{\lambda}_C^1, \bar{\lambda}_S^0, \bar{\lambda}_S^1, \bar{\lambda}_S'^0, \bar{\lambda}_S'^1, \bar{\lambda}_N^0, \bar{\lambda}_N^1, \bar{\lambda}_N'^0, \bar{\lambda}_N'^1, \bar{\lambda}_L^0, \bar{\lambda}_L^1, \bar{\lambda}_L'^0, \bar{\lambda}_L'^1 \rangle$ *and the exponential distribution assumption for factors* V, C, S, N, *and* L, *there exists a single optimal solution of* $\boldsymbol{\theta}$ *that maximizes the objective function defined in Equation* (12), *and the optimal solution is of closed form.*

Proof. See Online Appendix B for the proof and closed-form solution of $\boldsymbol{\theta}$.

The existence of a single closed-form solution of $\boldsymbol{\theta}$, as discovered by Theorem 1, is an attractive property because a closed-form solution of parameter estimates not only greatly simplifies the implementation of an EM-based algorithm but also considerably improves its computation efficiency (McLachlan and Krishnan 2007). Armed with Theorem 1, we propose an algorithm to learn $\boldsymbol{\theta}$—namely, the Bayesian network learning with continuous latent factor (BNLF) algorithm. As shown in Figure 4, the algorithm starts with an initial estimation $\boldsymbol{\theta}_0$ of $\boldsymbol{\theta}$ and iteratively updates its estimation according to Theorem 1 until convergence. To obtain $\boldsymbol{\theta}_0$, we follow a common method of parameter initialization, repeated random initialization (Duda and Hart 1973). Specifically, we randomly sample one-third of the training data and estimate parameters in $\boldsymbol{\theta}_0$ according to the sample. Details of the $\boldsymbol{\theta}_0$ estimation are given in Online Appendix C. We then run the BNLF algorithm with $\boldsymbol{\theta}_0$ and obtain one $\hat{\theta}$. The above process is repeated three times. We finally choose the $\hat{\theta}$ that has the largest log-likelihood $H(O \mid \hat{\theta})$ among the three obtained $\hat{\theta}$ values.

4.3. Predicting Recommendation Probability

Having obtained parameter estimation $\hat{\boldsymbol{\theta}} = \langle \hat{p}_0, \hat{p}_1, \hat{\lambda}_V^0, \hat{\lambda}_V^1, \hat{\lambda}_C^0, \hat{\lambda}_C^1, \hat{\lambda}_S^0, \hat{\lambda}_S^1, \hat{\lambda}_S'^0, \hat{\lambda}_S'^1, \hat{\lambda}_N^0, \hat{\lambda}_N^1, \hat{\lambda}_N'^0, \hat{\lambda}_N'^1, \hat{\lambda}_L^0, \hat{\lambda}_L^1, \hat{\lambda}_L'^0, \hat{\lambda}_L'^1 \rangle$, we are ready to predict the recommendation

Figure 4. The BNLF Algorithm for Learning $\boldsymbol{\theta}$

BNLF (O, ϵ)
  O: training data
  ϵ: predefined convergence threshold
Initialize $\boldsymbol{\theta}_0$. // $\boldsymbol{\theta}_0$: initial estimation of $\boldsymbol{\theta}$
$k = -1$.
Do
  $k = k + 1$.
  Obtain $\boldsymbol{\theta}_{k+1}$ according to Equation (11) and Theorem 1.
While $(|H(O \mid \boldsymbol{\theta}_{k+1}) - H(O \mid \boldsymbol{\theta}_k)| > \epsilon)$
  // $H(\cdot)$: log-likelihood defined in Equation (10)
$\hat{\theta} = \boldsymbol{\theta}_{k+1}$. // $\hat{\theta}$: final estimation of $\boldsymbol{\theta}$
Return $\hat{\theta}$.

probability for each potential link. In an online social network, for a potential link a at current time t, i.e., a link that has not been established by current time t, we can calculate its V_a, N_a, and S_a according to Equations (5), (7), and (8), respectively; we estimate its C_a based on the status of the network at current time t. We define the recommendation probability for potential link a as the probability of recommending it given its V_a, C_a, S_a, N_a, and parameter estimation $\hat{\theta}$; i.e., $P(R_a = 1 \mid V_a, C_a, S_a, N_a, \hat{\theta})$. In Online Appendix D, we show how to compute recommendation probability $P(R_a = 1 \mid V_a, C_a, S_a, N_a, \hat{\theta})$. Applying the formula for computing recommendation probability given in Online Appendix D, we can predict recommendation probability for each potential link and recommend the top K potential links with the highest recommendation probabilities.

5. Empirical Evaluation

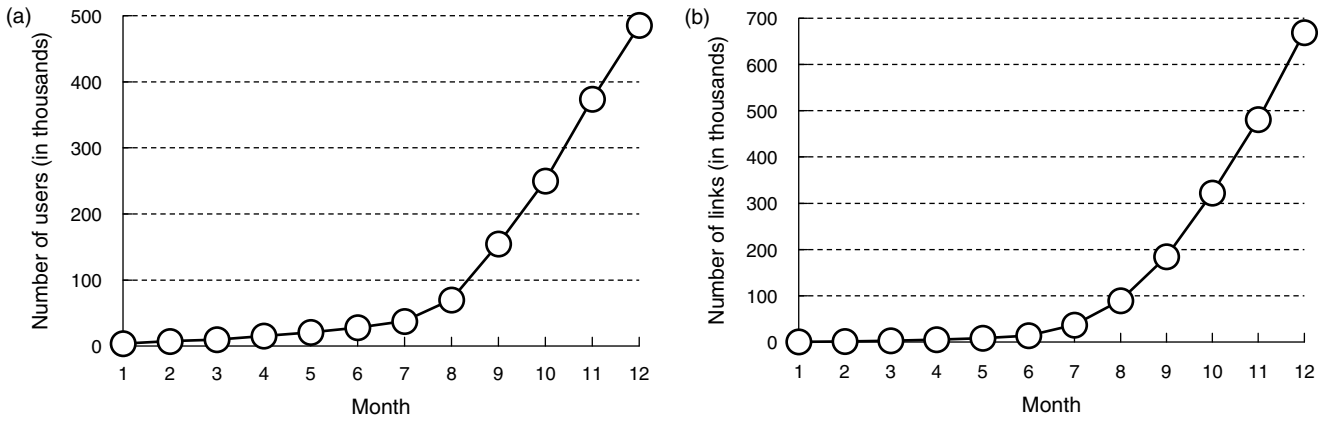
We conducted experiments to evaluate our method using real-world social network data. In this section, we describe the data and parameter calibration; detail our experimental procedure, evaluation metrics, and benchmark methods; and report experimental results.

5.1. Data and Parameter Calibration

We collected data from a major U.S. online social network over a one-year period, starting from the launch of the network. One collected data set describes who registered on what date as a user of the online social network; another data set contains data on who is linked to whom and when the linkage was established. As shown in Figure 5, both the number of users and the number of links grow rapidly over time. At the end of the one-year period, the online social network had 485,608 users, connected by 669,524 links. For each user, data about his or her profile were also collected. Because of privacy concerns, the profile of a user consists of a set of encoded terms, each corresponding to a characteristic of the user. During the one-year period, no link recommendation mechanism had been deployed in the online social network. Thus, our data provide a natural test bed for evaluating different link recommendation methods.

We then calibrated parameters for the utility-based link recommendation problem. Following Jackson (2008), we set decay factor α in Equation (2) to 0.5. Parameter m_j in Equation (2) was estimated as the revenue generated if an advertisement initiated by user w_j reaches a (direct or indirect) neighbor of w_j, multiplied by the average number of advertisements initiated by a user in a month.[5] Cost C_{jh} in Equation (6) was initially treated as the opportunity cost of not being able to recommend another potential link to user w_j or w_h because of recommending e_{jh}. We thus estimated C_{jh} as the average value of links actually established by user w_j or w_h, where the value of a link was computed using Equation (5). For robustness analysis, we conducted additional experiments with $\rho \times$ intial cost estimation, where $\rho = 0.5, 2$. Following a common practice in link recommendation (Backstrom and Leskovec 2011, Wang et al. 2011, Dong et al. 2012), we focus on potential links that, if established, would connect users who are two hops away. Compared with considering all possible potential links (i.e., potential links that would connect users two or more hops away), focusing on potential links that would connect users two hops away can greatly accelerate computation without sacrificing much on prediction (Backstrom and Leskovec 2011, Wang et al. 2011, Dong et al. 2012). In our experiments, the number of potential links that would connect users two hops away increases over time because the number of users increases over time.[6] Hence, rather than setting K to a static number, we set K as a percentage of potential links in a month. We therefore set $K = 0.5\% \times$ number of potential links in a month, i.e., recommending the top 0.5% of potential links in a month. To ensure the robustness of our empirical evaluation,

Figure 5. The Growth of Users (a) and Links (b) in the Online Social Network



we further conducted experiments with $K = 0.25\% \times$ number of potential links in a month and with $K = 0.75\% \times$ number of potential links in a month.

5.2. Experimental Design

Our experiments follow the procedure described below. Let month t be the current month, and let month $t+1$ be the prediction month. We use data by the current month t to train our method as well as each benchmark method to recommend the top K potential links[7] out of potential links in month t. Using data by prediction month $t+1$, we can verify whether a potential link in month t is actually accepted and established by users in month $t+1$, compute its utility with Equation (6), and identify true top K potential links that have the highest utilities. The performance of a method is then evaluated using *top K utility-based precision*, which is the fraction of recommended top K potential links that are true top K potential links. In addition, we also evaluate the *average utility* of links recommended by a method. After all, the objective of the utility-based link recommendation problem is to recommend links with the highest utilities. Therefore, a method that better achieves the objective recommends links with higher average utility.

We next discuss implementation details of our method and benchmark methods. For our method, nodal proximity between users was measured using Equation (7), for which we chose the Jaccard coefficient as the similarity function. In our experiments, the user profile is represented by a set of terms, and the Jaccard coefficient is suitable for measuring similarity between sets (Salton and McGill 1983). Specifically, the Jaccard coefficient is defined as

$$\mathrm{sim}(\mathbf{r}_j, \mathbf{r}_h) = \frac{|\mathbf{r}_j \cap \mathbf{r}_h|}{|\mathbf{r}_j \cup \mathbf{r}_h|}, \tag{18}$$

where $\mathbf{r}_j$ and $\mathbf{r}_h$ denote the set of profile terms of users w_j and w_h, respectively, and $|\cdot|$ is the cardinality of a set. Structural proximity between users was computed using Equation (8). Following a common practice (Liben-Nowell and Kleinberg 2007, Lichtenwalter et al. 2010), we set β in Equation (8) to 0.05.

We selected a representative method from each category of existing link recommendation methods as a benchmark. For the category of learning-based methods, we chose the SVM-based link recommendation method because of its outstanding performance among learning-based methods (Al Hasan et al. 2006). Our implementation of the SVM-based method followed its implementation described in Al Hasan et al. (2006), Lichtenwalter et al. (2010). Among structural proximity-based methods, the Katz index was selected because of its superior performance (Liben-Nowell and Kleinberg 2007). A suitable nodal proximity-based method for our experiments is the Jaccard coefficient (i.e., Equation (18)), thus chosen as a benchmark. Moreover, we benchmarked our method against a link recommendation method commonly used in practice, common neighbor.[8] In particular, the common neighbor CN_{jh} between users w_j and w_h is computed as the number of mutual neighbors (Newman 2001, Liben-Nowell and Kleinberg 2007):

$$CN_{jh} = |\Gamma_j \cap \Gamma_h|, \tag{19}$$

where Γ_j and Γ_h denote the set of direct neighbors of users w_j and w_h, respectively. We also benchmarked our method against a variation of the common neighbor, Adamic/Adar (Adamic and Adar 2003). Extended from the common neighbor, the Adamic/Adar AA_{jh} between users w_j and w_h is given by Adamic and Adar (2003), and Liben-Nowell and Kleinberg (2007):

$$AA_{jh} = \sum_{w_z \in \Gamma_j \cap \Gamma_h} \frac{1}{\log |\Gamma_z|}, \tag{20}$$

where Γ_j , Γ_h, and Γ_z denote the set of direct neighbors of users w_j, w_h, and w_z, respectively, and w_z is a common neighbor of users w_j and w_h. Table 1 summarizes the benchmark methods.

5.3. Experimental Results and Analysis

Following the experimental procedure, we conducted experiments with the parameter values as set in Section 5.1 and current month $t = 2, 3, \ldots, 11$.[9] As shown in Table 2, our method substantially outperforms each benchmark method in every prediction month, in terms of top K utility-based precision. Averaged across prediction months, the mean top K utility-based precision of our method is 0.40, which indicates that, on average, 40% of top K potential links recommended by our method are true top K links. On the other hand, the mean top K utility-based precision of SVM, the best-performing benchmark method in terms of top K utility-based precision, is only 0.27. On average, the top K utility-based precision of our method is 45.52% higher than that of SVM and 122.25% higher than that of CN, a link recommendation method commonly used in practice. The outperformance of our method over benchmark methods in terms of top K utility-based precision is due to its consideration of utility factors

Table 1. Summary of Benchmark Methods

Method	Category	Abbreviation
SVM-based link recommendation	Learning	SVM
Katz index	Structural proximity	Katz
Jaccard coefficient	Nodal proximity	Jaccard
Common neighbor	Structural proximity	CN
Adamic/Adar	Structural proximity	AA

Table 2. Top K Utility-Based Precision: Our Method vs. Benchmark Methods

Prediction month ($t+1$)	AA	CN	Jaccard	Katz	SVM	Our method
3	0.18	0.17	0.05	0.25	0.28	0.46
4	0.19	0.20	0.04	0.27	0.31	0.42
5	0.16	0.16	0.05	0.20	0.26	0.36
6	0.16	0.20	0.03	0.23	0.27	0.41
7	0.17	0.16	0.03	0.22	0.26	0.37
8	0.18	0.18	0.03	0.25	0.25	0.35
9	0.16	0.17	0.03	0.22	0.25	0.37
10	0.16	0.16	0.02	0.22	0.24	0.39
11	0.19	0.20	0.02	0.27	0.29	0.45
12	0.20	0.21	0.03	0.28	0.30	0.39
Mean	0.17	0.18	0.03	0.24	0.27	0.40
Std. dev.	0.01	0.02	0.01	0.03	0.02	0.04

such as value (V) and cost (C) as well as its better link prediction accuracy. Here, the link prediction accuracy of a method is the fraction of potential links recommended by this method that are actually established. For example, averaged across prediction months, our method outperforms SVM by 24.11% in terms of link prediction accuracy.

Table 3 compares the average utility of links recommended by our method against that by each benchmark method. As shown, our method significantly outperforms each benchmark method across prediction months, in terms of average utility. Averaged across prediction months, our method outperforms SVM (the best-performing benchmark method in terms of average utility) by 41.76% and CN (a method commonly used in practice) by 210.46%. For example, in prediction month 12, the average utility of links recommended by our method is \$1.87, which means that, on average, a link recommended by our method could bring \$1.87 to the operator of the online social network. In comparison, the average utility of links recommended by SVM is \$1.56, and the average utility of links recommended by CN is \$0.71. In other words, a link recommended by our method could generate an average of \$0.31 more revenue than a link recommended by SVM and an average of \$1.16 more revenue than a link recommended by CN. Considering the sheer number of links recommended in the online social network, successful implementation of our method could bring significant financial gains to the network's operator, compared with benchmark methods.

Table 3. Average Utility: Our Method vs. Benchmark Methods

Prediction month ($t+1$)	AA (\$)	CN (\$)	Jaccard (\$)	Katz (\$)	SVM (\$)	Our method (\$)
3	0.29	0.21	0.10	0.44	0.92	1.15
4	0.44	0.47	0.12	0.71	0.82	1.14
5	0.27	0.25	0.06	0.32	0.76	0.99
6	0.30	0.33	0.07	0.36	0.51	0.83
7	0.28	0.26	0.04	0.39	0.54	0.78
8	0.37	0.38	0.05	0.55	0.56	0.97
9	0.36	0.37	0.05	0.52	0.74	0.99
10	0.37	0.38	0.04	0.55	0.74	1.17
11	0.60	0.64	0.06	0.99	1.35	1.77
12	0.65	0.71	0.06	1.07	1.56	1.87
Mean	0.39	0.40	0.07	0.59	0.85	1.17
Std. dev.	0.13	0.16	0.02	0.26	0.35	0.37

The performance improvement by our method is attributed to its methodological design. Our method is designed to recommend a link based on its value, cost, and linkage likelihood—essential factors determining the utility of a link—whereas benchmark methods recommend a link on the basis of its linkage likelihood only. As a result, the links recommended by our method generally have higher utilities than those recommended by benchmark methods. While it is necessary to consider a link's value and cost when making link recommendation decision, it is also essential to take into account its linkage likelihood. In our method, linkage likelihood is treated as a latent factor. Without this latent factor, a simple solution to the utility-based link recommendation problem is a naïve Bayes (NB) method, which predicts the link recommendation decision (R) from input factors: value (V), cost (C), structural proximity (S), and nodal proximity (N). Table 4 compares the performance of our method against that of NB. Averaged across prediction months, our method outperforms NB by 80.38% in terms of top K utility-based precision and 20.22% in terms of average utility. Such improvements echo the necessity of linkage likelihood in link recommendation decision.

In summary, our empirical results show that our method substantially outperforms representative link recommendation methods from prior research. To ensure the robustness of our empirical findings, we

Table 4. Performance Comparison Between Our Method and NB

Prediction month ($t+1$)	Top K utility-based precision		Average utility (\$)	
	NB	Our method	NB	Our method
3	0.31	0.46	1.04	1.15
4	0.26	0.42	1.02	1.14
5	0.21	0.36	0.82	0.99
6	0.19	0.41	0.63	0.83
7	0.18	0.37	0.61	0.78
8	0.19	0.35	0.80	0.97
9	0.18	0.37	0.83	0.99
10	0.20	0.39	0.91	1.17
11	0.25	0.45	1.48	1.77
12	0.23	0.39	1.70	1.87
Mean	0.22	0.40	0.98	1.17
Std. dev.	0.04	0.04	0.35	0.37

conducted experiments with different cost estimations or K; experimental results reported in Online Appendix E further confirm the performance advantage of our method over benchmark methods. We also show the contribution of each component of our method to its performance in Online Appendix F and demonstrate the outperformance of our method over benchmark methods using another social network data set in Online Appendix G.

6. Conclusions

Link recommendation is a key functionality offered by major online social networks. Existing methods for link recommendation focus on the likelihood of linkage but overlook the benefit of linkage. Our study addresses this limitation and contributes the innovative idea of utility-based link recommendation to extant literature. First, we define the utility of recommending a potential link and formulate a new link recommendation problem. Second, we propose a novel utility-based link recommendation method that recommends links based on the value, cost, and linkage likelihood of a potential link, in contrast to existing link recommendation methods that focus solely on linkage likelihood. Our empirical evaluation demonstrates the performance advantage of our proposed method over prevalent link recommendation methods found in representative prior research. Third, our study also contributes a novel problem and method to the utility-based data mining literature (Weiss et al. 2008, Saar-Tsechansky et al. 2009).

Our study has several managerial implications. First, our research sheds light on the critical role of balancing operator's benefit and users' needs in link recommendation. After all, a potential link can benefit an operator only after it is established by users. In this vein, the effectiveness of link recommendation depends on the proper consideration of the value, cost, and linkage likelihood of a potential link. Failing to consider any of the three factors greatly reduces the effectiveness of a link recommendation method, as demonstrated in our empirical analysis. Our method innovatively integrates these three factors in link recommendation and offers great value to significant applications. For example, we show that, on average, a link recommended by our method can produce 41.76% more utility than a link recommended by the best-performing benchmark method. Given that advertisement alone is estimated to generate $12 billion revenue to operators of online social networks in 2014 (*eMarketer* 2012), such improvements could create huge financial gains for them.

Second, an important reason our method yields better performance over benchmark methods is the consideration of the network value that a user brings to an operator. In an online social network, users are neither isolated nor independent. On the contrary, they are connected and influence each other. As a result, a user's network value arises from his or her influence on his or her direct and indirect neighbors in a network. Since the major revenue source for an operator comes from advertisement, and pay-per-impression as well as pay-per-click continue to be prevalent business models for online social networks, network value is critical to the business performance of these networks. Therefore, the operator of an online social network should treat network value as a key factor in crucial decisions such as link recommendation and targeted marketing.

Third, our study highlights the important role of linkage likelihood in determining utilities of recommended links. In this light, the operator of an online social network could purposefully boost linkage likelihood such that both the operator and users are benefited. For example, an operator can enhance features that accentuate nodal and structural similarity between users, two factors that jointly determine linkage likelihood. This in turn enables a user to be more aware of other users who are similar to the user, which generally increases the likelihood of link formation. Another promising approach is to provide incentives to facilitate the establishment of potential links. Rather than passively waiting for users to make connections, an operator can actively intervene and provide incentives to lure users to connect. Such approach could be especially useful for potential links with moderate linkage likelihood but high value.

Our research has limitations that should be addressed in future research endeavors. In our Bayesian network model, we assume mutual independences among factors value (V), cost (C), structural proximity (S), and nodal proximity (N). Future work should examine how to relax this assumption. One viable approach is to learn dependency relationships among these factors from data (Heckerman 2008). Such an approach could improve the predictive effectiveness of our method, although it might increase its computational complexity. In addition, we assume that a better-connected social network is more effective in facilitating information diffusion over the network. Future work should also consider users' susceptibilities to information diffused over a network. Moreover, we assume that effective information diffusion over a social network is beneficial for the network's operator and study the positive side of link recommendation, i.e., the utility of link recommendation. Future work should examine the negative aspect of link recommendation, such as the rapid spread of negative sentiments about a product over a more connected social network as a result of link recommendation.

Additionally, there are research questions worthy of future exploration. First, it would be interesting to study how to recommend a set of potential links that

collectively have the highest utility. Second, the effectiveness of the learned Bayesian network in predicting recommendation probabilities could decline over time because it is learned from previous user linkage behaviors and does not capture new user linkage behaviors. Thus, another interesting question is how to maintain the currency of the learned Bayesian network over time. Prior research on knowledge refreshing and maintenance (Bensoussan et al. 2009, Fang et al. 2013a) could provide theoretical foundations for this question. Another area worthy of future investigation is to characterize potential links that can increase the value of a social network the most, where value increase as a result of the addition of a potential link is defined in Equation (5). Some recent exploration in this direction has found that a potential link that reduces the clustering coefficient of a social network could increase the value of the network (Zhao et al. 2012). Such a finding could be used in combination with linkage likelihood predicted by an existing link prediction method to recommend links with high utilities. It is also worthwhile to conduct field experiments to evaluate our method. In an experiment, we can observe in real time how users react to recommended links, which recommended links they actually establish, and the values of these established links. By combining evaluation results with archival data and field experimental results, we could produce more comprehensive evidence on the effectiveness of our method. Finally, it would be interesting to extend our method by considering a user's historical link adoption record, which documents all other users to whom the user has already linked, i.e., direct neighbors of the user. Understandably, the more similar user w_j to user w_h's direct neighbors, the more likely w_j is the kind of person with whom w_h likes to connect and the higher the linkage likelihood between w_j and w_h. Therefore, future work needs to incorporate the similarities between a user and each direct neighbor of another user into our method.

Acknowledgments

The authors thank department editor Anandhi Bharadwaj, the associate editor, and four anonymous reviewers for their guidance and constructive comments that have tremendously improved the paper. The authors are also grateful to the late Sandra Slaughter for her guidance in the early stage of this paper. Z. Li and X. Fang contributed equally to the paper.

Endnotes

[1] A potential link refers to a link that has not been established.

[2] LinkedIn offers two types of user accounts: basic and premium. Whereas a basic account is free, LinkedIn Corporation charges a membership fee for premium accounts.

[3] The K recommended links are for all users. A user receives a recommendation if a recommended link has the user as an end point.

[4] The status of a social network such as its structure evolves over time.

[5] Because of privacy concerns, we do not have data on the number of advertisements initiated by a specific user, only the average number across users.

[6] The number of potential links that would connect users two hops away is as follows: 110,475 (1), 176,216 (2), 252,963 (3), 483,948 (4), 709,991 (5), 931,623 (6), 1,333,526 (7), 3,291,007 (8), 8,808,065 (9), 15,988,577 (10), 26,317,250 (11), and 38,517,866 (12). Here, the month number is enclosed in the parentheses.

[7] The K recommended links are for all users. By simply adjusting its output, our method can target a user and recommend the same number of links (say, m links) to each user. Specifically, for a user, our method can be adapted to identify and recommend m potential links that have the highest recommendation probabilities among potential links with this user as an end point.

[8] Common neighbor is popularly used by major online social networks for link recommendation. For example, it is called "mutual friend" in Facebook and "shared connection" in LinkedIn.

[9] To construct training data for our method, we need data in month $t-1$ and month t. Thus, current month t in our experiments starts from month 2 instead of month 1.

References

Adali S, Sisenda F, Magdon-Ismail M (2012) Actions speak as loud as words: Predicting relationships from social behavior data. *Proc. 21st Internat. Conf. World Wide Web* (ACM, New York), 689–698.

Adamic LA, Adar E (2003) Friends and neighbors on the web. *Soc. Networks* 25(3):211–230.

Al Hasan M, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. *Proc. SIAM Workshop on Link Anal., Counterterrorism Security*, Society for Industrial and Applied Mathematics, Philadelphia.

Backstrom L, Leskovec J (2011) Supervised random walks: Predicting and recommending links in social networks. *Proc. 4th ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 635–644.

Ballester C, Calvó-Armengol A, Zenou Y (2006) Who's who in networks. Wanted: The key player. *Econometrica* 74(5):1403–1417.

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

Barabási A, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A: Statist. Mechanics Appl.* 311(3):590–614.

Benchettara N, Kanawati R, Rouveirol C (2010) Supervised machine learning applied to link prediction in bipartite social networks. Memon N, Alhajj R, eds. *Proc. 2nd Internat. Conf. Adv. Soc. Network Anal. Mining* (IEEE Computer Society, Washington, DC), 326–330.

Bensoussan AR, Mookerjee V, Mookerjee W, Yue T (2009) Maintaining diagnostic knowledge-based systems: A control theoretic approach. *Management Sci.* 55(2):294–310.

Bishop CM, Nasrabadi NM (2006) *Pattern Recognition and Machine Learning* (Springer, New York).

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. *Comput. Networks* 30(1–7):107–117.

Chen J, Geyer W, Dugan C, Muller M, Guy I (2009) Make new friends, but keep the old: Recommending people on social networking sites. Greenberg S, Hudson SE, Hinckley K, Morris MR, Olsen DR Jr, eds. *Proc. 27th Internat. Conf. Human Factors Comput. Systems* (ACM, New York), 201–210.

Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. *Proc. Natl. Acad. Sci. USA* 107(52):22436–22441.

Davenport T, Patil DJ (2012) Data scientist: The sexist job of the 21st century. *Harvard Bus. Rev.* 90(10):70–76.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39(1):1–38.

Domingos P, Richardson M (2001) Mining the network value of customers. *Proc. 7th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 57–66.

Dong Y, Tang J, Wu S, Tian J, Chawla NV, Rao J, Cao H (2012) Link prediction and recommendation across heterogeneous social networks. Zaki MJ, Siebes A, Yu JX, Goethals B, Webb G, Wu X, eds. *Proc. 12th Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 181–190.

Doreian P (1989) Two regimes of network autocorrelation. Kochen M, ed. *The Small World* (Ablex, Norwood, NJ), 280–295.

Duda RO, Hart PE (1973) *Pattern Classification and Scene Analysis* (John Wiley & Sons, New York).

Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "friends": Social capital and college students' use of online social network sites. *J. Comput.-Mediated Comm.* 12(4):1143–1168.

eMarketer (2012) Total worldwide social network ad revenues continue strong growth. (February 24), http://www.emarketer.com/Article/Total-Worldwide-Social-Network-Ad-Revenues-Continue-Strong-Growth/1008862.

Facebook Inc. (2013) Form 10-K for the fiscal year ended December 31, 2013. Accessed February 1, 2015, https://www.sec.gov/Archives/edgar/data/1326801/000132680114000007/fb-12312013x10k.htm.

Fang X, Sheng ORL, Goes P (2013a) When is the right time to refresh knowledge discovered from data? *Oper. Res.* 61(1):32–44.

Fang X, Hu P, Li Z, Tsai W (2013b) Predicting adoption probabilities in social networks. *Inform. Systems Res.* 24(1):128–145.

Fouss F, Pirotte A, Renders JM, Saerens M (2007) Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Trans. Knowledge Data Engrg.* 19(3):355–369.

Freund JE (1961) A bivariate extension of the exponential distribution. *J. Amer. Statist. Assoc.* 56(296):971–977.

Friedman N (1998) The Bayesian structural EM algorithm. Cooper G, Moral S, eds. *Proc. 14th Conf. Uncertainty Artificial Intelligence* (Morgan Kaufmann, San Francisco), 129–138.

Gong NZ, Talwalkar A, Mackey L, Huang L, Shin ECR, Stefanov E, Shi ER, Song D (2014) Joint link prediction and attribute inference using a social-attribute network. *ACM Trans. Intelligent Systems Technol. (TIST)* 5(2):27.

Granovetter MS (1973) The strength of weak ties. *Amer. J. Sociol.* 78(6):1360–1380.

Granovetter M (2005) The impact of social structure on economic outcomes. *J. Econom. Perspect.* 19(1):33–50.

Heckerman D (2008) A tutorial on learning with Bayesian networks. Holmes D, Jain L, eds. *Innovations in Bayesian Networks* (Springer-Verlag, Berlin), 33–82.

Heider F (1958) *The Psychology of Interpersonal Relations* (John Wiley & Sons, New York).

Hopcroft J, Lou T, Tang J (2011) Who will follow you back? Reciprocal relationship prediction. Berendt B, de Vries A, Fan W, MacDonald G, Ounis I, Ruthven I, eds. *Proc. 20th ACM Internat. Conf. Inform. Knowledge Management* (ACM, New York), 1137–1146.

Huberman BA, Romero DM, Wu F (2009) Social networks that matter: Twitter under the microscope. *First Monday* 14(1). http://firstmonday.org/ojs/index.php/fm/article/view/2317/2063.

Jackson MO (2008) *Social and Economic Networks* (Princeton University Press, Princeton, NJ).

Jackson MO, Rogers BW (2005) The economics of small worlds. *J. Eur. Econom. Assoc.* 3(2–3):617–627.

Jackson MO, Wolinsky A (1996) A strategic model of social and economic networks. *J. Econom. Theory* 71(1):44–74.

Jeh G, Widom J (2002) SimRank: A measure of structural-context similarity. *Proc. 8th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 538–543.

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. *Bus. Horizons* 53(1):59–68.

Katz L (1953) A new status index derived from sociometric analysis. *Psychometrika* 18(1):39–43.

Kunegis J, De Luca EW, Albayrak S (2010) The link prediction problem in bipartite networks. Hüllermeier E, Kruse R, Hoffmann F, eds. *Computational Intelligence for Knowledge-Based Systems Design*, Lecture Notes in Artificial Intelligence, Vol. 6178 (Springer-Verlag, Berlin), 380–389.

Kuo TT, Rui Y, Huang YY, Kung PH, Lin SD (2013) Unsupervised link prediction using aggregative statistics on heterogeneous social networks. Dhillon IS, Koren Y, Ghani R, Senator TE, Bradley P, Parekh R, He J, Grossman RL, Uthurusamy R, eds. *Proc. 19th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 775–783.

Liben-Nowell D, Kleinberg J (2007) The link prediction problem for social networks. *J. Amer. Soc. Inform. Sci. Tech.* 58(7):1019–1031.

Lichtenwalter RN, Lussier JT, Chawla NV (2010) New perspectives and methods in link prediction. *Proc. 16th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 243–252.

LinkedIn Corporation (2014) Form 10-K for the fiscal year ended December 31, 2013. Accessed February 1, 2015, https://www.sec.gov/Archives/edgar/data/1271024/000144530514000439/a20131231-10xkdocument.htm.

McLachlan G, Krishnan T (2007) *The EM Algorithm and Extensions* (John Wiley & Sons, Hoboken, NJ).

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

Mitchell TM (1997) *Machine Learning* (McGraw-Hill, New York).

Newman MEJ (2001) The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci. USA* 98(2):404–409.

Newman MEJ, Watts DJ, Strogatz SH (2002) Random graph models of social networks. *Proc. Natl. Acad. Sci. USA* 99(1):2566–2572.

O'Madadhain J, Hutchins J, Smyth P (2005) Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations* 7(2):23–30.

Popescul A, Ungar L (2003) Statistical relational learning for link prediction. Gottlob G, Walsh T, eds. *Workshop Learn. Statist. Models Relational Data Internat. Joint Conf. Artificial Intelligence* (Morgan Kaufmann, San Francisco), 81–90.

Quercia D, Capra L (2009) FriendSensing: Recommending friends using mobile phones. *Proc. 3rd ACM Conf. Recommender Systems* (ACM, New York), 273–276.

Saar-Tsechansky M, Melville P, Provost F (2009) Active feature-value acquisition. *Management Sci.* 55(4):664–684.

Salton G, McGill MJ (1983) *Introduction to Modern Information Retrieval* (McGraw-Hill, New York).

Scellato S, Noulas A, Mascolo C (2011) Exploiting place features in link prediction on location-based social networks. *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 1046–1054.

Schifanella R, Barrat A, Cattuto C, Markines B, Menczer F (2010) Folks in folksonomies: Social link prediction from shared metadata. *Proc. 3rd ACM Internat. Conf. Web Search Data Mining* (ACM, New York), 271–280.

Shen D, Sun JT, Yang Q, Chen Z (2006) Latent friend mining from blog data. Clifton CW, Zhong N, Liu J, Wah BW, Wu X, eds. *Proc. 6th IEEE Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 552–561.

Tan P-N, Steinbach M, Kumar V (2005) *Introduction to Data Mining* (Addison-Wesley, Boston).

Tong H, Faloutsos CC, Pan JY (2006) Fast random walk with restart and its applications. Clifton CW, Zhong N, Liu J, Wah BW, Wu X, eds. *Proc. 6th IEEE Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 613–622.

Wang C, Satuluri V, Parthasarathy S (2007) Local probabilistic models for link prediction. Ramakrishnan N, Zaïane OR, Shi Y, Clifton CW, Wu X, eds. *Proc. 7th IEEE Internat. Conf. Data Mining* (IEEE Computer Society, Washington, DC), 322–331.

Wang D, Pedreschi D, Song C, Giannotti F, Barabasi AL (2011) Human mobility, social ties, and link prediction. *Proc. 17th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining* (ACM, New York), 1100–1108.

Watts A (2001) A dynamic model of network formation. *Games Econom. Behav.* 34(2):331–341.

Weiss G, Zadrozny B, Saar-Tsechansky M (2008) Special issue on utility-based data mining. *Data Mining Knowledge Discovery* 17(2): 129–135.

Yang S-H, Long B, Smola A, Sadagopan N, Zheng Z, Zha H (2011) Like like alike: Joint friendship and interest propagation in social networks. *Proc. 20th ACM Internat. Conf. World Wide Web* (ACM, New York), 537–546.

Zhang B, Thomas A, Doreian P, Krackhardt D, Krishnan R (2013) Contrasting multiple social network autocorrelations for binary outcomes, with applications to technology adoption. *ACM Trans. Management Inform. Systems* 3(4): Article 18.

Zhao K, Yen J, Ngamassi L, Maitland C, Tapia A (2012) Simulating inter-organizational collaboration networks: A multi-relational and event-based approach. *Simulation* 88(5):617–633.

Zheleva E, Getoor L, Golbeck J, Kuter U (2010) Using friendship ties and family circles for link prediction. Giles L, Smith M, Yen J, Zhang H, eds. *Advances in Social Network Mining and Analysis*, Lecture Notes in Computer Science, Vol. 5498 (Springer-Verlag, Berlin), 97–113.

Zheng Z, Pavlou P (2010) Toward a causal interpretation from observational data: A new Bayesian networks method for structural models with latent variables. *Inform. System Res.* 21(2): 365–391.