A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions

ZHEPENG (LIONEL) LI, York University
XIAO FANG, University of Delaware
OLIVIA R. LIU SHENG, University of Utah

Link recommendation has attracted significant attention from both industry practitioners and academic researchers. In industry, link recommendation has become a standard and most important feature in online social networks, prominent examples of which include "People You May Know" on LinkedIn and "You May Know" on Google + . In academia, link recommendation has been and remains a highly active research area. This article surveys state-of-the-art link recommendation methods, which can be broadly categorized into learning-based methods and proximity-based methods. We further identify social and economic theories, such as social interaction theory, that underlie these methods and explain from a theoretical perspective why a link recommendation method works. Finally, we propose to extend link recommendation research in several directions that include utility-based link recommendation, diversity of link recommendation, link recommendation from incomplete data, and experimental study of link recommendation.

CCS Concepts: • **Information systems** → **Social recommendation**;

Additional Key Words and Phrases: Link Recommendation, Social Network, Network Formation

ACM Reference format:
Zhepeng (Lionel) Li, Xiao Fang, and Olivia R. Liu Sheng. 2017. A Survey of Link Recommendation for Social Networks: Methods, Theoretical Foundations, and Future Research Directions. *ACM Trans. Manage. Inf. Syst.* 9, 1, Article 1 (October 2017), 26 pages.
https://doi.org/10.1145/3131782

1 INTRODUCTION

A social network can be represented as $G = \langle V, E \rangle$, where V denotes the set of social entities in the network and E represents the set of existing links each of which connects a pair of social entities. For example, a social network on Facebook consists of a set of users (i.e., social entities) connected by pairwise friendship links among them. Let $\bar{E}$ be the set of potential links that have not been established in a social network; that is, $\bar{E} = V \times V - E$. We use $F(\bar{e})$ to denote the value of

Author's addresses: Z. (Lionel) Li, Schulich School of Business, York University, Toronto, Ontario M3J 1P3, Canada; email: zli@schulich.yorku.ca; X. Fang (corresponding author), Lerner College of Business and Economics, University of Delaware, Newark, Delaware 19716; email: xfang@udel.edu; O. R. L. Sheng, David Eccles School of Business, University of Utah, Salt Lake City, Utah 84112; email: olivia.sheng@eccles.utah.edu.

https://doi.org/10.1145/3131782

a potential link $\bar{e} \in \bar{E}$. The link recommendation problem is to estimate the value of each potential link, rank potential links in decreasing order of value, and recommend top-ranked potential links.

Link recommendation has attracted significant attentions from both industry practitioners and academic researchers over the years. In industry, link recommendation has become a standard and most important feature in online social networks since its early success at LinkedIn (Davenport and Patil 2012). Prominent examples of link recommendation include "People You May Know" on Facebook and LinkedIn as well as "You May Know" on Google + . In academia, link recommendation has been and remains a highly active research area. Given the tremendous academic and practical interests in link recommendation, there is a need for a review of the state-of-the-art knowledge in this area as well as the identification of significant and interesting questions for future research. This survey aims to address this need. Our survey differs from prior surveys of link prediction for social networks (Hasan and Zaki 2011; Lü and Zhou 2011) and contributes to the literature in the following ways.

- While prior surveys focus primarily on either learning-based link recommendation methods (Hasan and Zaki 2011) or proximity-based link recommendation methods (Lü and Zhou 2011), this survey reviews representative methods in both categories and is more comprehensive.
- Prior surveys have not examined social and economic theories underlying link recommendation methods. Our survey identifies these theories (e.g., social interaction theory) and explains from a theoretical perspective why a link recommendation method works.
- Our survey suggests a unique set of research directions worthy of future exploration.

In general, existing link recommendation methods operationalize $F(\bar{e})$ as the likelihood that potential link $\bar{e}$ will be established in the future and recommend the links that are most likely to be established (Kashima and Abe 2006; Hasan et al. 2006; Liben-Nowell and Kleinberg 2007; Yin et al. 2010; Backstrom and Leskovec 2011; Gong et al. 2014). Therefore, the core of these methods is the prediction of linkage likelihood. According to different prediction approaches used, existing link recommendation methods can be broadly categorized into learning-based methods and proximity-based methods, which we review in Sections 2 and 3 respectively. We summarize link recommendation methods in Section 4. We then identify social and economic theories underlying link recommendation methods in Section 5 and suggest important future research directions in Section 6. Section 7 concludes the article.

2 LEARNING-BASED METHODS

Learning-based methods learn a model from training data constructed from observed link establishments and use the learned model to predict the linkage likelihood for each potential link. A model can be learned using classification approaches, probabilistic models, or relational learning approaches. Learning-based methods can thus be categorized as classification-based methods, probabilistic model-based methods, and relational learning-based methods.

2.1 Classification-Based Methods

Given a social network, we can construct training data from observed link establishments in the network. In general, each record of the training data has the format $\langle f_1, f_2, \ldots, f_m, l\rangle$, where $f_1, f_2, \ldots, f_m$ represent features that affect link establishments in the network and l is the class label. We note that each training record is about a pair of social entities, and each feature of the record is defined on the pair. The class label l is 1 for an existing link, and it is 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 intrinsic characteristics of individual social entities. Topological features include neighbor-based features and path-based features. Neighbor-based features characterize neighborhoods of social entities, for example, the number of common neighbors between social entities (Hasan et al. 2006; Lichtenwalter et al. 2010; Benchettara et al. 2010), while path-based features describe paths connecting social entities in a social network, for example, the shortest distance between social entities (Hasan et al. 2006; Lichtenwalter et al. 2010; O'Madadhain et al. 2005; Benchettara et al. 2010; Wang et al. 2007). Nodal features can be computed from intrinsic characteristics of social entities, including demographic characteristics (Zheleva et al. 2008), geographic characteristics (O'Madadhain et al. 2005; Scellato et al. 2011; Wang et al. 2011), and semantic characteristics (Hasan et al. 2006; Wang et al. 2007).

Typical classification approaches can be applied to the constructed training data to predict the linkage likelihood of a potential link. O'Madadhain et al. (2005) employ logistic regression to predict the likelihood of interactions between social entities using data regarding CiteSeer articles, AT&T telephone calls, and Enron emails. Wang et al. (2007) combine a local Markov random field model and logistic regression to predict the likelihood of co-authorship using DBLP and PubMed article datasets. Benchettara et al. (2010) also predict the likelihood of co-authorship but employ a decision tree classifier enhanced with Adaboost for making the prediction. Gong et al. (2014) predict new and missing links in Google + using support vector machine (SVM). Comparing popular classification approaches for predicting linkage likelihood, including Naïve Bayes, decision tree, SVM, and k-nearest neighbor, Hasan et al. (2006) find that SVM seems to be the most effective among these approaches and that SVM with a radial basis function (RBF) kernel outperforms SVM with a linear kernel.

The performance of classification-based methods can be enhanced in a number of ways. One way to enhance the effectiveness and efficiency of these methods is through judicious selection of relevant features. In this vein, Xu and Rockmore (2012) and Bliss et al. (2014) design feature selection frameworks for ranking and weighting features. In addition, Lichtenwalter et al. (2010) suggest that the performance of classification-based methods can be improved by carefully sampling training data. According to Lichtenwalter et al. (2010), to predict the likelihood of linkage between social entities that are n hops away, it is more effective and efficient to use a training sample of entities that are n hops apart than to use the entire training data. Yet another way to enhance performance is to enrich a social network with additional information that is useful for linkage likelihood prediction. A social network can be enriched by incorporating additional relationships among its social entities. For example, Zheleva et al. (2008) enrich a friendship social network of pets with their family affiliations. A social network can also be enriched by adding additional nodes. For instance, Gong et al. (2014) propose a social-attribute network whose nodes consist of both social entities and attributes of these entities. Finally, traditional ways of improving the performance of classification approaches are beneficial, too. For example, ensemble methods, such as AdaBoost and Bagging, have been employed to improve the performance of linkage likelihood prediction (Benchettara et al. 2010; Hasan et al. 2006; Lichtenwalter et al. 2010), and Doppa et al. (2010) develop a cost-sensitive method to address the class imbalance issue in linkage likelihood prediction.

2.2 Probabilistic Model-Based Methods

In general, probabilistic model-based methods (Kashima and Abe 2006; Clauset et al. 2008; Lu et al. 2010; Backstrom and Leskovec 2011; Hopcroft et al. 2011; Yang et al. 2011; Kim and Leskovec 2011; Barbieri et al. 2014; Dong et al. 2015; Song et al. 2015) predict the linkage likelihood L_{ij} between social entities v_i and v_j as

$$L_{ij} = g\left(\boldsymbol{f}_{ij}, \boldsymbol{\theta}^*\right), \tag{1}$$

where $g(\cdot)$ is a prediction function, $\boldsymbol{f}_{ij}$ is a vector of features gathered from social entities v_i and v_j, and $\boldsymbol{\theta}^*$ is a vector of parameters that can be learned from observed link establishments. Given training data D constructed from observed link establishments, $\boldsymbol{\theta}^*$ is estimated as the parameter vector that best explains D, that is,

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{argmax}\, Obj\,(D|M, \boldsymbol{\theta}), \tag{2}$$

where the objective function $Obj(D|M, \boldsymbol{\theta})$ defines how well observed link establishments in D can be explained by a probabilistic model M with a parameter vector $\boldsymbol{\theta}$. We next discuss representative probabilistic model-based methods.

The method proposed by Kashima and Abe (2006) is based on the intuition that a social entity's linkage decision is influenced by his or her neighbors in a social network. Accordingly, they predict the linkage likelihood ϕ_{ij}^{t+1} between social entities v_i and v_j at time $t + 1$ as

$$\phi_{ij}^{t+1} = \frac{1}{|V| - 1}\left(\sum_{k \neq i,j} \theta_{kj}\phi_{ki}^t + \theta_{ki}\phi_{kj}^t\right) + \left(1 - \frac{1}{|V| - 1}\sum_{k \neq i,j} \theta_{kj} + \theta_{ki}\right)\phi_{ij}^t, \tag{3}$$

where V is the set of social entities in a social network, entity k is a neighbor of entity i or j, parameter vector $\boldsymbol{\theta} =< \theta_{xy}|\forall x, y \in V\rangle$. The stationary likelihood of linkage between a pair of social entities is obtained by iteratively updating their linkage likelihood according to Equation (3) until convergence. The objective function $Obj(\cdot)$ is defined as the total stationary linkage likelihood summed over all existing links, and $\boldsymbol{\theta}^*$ is then estimated as the parameter vector that maximizes the objective function.

Guimerà and Marta (2009) propose a stochastic block model in which social entities are partitioned into groups. Given any two groups η and η', there is a connection between them if there exists a link between entities $v_a \in \eta$ and $v_b \in \eta'$. Let $P_{\eta\eta'}$ be the probability of a connection existing between groups η and η'. Let A denote the adjacency matrix of a social network, where $A_{ab} = 1$ if there is a link between entities v_a and v_b in the network and $A_{ab} = 0$ otherwise. Given a partition model M, the likelihood of A is given by

$$L(A|M) = \prod_{\eta \leq \eta'} P_{\eta\eta'}^{l_{\eta\eta'}}\left(1 - P_{\eta\eta'}\right)^{r_{\eta\eta'} - l_{\eta\eta'}}, \tag{4}$$

where $l_{\eta\eta'}$ is the number of existing connections between groups η and η' and $r_{\eta\eta'}$ is the number of possible connections between groups η and η'. Using Equation (4) as the objective function, we obtain the maximum likelihood estimation of $P_{\eta\eta'}$ as

$$P_{\eta\eta'}^* = \frac{l_{\eta\eta'}}{r_{\eta\eta'}}. \tag{5}$$

The linkage likelihood L_{ij} between social entities v_i and v_j is then given by

$$L_{ij} = \int_{\mathbf{M}} P(A_{ij} = 1|M)L\,(\mathrm{M}|A)\,dM, \tag{6}$$

where $\boldsymbol{M}$ is the space of all possible partition models. By applying Bayes's Theorem, we can rewrite Equation (6) as

$$L_{ij} = \frac{\int_{\mathbf{M}} P(A_{ij} = 1|M)L(A|M)P\,(M)\,dM}{\int_{\mathbf{M}} L\,(A|M)\,P\,(M)\,dM}, \tag{7}$$

where $P(M)$ is the prior probability of partition model M and $L(A|M)$ can be obtained using Equation (4) with $P_{\eta\eta'}^*$. In practice, it is not possible to enumerate all possible partition models and the Metropolis algorithm is applied to sample partition models (Metropolis et al. 1953). Clauset et al. (2008) propose a similar method, in which social entities are grouped into a hierarchical structure.

Huang (2010) argues that the linkage likelihood between social entities v_i and v_j depends on the number of existing paths connecting them. Let c_z be the probability of linking v_i and v_j given that there exists a length-z path connecting them, $z \geq 2$. According to Huang (2010), the linkage likelihood L_{ij} between social entities v_i and v_j can be predicted as

$$L_{ij} = 1 - (1 - c_2)^{m_2}(1 - c_3)^{m_3} \cdots (1 - c_k)^{m_k}, \tag{8}$$

where m_z denotes the number of existing length-z paths connecting v_i and v_j, $z = 2, 3, \ldots, k$. To compute L_{ij}, we need to estimate parameters in $\boldsymbol{\theta} = \langle c_2,\ c_3, \ldots, c_k \rangle$. Huang (2010) proposes to estimate $\boldsymbol{\theta}$ based on a generalized variant of the clustering coefficient, which indicates the tendency of a path to form a cycle (Watts and Strogatz 1998; Newman et al. 2001). Considering a social network as a graph, Huang (2010) computes its clustering coefficient CC_k as

$$CC_k = \frac{\left|cycle^{\langle k \rangle}\right|}{\left|path^{\langle k \rangle}\right|}, \tag{9}$$

where $path^{\langle k \rangle}$ denotes the set of length-k paths and $cycle^{\langle k \rangle}$ represents the set of cycles each of which is formed by adding a link to a length-k path in $path^{\langle k \rangle}$. By adding a link connecting v_i and v_j with probability L_{ij}, the expected clustering coefficient of the social network is a function of $\boldsymbol{\theta}$, that is, $f(c_2, c_3, \ldots, c_k)$. According to Huang (2010), we have $f(c_2, c_3, \ldots, c_k) \approx CC_k$. Therefore, each parameter in $\boldsymbol{\theta}$ can be estimated as

$$c_z^* = \underset{c_z}{argmin}\left(\left|f\left(c_2, c_3, \ldots, c_i, \ldots, c_k\right) - CC_k\right|\right), \tag{10}$$

where $z = 2, 3, \ldots, k$.

Hopcroft et al. (2011) develop a method based on social network theories such as homophily theory. According to Hopcroft et al. (2011), the linkage likelihood L_{ij} between social entities v_i and v_j is predicted as

$$L_{ij} = P\left(y_{ij} = 1 | \boldsymbol{f}_{ij}, \boldsymbol{\theta}^*\right), \tag{11}$$

where $y_{ij} = 1$ indicates the linkage between social entities v_i and v_j, feature vector $\boldsymbol{f}_{ij}$ is constructed based on social network theories, and $\boldsymbol{\theta}^*$ is the parameter vector to be learned. Specifically, $\boldsymbol{\theta}^*$ is estimated as the parameter vector that best explains observed link establishments using a gradient descent method (Hopcroft et al. 2011).

Backstrom and Leskovec (2011) propose a supervised random walk method to predict linkage likelihood. They define a transition matrix with the social entities in a social network as indices. Each element of the transition matrix is a transition probability from one entity to another and it is defined as an exponential function

$$a_{ij} = \exp\left(\boldsymbol{f}_{ij} \cdot \boldsymbol{\theta}^*\right) \tag{12}$$

or a logistic function

$$a_{ij} = \frac{1}{1 + \exp(-\boldsymbol{f}_{ij} \cdot \boldsymbol{\theta}^*)}, \tag{13}$$

where $\boldsymbol{f}_{ij}$ is a feature vector. To compute the linkage likelihood L_{ij} between entities v_i and v_j, a random walk with restart is started from v_i according to the transition matrix and L_{ij} is the stationary probability of reaching v_j from v_i. Backstrom and Leskovec (2011) estimate $\boldsymbol{\theta}^*$ as the parameter vector that minimizes a loss function defined based on the difference between the stationary probability from v_i to one of its currently linked social entities and the stationary probability from v_i to

PAPER Class

PaperID	Area	Keyword	Journal
P1	AI	Representation	J1
P2	Data Mining	Clustering	J2
P3	Data Mining	Association Rule	J1
P4	AI	Reasoning	J1

Citation

Citing	Cited
P1	P4
P2	P3
P3	P1

Fig. 1. An example of Relational Data Model (M_D).

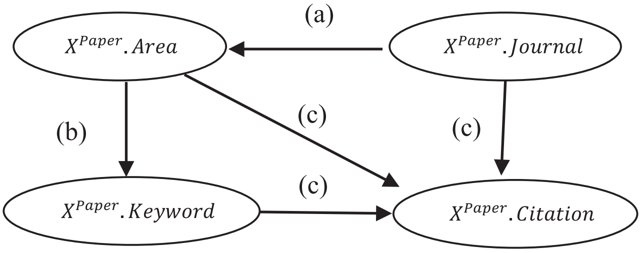

Fig. 2. An example of Dependency Structure Model (M_S).

one of the social entities that it is currently not linked to. Yang et al. (2011) develop a similar linkage prediction method that combines a random walk model with collaborative filtering techniques.

2.3 Relational Learning-Based Methods

Relational learning represents data using a relational model; it then learns and infers based on the model (Friedman et al. 1999; Getoor et al. 2001; Taskar et al. 2003; Heckerman et al. 2004). Specifically, a relational model consists of two components: the relational data model (M_D) and the dependency structure model (M_S). A relational data model consists of a set of classes $X = \{X^c\}$, each of which is described by a set of descriptive attributes $A(X^c)$ (Getoor and Taskar 2007). For example, as shown in Figure 1, PAPER class X^{paper} has descriptive attributes such as topic area, keyword, and journal, that is, $A(X^{paper}) = \{Area,\ Keyword,\ Journal\}$. Each class X^c is associated with a set of linkage relationships $R(X^c)$. In Figure 1, a linkage relationship $x^{paper}.\rho \in R(X^{paper})$ of a paper $x^{paper} \in X^{paper}$ represents its citation of other papers.

A dependency structure model represents the dependencies among the descriptive attributes and linkage relationships of a class. Using the PAPER class in Figure 1 as an example, a dependency structure model shown in Figure 2 captures the following dependencies: (a) the area of a paper depends on the journal in which it is published, (b) the keyword of a paper depends on its area, and (c) linkage relationships (e.g., citations) among papers depends on their areas, keywords, and published journals. Dependencies can be learned from data or constructed using background knowledge (Getoor and Taskar 2007).

Given a dependency structure model, the distribution of a descriptive attribute $x^c.A$ and the distribution of a linkage relationship $x^c.\rho$ can be derived as $P(x^c.A|Pa(x^c.A))$ and $P(x^c.\rho|pa(x^c.\rho)$, respectively, where $Pa(\cdot)$ denotes the parents of a descriptive attribute or a linkage relationship in a dependency structure model. The objective of relational learning is to obtain a vector of parameters θ such that the following data likelihood function for an instance I of a relational model is maximized:

$$l\left(\boldsymbol{\theta}|M_D, M_S, \boldsymbol{I}\right) = \prod_{X^c \in \boldsymbol{X}} \prod_{x^c \in X^c} \prod_{A \in \boldsymbol{A}(X^c)} P(x^c.A|Pa\left(x^c.A\right)) \prod_{\rho \in \boldsymbol{R}(X^c)} P(x^c.\rho|Pa\left(x^c.\rho\right)). \tag{14}$$

The obtained optimal parameters $\boldsymbol{\theta}^*$ can then be used to infer the probability of a linkage relationship in relational learning (e.g., citation) and the distribution of a missing descriptive attribute (Getoor et al. 2003).

To apply relational learning to the link recommendation problem, we can treat linkage likelihood in link recommendation as the probability of a linkage relationship in relational learning (Getoor et al. 2003). In this vein, Bilgic et al. (2007) show that inferring the distributions of missing descriptive attributes can help the estimation of linkage likelihood. In addition to Bayesian Network-based relational learning depicted in Equation (14), there are other relational learning methods that can be applied to link recommendation. For example, Markov Network-based relational learning has been used for link recommendation (Taskar et al. 2003; Mihalkova et al. 2011). Moreover, Heckerman et al. (2001) introduce a relational dependency network for relational learning, Popescul and Ungar (2003) develop a relational learning method based on the structural logistic regression model, Heckerman et al. (2004) design a relational learning model based on the Entity-Relationship diagram, and Yu et al. (2006) propose a non-parametric stochastic approach to relational learning.

3 PROXIMITY-BASED METHODS

Proximity-based methods surrogate the linkage likelihood between social entities using the proximity between them (Liben-Nowell and Kleinberg 2007; Chen et al. 2009; Crandall et al. 2010; Lü and Zhou 2011). According to homophily theory (McPherson et al. 2001), similar social entities are of high tendency to link to each other. In light of this theory, the greater the proximity between social entities the higher the linkage likelihood between them. Proximity-based methods can generally be grouped into nodal proximity-based methods and structural proximity-based methods.

3.1 Nodal Proximity-Based Methods

Nodal proximity-based methods surrogate the linkage likelihood L_{ij} between social entities $\mathrm{v_i}$ and v_j using their nodal proximity $S(Y_i, Y_j)$, where Y_i and Y_j denote the profile of social entities v_i and v_j, respectively, and $S(\cdot)$ is a similarity function. The profile of a social entity consists of its intrinsic characteristics, including demographic, geographic, and semantic characteristics. Demographic characteristics, such as age, education, and occupation, are commonly used (Zheleva et al. 2008; Xu and Rockmore 2012). Geographic characteristics, such as co-location and distance, have also been employed to capture the closeness between two social entities in a physical space (Quercia and Capra 2009; Crandall et al. 2010; Yin et al. 2010; Scellato et al. 2011; Wang et al. 2011). Semantic characteristics, such as keywords, annotation tags, and communication descriptions, are used to measure the similarity between social entities in terms of their semantic patterns (Shen et al. 2006; Chen et al. 2009; Schifanella et al. 2010; Yin et al. 2010; Makrehchi 2011; Adali et al. 2012; Kuo et al. 2013; Yuan et al. 2014).

A number of similarity functions have been applied to measure nodal proximity. For profiles with numerical characteristics, cosine similarity has been employed (Chen et al. 2009; Shen et al. 2006; Schifanella et al. 2010; Wang et al. 2011). According to Salton (1989), the cosine similarity between social entities v_i and v_j is given by

$$CS(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \frac{\sum_k Y_{ik} Y_{jk}}{\sqrt{\sum_k Y_{ik}^2 \sum_k Y_{jk}^2}}, \tag{15}$$

where Y_{ik} is the k^{th} characteristic in $\mathbf{Y}_i$. Another similarity function suitable for numerical characteristics is *KL*-divergence (Shen et al. 2006). Concretely, the *KL*-divergence between social entities v_i and v_j is computed as

$$KL(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_k h_{\text{ik}} \log \frac{h_{ik}}{h_{jk}} + h_{jk} \log \frac{h_{jk}}{h_{ik}}, \tag{16}$$

where h_{ik} is the probability of the k^{th} characteristic in $\mathbf{Y}_i$. Manhattan distance has been used to gauge the *dissimilarity* between social entities (Wang et al. 2011; Adali et al. 2012). Specifically, the Manhattan distance between social entities v_i and v_j is defined as

$$MD(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_k |Y_{ik} - Y_{jk}|. \tag{17}$$

For profiles with nominal characteristics, Jaccard's coefficient is a suitable similarity function (Schifanella et al. 2010; Scellato et al. 2011; Wang et al. 2011; Xu and Rockmore 2012; Kuo et al. 2013). In particular, the Jaccard's coefficient between social entities v_i and v_j is defined as

$$JC(\mathbf{Y}_i, \mathbf{Y}_j) = \frac{|\mathbf{Y}_i \cap \mathbf{Y}_j|}{|\mathbf{Y}_i \cup \mathbf{Y}_j|}. \tag{18}$$

In addition to the above-reviewed similarity functions, other similarity functions such as match count have also been used (Kahanda and Neville 2009; Xiang et al. 2010).

3.2 Structural Proximity-Based Methods

Structural proximity measures the proximity between two social entities based on their structural features in a social network (Jeh and Widom 2002; Liben-Nowell and Kleinberg 2007; Lü et al. 2009; Lü and Zhou 2011; Liu and Lü 2010). Structural proximity-based methods surrogate the linkage likelihood L_{ij} between social entities v_i and v_j with their structural proximity (Liben-Nowell and Kleinberg 2007), which can be classified into neighborhood-based structural proximity and path-based structural proximity.

3.2.1 Neighborhood-Based Structural Proximity. Structural proximity between social entities can be measured based on their neighborhoods. Common neighbor is a widely used neighborhood-based structural proximity measure (Newman 2001; Liben-Nowell and Kleinberg 2007). The common neighbor CN_{ij} between social entities v_i and v_j is computed as the number of their mutual neighbors, that is,

$$CN_{ij} = |\Gamma_i \cap \Gamma_j|, \tag{19}$$

where Γ_i and Γ_j denote the set of direct neighbors of entities v_i and v_j, respectively, and $|\cdot|$ is the cardinality of a set. Extended from the common neighbor measure, the Adamic/Adar measure assigns less weight to more connected common neighbors (Adamic and Adar 2003; Liben-Nowell and Kleinberg 2007). Specifically, the Adamic/Adar AA_{ij} between social entities v_i and v_j is given by

$$AA_{ij} = \sum_{v_z \in \Gamma_i \cap \Gamma_j} \frac{1}{\log |\Gamma_z|}, \tag{20}$$

where v_z is a common neighbor of v_i and v_j and Γ_z denotes the set of direct neighbors of v_z. It has been shown theoretically and empirically that the linkage likelihood between social entities is highly correlated with their neighborhood sizes (Barabási and Albert 1999; Newman 2001;

Barabási et al. 2002; Liben-Nowell and Kleinberg 2007). Motivated by these theoretical and empirical findings, the preferential attachment PA_{ij} between social entities v_i and v_j is defined as

$$PA_{ij} = |\Gamma_i| \times \left|\Gamma_j\right|. \tag{21}$$

Observing that two social entities are similar if their neighbors are similar, Jeh and Widom (2002) propose the SimRank measure. The SimRank score SR_{ij}between social entities v_i and v_j is defined as (Jeh and Widom 2002; Liben-Nowell and Kleinberg 2007; Liu and Lü 2010)

$$SR_{ij} = \frac{\gamma \cdot \sum_{v_z \in \Gamma_i} \sum_{v_{z'} \in \Gamma_j} SR_{zz'}}{\left|\Gamma_i \right| \cdot \left| \Gamma_j \right|}, \tag{22}$$

where $\gamma \in (0, 1)$ is a decay factor. A variation of the SimRank measure is proposed by Leicht et al. (2006). In addition to the measures reviewed in this subsection, there are other neighborhood-based measures, such as the Sørensen Index (Sørensen 1948), the Salton Measure (Salton and McGill 1986), and the Hub Promoted (HP)/Hub Depressed (HD) Index (Ravasz et al. 2002).

3.2.2 Path-based Structural Proximity. Going beyond neighborhoods, path-based structural proximity measures target paths connecting social entities. The Katz index measures the structural proximity between social entities using the number of paths connecting them, weighted by their lengths (Katz 1953). Originally developed for measuring the social status of a social entity, the Katz index has been shown to be effective in predicting linkage between social entities (Liben-Nowell and Kleinberg 2007). The Katz index KZ_{ij} between social entities v_i and v_j is given by Katz (1953),

$$KZ_{ij} = \sum_k \beta^k \left|path_{ij}^{\langle k \rangle}\right|, \tag{23}$$

where $path_{ij}^{\langle k \rangle}$ represents the set of length-k paths connecting v_i and v_j and weight β is between 0 and 1. According to Equation (23), the contribution of a path to the Katz index decreases as its length increases. The local path index (Zhou et al. 2009) is a localized version of the Katz index, where k in Equation (23) is upper bounded.

Considering link establishment between social entities as a random walk from one to the other, the PageRank algorithm (Brin and Page 1998) can be adapted to compute structural proximity between them (Haveliwala 2002; Jeh and Widom 2003; Tong et al. 2006). Let $\boldsymbol{P}$ be a transition matrix with element P_{xy} representing the transition probability from social entity v_x to entity v_y, where $P_{xy} = \frac{1}{|\Gamma_x|}$ if $e_{xy} \in E$ and $P_{xy} = 0$ otherwise. Let $\boldsymbol{q_i}$ be a vector of probabilities, each element of which represents the probability from social entity v_i to another entity in a social network. For example, element q_i^j represents the probability from v_i to v_j. According to Tong et al. (2006), we have

$$\boldsymbol{q}_i = \alpha \boldsymbol{P}^T \boldsymbol{q}_i + (1 - \alpha)\, \epsilon_{\boldsymbol{i}}, \tag{24}$$

where the i^{th} element in vector $\epsilon_{\boldsymbol{i}}$ is 1 and all other elements are 0. We can iteratively compute $\boldsymbol{q}_i$ according to Equation (24) until convergence. The PageRank-based structural proximity PR_{ij} between v_i and v_j is calculated as

$$PR_{ij} = q_i^j + q_j^i. \tag{25}$$

Based on a similar idea of treating link establishment as a random walk, Fouss et al. (2007) define the hitting time H_{ij} as the expected number of steps needed to reach social entity v_j from entity v_i for the first time. Let b_{xy} denote the weight between entities v_x and v_y, where $b_{xy} > 0$ if v_x and

v_y are linked and $b_{xy} = 0$ otherwise. To reach v_j from v_i, one needs to go to a neighbor v_k of v_i and then proceeds from v_k to v_j. Accordingly, the hitting time H_{ij} can be measured as

$$H_{ij} = 1 + \sum_k P_{ik} H_{kj}, \tag{26}$$

where P_{ik} is the probability of moving from v_i to its neighbor v_k and $P_{ik} = b_{ik} / \sum_z b_{iz}$. The hitting time H_{ij} can be computed recursively according to Equation (26) using a Markovian algorithm (Kemeny and Snell 1976) or can be obtained using the pseudoinverse of the Laplacian matrix of a social network (Fouss et al. 2007). The average commute time $\mathrm{ACT_{ij}}$ is a symmetric variation of H_{ij}. Specifically, the average commute time ACT_{ij} is defined as the expected number of steps needed to reach v_j from v_i for the first time and then to go back to v_i from v_j (Fouss et al. 2007). Thus, ACT_{ij} is measured as (Fouss et al. 2007)

$$ACT_{ij} = H_{ij} + H_{ji}. \tag{27}$$

Liben-Nowell and Kleinberg (2007) define the normalized average commute time ACT'_{ij} as

$$ACT'_{ij} = H_{ij} \cdot \pi_j + H_{ji} \cdot \pi_i, \tag{28}$$

where π_j and π_i represent the stationary probability of reaching v_j and v_i, respectively. Since H_{ij}, ACT_{ij}, and ACT'_{ij} are all distance-based measures, a higher value of these measures indicates less linkage likelihood (Liben-Nowell and Kleinberg 2007).

3.2.3 Performance Improvement. Several approaches have been proposed to improve the performance of structural proximity-based methods. A social network can be represented as an adjacency matrix, each element of which represents whether two social entities are linked or the strength of their linkage. Inspired by the success of matrix factorization techniques in information retrieval and recommender systems (Deerwester et al. 1990; Koren et al. 2009; Rennie and Srebro 2005), Liben-Nowell and Kleinberg (2007) apply singular value decomposition, a matrix factorization technique, to reduce the rank of an adjacency matrix that represents a social network. They show that structural proximities computed using the rank-reduced adjacency matrix can predict linkage more accurately than those calculated from the original adjacency matrix. Other approaches to improving the prediction accuracy of structural proximity-based methods include the supervised matrix factorization (Menon and Elkan 2011), aggregation approach (Liben-Nowell and Kleinberg 2007), and the kernel approach (Kunegis and Lommatzsch 2009).

Given the huge size of real-world social networks, it is imperative to develop efficient approaches to computing structural proximities. One approach (Song et al. 2009; Shin et al. 2012) is developed based on approximation techniques. It trades accuracy for efficiency and produces approximate structural proximities. Another approach calculates structural proximities from a local social network rather than a complete social network and thus improves the efficiency of computing structural proximities (Zhou et al. 2009; Lü et al. 2009; Lichtenwalter et al. 2010; Liu and Lü 2010; Fire et al. 2011). Yet another improvement adds a time dimension to a social network and computes structural proximities from an evolving social network. For example, Huang and Lin (2009) combine time series models and structural proximity measures to predict linkage likelihood at a particular time. Acar et al. (2009) and Dunlavy et al. (2011) create a tensor from a series of adjacency matrices, each of which represents a social network at a particular time; they then compute the Katz index from the tensor.

4 SUMMARY OF LINK RECOMMENDATION METHODS

In summary, link recommendation methods predict the linkage likelihood for each potential link and recommend potential links with the highest linkage likelihoods. In particular, these methods

Table I. Categorization of Link Recommendation Methods

Method		Representative Work
Learning-based Method	Classification-based Method	– O'Madadhain et al. (2005) – Hasan et al. (2006) – Wang et al. (2007) – Lichtenwalter et al. (2010) – Scellato et al. (2011) – Gong et al. (2014) – Bliss et al. (2014) – Zhang et al. (2014)
	Probabilistic Model-based Method	– Kashima and Abe (2006) – Guimerà and Marta (2009) – Huang (2010) – Backstrom and Leskovec (2011) – Hopcroft et al. (2011) – Yang et al. (2011) – Barbieri et al. (2014) – Dong et al. (2015)
	Relational Learning-based Method	– Getoor et al. (2003) – Popescul and Ungar (2003) – Taskar et al. (2003) – Yu et al. 2006 – Mihalkova et al. (2011)
Proximity-based Method	Nodal Proximity-based Method	– Chen et al. (2009) – Crandall et al. (2010) – Schifanella et al. (2010)
	Structural Proximity-based Method (Neighborhood)	– Newman (2001) – Barabási et al. (2002) – Jeh and Widom (2002) – Adamic and Adar (2003) – Liben-Nowell and Kleinberg (2007)
	Structural Proximity-based Method (Path)	– Tong et al. (2006) – Liben-Nowell and Kleinberg (2007)

predict linkage likelihood using machine-learning approaches, including classification, probabilistic models, and relational learning, or they surrogate linkage likelihood with proximity measures, such as nodal and structural proximity measures. We therefore categorize link recommendation methods accordingly and list representative works in each category in Table 1.

In comparison to proximity-based methods, learning-based methods have the following advantages. First, the output of a learning-based method is (or can be easily transformed to) the probability that a potential link will be established in the future. For example, a Naïve Bayes-based method outputs the probability of linkage for each potential link. Most proximity-based methods, on the other hand, do not produce probabilities of linkage. Taking nodal proximity-based methods as an example, these methods yield similarity scores between social entities rather than probabilities of linkage. While the outputs of proximity-based methods are adequate for ranking potential links, they are insufficient for more advanced applications that require probabilities of linkage. For example, in many applications, the cost of recommending a wrong potential link that will not be established is different from the cost of missing a true potential link that will be established. In these applications, knowing probabilities of linkage is essential to cost-sensitive link recommendation decisions. Second, learning-based methods learn to predict future linkage from the ground truth of prior link establishments. Therefore, well-designed learning-based methods can predict future linkage more accurately than proximity-based methods (Lichtenwalter et al. 2010; Backstrom and Leskovec 2011). In addition, their prediction performance is more stable across different social networks than that of proximity-based methods.

However, learning-based methods, especially classification-based methods, suffer from the imbalance issue "due to the inherent disproportion of links that can form to links that do form" (Lichtenwalter et al. 2010). One solution addresses this issue by focusing on potential links that would connect users two hops away rather than considering all possible potential links (Backstrom and Leskovec 2011; Wang et al. 2011; Dong et al. 2012). Another solution divides a link recommendation problem into sub-problems, each of which predicts the likelihood of linkage between social entities that are *n* hops away using training data of entities that are *n* hops apart (Lichtenwalter et al. 2010). This solution further resample training data of each sub-problem to address the imbalance issue. Proximity-based methods have their own merits. First, proximity-based methods do not need a training phase, which is required for learning-based methods. As a consequence, compared to learning-based methods, proximity-based methods save the cost of constructing training data and the cost of learning from training data. Second, many proximity-based methods are easy to implement and widely applied in practice. For example, common neighbor is popularly used by major online social networks for link recommendation. It is called "mutual friend" in Facebook and "shared connection" in LinkedIn.

Within the category of proximity-based methods, choosing a proper method for a link recommendation task on hand needs to consider the following two factors: (1) network evolution and data availability and (2) user linkage behavior. A social network normally begins with few nodes connected by a sparse linkage structure; it evolves as new nodes are added and new links are formed among its nodes. At the early stage of a social network, there is little structural information and thus the performance of structural proximity-based methods is significantly impaired, especially for those that rely heavily on rich structural information (e.g., path-based structural proximities). Therefore, it is more effective to use nodal proximity-based methods at this stage. As more links are formed over time, the performance of structural proximity-based methods picks up. For this situation, it makes more sense to use structural proximity-based methods, especially when there lacks nodal features essential for nodal proximity-based methods (partly due to privacy concerns). Another factor to consider when choosing a proximity-based method is user linkage behavior. Users demonstrate various link formation behaviors (Lichtenwalter et al. 2010): some primarily looking for local community building while others mainly seeking complementary skills from distant individuals. In the case of local community building, neighborhood-based structural proximities are effective; when users seek to build links over long distances, path-based structural proximities are proper.

To choose an appropriate learning-based method, one should consider the following aspects. First, due to the large scale of social networks, learning-based methods usually incur high training cost. Thus, when choosing a learning-based method, one needs to consider the tradeoff between the cost of training and the accuracy of link recommendation. For example, while a naïve Bayesian-based method could incur less training cost than a Bayesian network-based method, the latter could be more accurate in recommending links than the former. Second, one needs to consider the tradeoff between generalizability and recommendation performance of a model produced by a learning-based method. For example, probabilistic model-based methods are often tailored to handle specific characteristics of a focal network. Models produced by these methods could perform well in the focal network but may not generalize to other social networking scenarios. Classification-based methods, on the other hand, enjoy more generalizability than probabilistic model-based methods.

Recent developments in link recommendation include link recommendation in signed networks and reciprocal networks. Links in a social network could have positive or negative sign, such as trust vs. distrust (Guha et al. 2004), friend vs. foe (Kunegis et al. 2009), and favor vs. against (Leskovec et al. 2010a). Link recommendation in signed social networks aims at predicting not only the likelihood of linkage but also the sign of linkage. Some studies develop learning-based methods for link recommendation in signed networks. In this vein, Leskovec et al. (2010a) empirically discover that signed linkage formation in a social network can be explained by social balance and social focus theories. In a follow-up study, they define and solve the problem of edge sign prediction, which aims to estimate the sentiment (e.g., favor or against) between two individuals in a social network (Leskovec et al. 2010b). Informed by social balance and social status theories, they define path-based and neighborhood-based structural features that affect signed linkage formation. They further propose a logistic regression method that takes these features as inputs to predict edge signs. Yang et al (2012) design a method to infer hidden factors that characterize individuals from their behaviors (e.g., their evaluation of items); the method further predicts the existence of linkage and the sign of linkage between individuals based on the correlation of their hidden factors. Ye et al. (2013) target a specific link recommendation problem in signed social networks, where there are few signed links in the network. They approach the problem from the perspective of transfer learning (Pan et al. 2010) and propose a method to infer the sign of linkage for a social network with few signed links (i.e., target network) from a social network with large number of signed links (i.e., source network). Tang et al. (2015) develop a method to infer linkage signs from users' posts and other users' opinions to these posts in a social media site. In a recent study, Song et al. (2015) advocate the use of Generalized AUC in the evaluation of link recommendation performance in signed social networks. Different from AUC, Generalized AUC accounts for three classes (e.g., positive link, negative link, or unlinked) and it is particular useful for evaluating link recommendation performance in signed social networks.

In addition to learning-based approaches, methods based on matrix operations have been developed to predict linkage signs. These methods target the structure of a signed social network. For example, Guha et al. (2004) define four atomic propagations that reflect how signed relationships are formulated in a social network and apply these atomic propagations to predict linkage signs in a social network. Kunegis et al. (2009) define an adjacency matrix for a signed network, with 1 for positive link, -1 for negative link, and 0 representing no linkage. Matrix operations such as singular value decomposition are then applied to the adjacency matrix to infer linkage signs. Moreover, Symeonidis et al. (2010) propose a neighborhood-based proximity measure that can be extended to link recommendation in signed social networks. Papadimitriou et al. (2012) define a path-based proximity measure that approximates the linkage likelihood between a pair of social entities with the weighted count of paths between them. The number of paths between a pair of social entities is obtained by iteratively multiplying the adjacency matrix for a signed social network.

Many social networks, such as Twitter and online dating platforms, are comprised of two-way relationships. For example, a Twitter user responds to an initial linkage (i.e., following) by establishing a reciprocal link (i.e., following back). Singhal et al. (2013) and Jiang et al. (2015) define the reciprocation (reciprocity) of a network as the percentage of its links with a reciprocal link. Singhal et al. (2013) find that different types of networks could have different reciprocations; for example, the reciprocation of a chat network is normally higher than that of a trust network. Jiang et al. (2015) give the upper bound of the reciprocity of a network. Cheng et al. (2011) formulate two major reciprocity prediction problems. One problem aims at predicting the likelihood that a link e_{ij}

from v_i to v_j will be established given that a link e_{ji} from v_j to v_i already exists. The other problem is to predict the coexistence of e_{ij} and e_{ji}. Hopcroft et al. (2011) and Lou et al. (2013) target the first problem. They propose a triad factor graph model to predict reciprocal relationships on Twitter, based on the linkage structure of Twitter and factors that affect the establishment of reciprocal relationships such as geographical distance and profile similarity between Twitter users. Zhao et al. (2014) study the second problem. They propose a method to predict reciprocal relationships in an online dating platform using taste similarity, attractiveness similarity, and unattractiveness similarity between users of the network. A dating network is an instance of bipartite networks, each of which consists of two sets of social entities and links connecting an entity in one set to an entity in the other. In general, link prediction in bipartite networks can be enhanced by considering similarities between entities belonging to a same set using techniques such as collaborative filtering (Akehurst et al. 2011; Cai et al. 2012; Zhao et al. 2014) and latent semantic analysis (Hoffman 2004). Beyond direct reciprocity, indirect reciprocity occurs when there is a path from v_i to v_j given that a link from v_j to v_i already exists. Although indirect reciprocal behaviors have been observed in altruistic and cooperative interactions among human beings (Nowak and Sigmund 2005), little is known about how to predict indirect reciprocity. Therefore, indirect reciprocity prediction could be an interesting future research area.

5 THEORETICAL FOUNDATIONS FOR LINK RECOMMENDATION METHODS

Our survey of link recommendation methods suggests that many of these methods are developed based on observations of social phenomena but overlook social and economic theories underlying these phenomena. The link between link recommendation methods and their foundational theories is largely missing in existing link recommendation studies. In general, a theory explains why a mechanism works and under what circumstances it works (Parsons 1938; Bacharach 1989). In particular, identifying social and economic theories underlying link recommendation methods has the following benefits. First, it helps us understand why and under what situations a link recommendation method works. With this understanding, we can effectively select appropriate methods for link recommendation tasks at hand. Second, it helps us identify limitations of existing methods and design more advanced methods. Third, theories inform us about generic factors that actually affect linkage decisions, which we can use to design novel and effective link recommendation methods. In this section, we identify social and economic theories underlying link recommendation methods and uncover the missing link between them.

Observations that have inspired existing link recommendation methods can be broadly grouped into three categories. Specifically, it is observed that the decision of linkage between two social entities is affected by (1) the degree of similarity between them (O'Madadhain et al. 2005; Hasan et al. 2006; Wang et al. 2007; Chen et al. 2009; Crandall et al. 2010; Schifanella et al. 2010; Mihalkova et al. 2011; Hopcroft et al. 2011; Scellato et al. 2011; Yang et al. 2011; Gong et al. 2014), (2) linkage decisions of their social neighbors (Newman 2001; Barabási et al. 2002; Jeh and Widom 2002; Adamic and Adar 2003; O'Madadhain et al. 2005; Kashima and Abe 2006; Lichtenwalter et al. 2010; Scellato et al. 2011; Gong et al. 2014), and (3) paths connecting them in a social network (Katz 1953; Hasan et al. 2006; Tong et al. 2006; Wang et al. 2007; Huang 2010; Lichtenwalter et al. 2010; Hopcroft et al. 2011; Gong et al. 2014). These observations can be explained by homophily theory (McPherson et al. 2001), social interaction theory (Becker 1974), and cognitive balance theory (Heider 1958) respectively. We next review these theories and explain why a link recommendation method works based on these theories.

Homophily theory states that "a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson et al. 2001, p. 416). After all, people like to hang out with others who are like them; that is, birds of a feather flock together. Theoretically, the flow of information from one person to another is a decreasing function of the distance between them in the Blau space, which is defined by socio-demographic dimensions (McPherson 1983). As a consequence, people separated by greater distances in the Blau space (i.e., less similar in terms of socio-demographics) are less likely to be linked (McPherson 1983). Furthermore, people self-select their "social world" by choosing with whom to interact. Similarity localizes communications in a social network and can lead to distinct social niches (McPherson and Ranger-Moore 1991). These self-selection effects and social niches imply that similarity contributes to the establishment of linkage in social networks (McPherson and Ranger-Moore 1991).

Homophily theory is the theoretical foundation for a number of link recommendation methods. According to this theory, a person tends to link to another person similar to him or her and the linkage between them further strengthens their similarity (McPherson and Ranger-Moore 1991). This mutual reinforcement between similarity and linkage is the foundation for the link recommendation methods developed by Yang et al. (2011) and Gong et al. (2014). Moreover, homophily theory predicts that "similarity breeds connection" (McPherson et al. 2001), which explains the effectiveness of nodal proximity-based methods (Chen et al. 2009; Crandall et al. 2010; Schifanella et al. 2010). Furthermore, using similarities between user profiles, one can build a Markov network for relational learning-based link recommendation (Mihalkova et al. 2011). In addition, homophily theory also justifies the effectiveness of similarity-based input features for link recommendation methods, including semantic similarities (O'Madadhain et al. 2005; Hasan et al. 2006; Wang et al. 2007), geographic proximities (O'Madadhain et al. 2005; Scellato et al. 2011), similarities on personal interests (Yang et al. 2011), social status similarities (Hopcroft et al. 2011), and demographic similarities (Gong et al. 2014).

According to social interaction theory (Becker 1974), when making a social decision, a social entity's utility depends on the decisions of his or her social neighbors. Specifically, a social decision is a decision that has impacts on the decision maker's social neighbors (Akerlof 1997). For link recommendation, the decision of linkage is a social decision. Similar to social interaction theory, social decision theory states that the welfare of a decision maker is affected by the outcomes of his or her social neighbors (Akerlof 1997; Maccheroni et al. 2012). Both social interaction theory and social decision theory can be reasoned from the perspective of social information processing (Salancik and Pfeffer 1978). That is, a social entity relies on social information, that is, information about what others think, to make a social decision (Kelley 1971; Berger and Calabrese 1975; Pfeffer et al. 1976; Salancik and Pfeffer 1978). In particular, a social entity's decision is affected by the decisions of his or her social neighbors in two ways (Salancik and Pfeffer 1978). First, the entity tends to fit into a social environment by agreeing with his or her social neighbors. Second, the entity's social neighbors can pass selective information to the entity, thereby affecting the entity's decision.

In short, social interaction theory, social decision theory, and social information processing theory all suggest that a social entity's decision depends on the decisions of the entity's social neighbors, which is the theoretical foundation for a group of link recommendation methods. For example, Kashima and Abe (2006) develop a link recommendation method based on the idea that a user's linkage decision in a social network depends on the linkage decisions of his or her direct friends in the network. Moreover, the influence of a social entity's neighborhood on the entity's decision, as suggested by these theories, is also encoded in the development of neighborhood-based structural proximities, which function as independent predictors for link recommendation (Newman 2001; Barabási et al. 2002; Jeh and Widom 2002; Adamic and Adar 2003) or as input features for a link recommendation method (O'Madadhain et al. 2005; Scellato et al. 2011; Gong et al. 2014).

Network transitivity distinguishes social networks from other types of networks (Newman and Park 2003). In general, network transitivity refers to the social phenomenon that indirectly associated social entities tend to tie to each other directly (Davis and Leinhardt 1971; Huang 2010). This phenomenon can be explained by cognitive balance theory from a psychological perspective (Heider 1958) or by social focus theory from a sociological standpoint (Homans 1950). According to cognitive balance theory, sentiments (or attitudes) of indirectly associated social entities could become consistent gradually, which in turn could drive them to link to each other (Heider 1958). Social focus theory suggests that people connect with those who share their foci (Homans 1950). Here, a focus is defined as "a social, psychological, legal, or physical entity around which joint activities are organized" (Feld 1981). For example, foci can be workplaces or hangouts. For two indirectly associated social entities, intermediate entities connecting them provide or create foci for them to meet and interact; as a consequence, these indirectly associated social entities are likely to be linked to each other (Feld 1981). Cognitive balance theory and social focus theory are theoretical foundations for link recommendation methods using path-based structural proximities (Katz 1953; Tong et al. 2006) and the probabilistic model-based method proposed by Huang (2010). All of these methods are developed based on network transitivity suggested by these theories. In addition, features constructed based on network transitivity have been used as inputs for several classification-based link recommendation methods (Hasan et al. 2006; Wang et al. 2007; Lichtenwalter et al. 2010; Gong et al. 2014) and the probabilistic model-based link recommendation method proposed by Hopcroft et al. (2011).

In summary, we identify major social and economic theories underlying link recommendation methods and explain working mechanisms of link recommendation methods using these theories, thereby uncovering the link between link recommendation methods and their theoretical foundations. Table 2 summarizes these theories and representative link recommendation methods that are grounded in the theories.

6 FUTURE RESEARCH DIRECTIONS

We propose to extend current link recommendation research in several directions, including utility-based link recommendation, diversity of link recommendation, link recommendation from incomplete data, and experimental study of link recommendation. These research directions focus on novel design and evaluation of link recommendation methods. In particular, we suggest incorporating additional objectives such as utility and diversity into the design of link recommendation methods, besides their traditional objective of accuracy. We also suggest the design of novel link recommendation methods that consider both observed and unobserved features driving link establishments in a social network. Finally, we recommend evaluating link recommendation methods (no matter traditional methods or novel methods suggested in this section) using experiments. In the following subsections, we identify promising research questions for each of these directions.

6.1 Utility-Based Link Recommendation

Existing link recommendation methods predict the linkage likelihood for each potential link and recommend potential links with the highest linkage likelihoods. Thus, existing methods focus on the accuracy of link recommendation. In addition to accuracy, there are other factors that need to be considered when recommending links. One is the value of link recommendation, that is, the value of a recommended link if it is established. We illustrate the value of link recommendation using an example of Facebook, whose operator (i.e., Facebook Inc.) harvests the majority of its $7.9 billion revenue from advertisements on the network (Facebook 10-K 2013). Facebook allows an advertisement being placed on the Facebook page of selected users. A Facebook user could interact with the advertisement through actions such as click, comment, like, and share. Such interactions propagate the advertisement to the user's friends, who could also interact with the advertisement and further propagate it to their friends. As this propagation process continues, the

Table 2. Theoretical Foundations for Link Recommendation Methods

Theory	Representative Work Grounded in Theory
Homophily Theory (McPherson et al. 2001)	Nodal Proximity-based Method – Chen et al. (2009) – Crandall et al. (2010) – Schifanella et al. (2010) Classification-based Method – O'Madadhain et al. (2005) – Hasan et al. (2006) – Wang et al. (2007) – Scellato et al. (2011) – Gong et al. (2014) Probabilistic Model-based Method – Hopcroft et al. (2011) – Yang et al. (2011) Relational Learning-based Method – Mihalkova et al. (2011)
Social Interaction Theory (Becker 1974) Social Information Processing Theory (Salancik and Pfeffer 1978) Social Decision Theory (Akerlof 1997)	Probabilistic Model-based Method – Kashima and Abe (2006) Structural Proximity-based Method (Neighborhood) – Newman (2001) – Barabási et al. (2002) – Jeh and Widom (2002) – Adamic and Adar (2003) Classification-based Method – O'Madadhain et al. (2005) – Lichtenwalter et al. (2010) – Scellato et al. (2011) – Gong et al. (2014)
Social Focus Theory (Homans 1950) Cognitive Balance Theory (Heider 1958)	Structural Proximity-based Method (Path) – Tong et al. (2006) Probabilistic Model-based Method – Huang (2010) – Hopcroft et al. (2011) Classification-based Method – Hasan et al. (2006) – Wang et al. (2007) – Lichtenwalter et al. (2010) – Gong et al. (2014)

advertisement could reach a much larger number of users than those that were initially selected. Facebook Inc. reaps revenue each time the advertisement reaches a Facebook user. Understandably, a recommended link, if established, leads to a more connected network among Facebook users, which in turn, drives advertisements to reach more users and bring more advertisement revenue to Facebook Inc. In this example, the value of link recommendation is the advertisement revenue brought in by a recommended and established link. Additionally, link recommendation is not costless. For example, in the context of online social network, ineffective link recommendations suggest new friends that are mostly irrelevant for the users. Consequently, new friendships recommended may not be established, and users may feel disappointed and eventually leave the network. Due to the failure of establishing new friendships and the loss of users, an online social network becomes smaller and less connected, resulting in less advertisement revenue being produced. Hence, another critical factor for link recommendation is the cost of link recommendation; that is, the cost incurred if a recommended link is not established.

We propose the utility of link recommendation that integrates the accuracy, value, and cost of link recommendation. Specifically, the utility of a recommended link depends on its linkage likelihood, the value brought in by the link if it is established, and the cost incurred if it is not established. An interesting research question is therefore how to recommend potential links with the highest utilities rather than the highest linkage likelihoods. This research question could be framed as a classification problem with the linkage likelihood, value, and cost of a potential link as predictors (Li et al. 2017). Alternatively, this question could be treated as a cost-sensitive learning problem (Fang 2013), in which, the linkage likelihood of a potential link is predicted using an existing link recommendation method and a cost matrix is constructed based on the value and cost of a potential link.

6.2 Diversity of Link Recommendation

Besides accuracy and utility, another important objective of link recommendation is diversity, for example, recommending friends with diverse backgrounds to a user. Diverse link recommendation benefits both individual users and a social network as a whole. A user who receives friend recommendations with diverse backgrounds could gain accesses to different social communities in a social network (Brandão et al. 2013), thereby obtaining significant social benefits such as a variety of information sources and competitive advantages (Burt 1992). A social network with high structural diversity among its users effectively facilitates the diffusion of information over the network (Ugander et al. 2012). Moreover, such a network is robust in terms of information diffusion efficiency, even after the removal of its well-connected users (Albert et al. 2000; Tanizawa et al. 2005). In addition, according to Eagle et al. (2010), high structural diversity in a social network has been shown to be positively correlated with a high level of socioeconomic well-being of its users (e.g., high income).

Despite these benefits, link recommendation diversity, however, has not received much research attention. In this subsection, we propose several research questions in this fruitful area. First, it is necessary to design metrics to gauge the diversity of link recommendation. While a number of measures have been proposed to evaluate the recommendation diversity of recommender systems (e.g., Zhang and Hurley (2008) and Vargas and Castells (2011)), many of these measures are not readily applicable to measuring the diversity of link recommendation, because a recommender system is a user-item bipartite graph whereas link recommendation involves a user-user graph. We therefore suggest some ideas for designing metrics for link recommendation diversity. One possible metric would be the number of connected components among friends recommended to a user. Understandably, a larger number of connected components indicates that recommended friends are distributed in more social groups and hence greater link recommendation diversity. We could also evaluate link recommendation diversity by measuring the diversity of a social network after adding recommended links with classical network metrics such as clustering coefficient (Watts and Strogatz 1998), network diameter (Wasserman and Faust 1994), and number of structural holes (Burt 1992). Second, it would be interesting to empirically evaluate the recommendation diversity of popular link recommendation methods using the designed metrics. Third, methods that balance multiple objectives of link recommendation, such as diversity, utility, and accuracy, are needed. In this vein, we could formulate a problem of maximizing the diversity of link recommendation while maintaining the accuracy (or utility) of link recommendation at a certain level or a problem of maximizing the accuracy (or utility) of link recommendation while maintaining the diversity of link recommendation at a certain level, and develop methods to solve these problems.

6.3 Link Recommendation from Incomplete Data

Current link recommendation methods employ observed features of a social network to predict linkage likelihood. Some of them target nodal features (e.g., Backstrom and Leskovec (2011)), while others using structural features (e.g., Liben-Nowell and Kleinberg (2007)). However, unobserved factors for which we do not have data could also drive link establishments in a social network. Consider the following examples:

- Parents become friends on Facebook because their children attend the same kindergarten. These parents have quite different profiles and would never become friends if their children were not attending the same kindergarten.
- A group of researchers who have different backgrounds and are located in geographically separated areas are temporarily assigned to work together on an interdisciplinary project. Some of them then link to each other on LinkedIn.

The factors driving the link establishments in these examples are children's attendance at the same kindergarten and temporary project assignments. We do not have data about these factors (i.e., they are unobserved), partly for the following reasons. First, users may not disclose sensitive information such as which kindergarten their children attend due to privacy concerns. Second, users live in a social environment that is much broader than a social network. As a result, many factors extrinsic to a social network, such as temporary project assignments, could drive link establishments in the network. The sheer number of extrinsic factors that might affect link establishments in a social network make it extremely difficult to collect data about all of them. Existing methods that rely on similarities between observed user profiles fail to predict link establishments in these examples, because observed user profiles in these examples are quite dissimilar. When link establishments in a social network are primarily driven by observed features of the network, considering unobserved factors may not improve link recommendation performance. However, when link establishments in a social network are partly (or even primarily) driven by unobserved factors (e.g., examples discussed in this subsection), considering them is important for enhancing link recommendation performance. Therefore, an interesting research question is how to integrate both observed features and unobserved factors for link recommendation. Toward this end, methods built on the expectation-maximization (EM) framework (Dempster et al. 1977; Fang et al. 2013a), a classical framework for learning from incomplete data, are needed for link recommendation from incomplete data. In general, an EM-based method is an iterative procedure starting from an initial paramter estimation for both obseved features and unobserved factors. Parameters for unobserved factors are randomly set intially. Each iteration of an EM-based method consists of the expectation step based on current paramete estimation and the maximization step, which maximizes the expectation and computes a revised parameter estimation. The iterative procedure terminates and outputs the final paramter estimation for both obseved features and unobserved factors.

6.4 Experimental Study of Link Recommendation

Most link recommendation methods are evaluated using archival data. However, simply using archival data, we cannot differentiate between link establishments due to organic growth and link establishments arising from link recommendation. Furthermore, with archival data alone, it is difficult to assess users' behavioral and emotional reactions to recommended links. Therefore, it is worthwhile to conduct laboratory or field experiments to evaluate link recommendation methods. In an experiment, we can observe in real time how users react to recommended links, both behaviorally and emotionally, which recommended links they actually establish, and the value of the established links along with the costs of links that are not established. Moreover, we could use laboratory or field experiments to study the delivery mechanisms for link recommendation. Interesting questions in this direction include: How should the user interface that presents recommended links be designed? What are the best timing and frequency for recommending links that will maximize the chance of link establishment? How many links should be recommended to a user? And what incentives can be used to facilitate the establishment of recommended links? In addition, we could design experiments to study the impact of link recommendation on social networks. Interesting questions in this direction include: how link recommendation impacts the diffusion of information, opinion, or advertisements in a social network and how link recommendation affects user engagement in a social network. In short, experimental study provides a means to address interesting link recommendation questions, many of which cannot be answered using archival data alone. By combining laboratory or field experiments and evaluations with archival data, we could produce more comprehensive and convincing evaluations of link recommendation methods.

Let us discuss one experimental study in detail. In general, a social network consists of two types of links: short-range and long-range links. While short-range links build local communities among similar individuals, long-range links connect separated individuals who have complementary skills or characteristics. Long-range links are valuable, because they are necessary for reducing the diameter of a network (Kleinberg 2000). However, long-range links are harder to predict compared to their short-range counterparts (Lichtenwalter et al. 2010). Therefore, recommending long-range links is riskier but can generate greater benefits when successful, while recommending short-range links is less risky but generates fewer benefits. Given these differences, it remains to be determined which type of links would yield better business outcomes, for example, user engagement, for a social network. To address this question, we can formulate the following hypothesis:

> *Hypothesis: Users are more engaged in a social network when recommended with long-range links than short-range links.*

To test this hypothesis with an experiment, it is necessary to control for several important variables, including selection bias, homophily effects, confounding factors, and network effects. Selection bias occurs when treatments are assigned to recipients with an uneven distribution. Heterogeneity in the content or valence of the treatment is another source of selection bias (Aral and Walker 2012). Similar social entities tend to behave similarly, which creates homophily bias in network contexts (Aral et al. 2009). Homophily effects can be eliminated by controlling for nodal characteristics in a social network. Confounding factors, such as email campaign and advertisement, constitute another important source of bias (Aral and Walker 2012; Xu et al. 2015). Network effects refer to the possible occasions that the effect of a treatment will spill over to the recipient's social circles, regardless of whether these social neighbors are in the control or treatment group (Xu et al. 2015). Network effects are ubiquitous and pose a special challenge for network-based experiments.

6.5 Other Future Research Questions

There are other interesting future research questions, which we discuss briefly for length consideration. The effectiveness of a link recommendation method decreases over time, because the method was developed based on previous user linkage behaviors and does not capture new user linkage behaviors. Thus, an interesting question is how to maintain the currency of a link recommendation method over time. Prior studies on data stream mining (Aggarwal 2007) and knowledge refreshing (Fang et al. 2013b) provide methodological and theoretical foundations for answering this question. In addition, current link recommendation methods rank potential links by their linkage likelihoods and recommend top-ranked links. Thus, existing methods essentially solve a ranking problem. However, many real-world link recommendation situations may be modeled as combinatorial rather than ranking problems. For example, how to recommend a set of K potential links that will collectively bring in the largest advertisement revenue, among all possible combinations of K potential links, is a combinatorial problem. Thus, it is interesting to study how to model and solve combinatorial link recommendation problems. There are also other challenges for link recommendation, such as cold start (Leroy et al. 2010), class imbalance (Rattigan and Jensen 2005; Lichtenwalter et al. 2010), and link recommendation across multiple (heterogeneous) social networks (Zhang et al. 2014; Dong et al. 2015), which are well documented in the literature and hence are not discussed here.

7 CONCLUSIONS

We review state-of-the-art link recommendation methods in this survey. We also identify social and economic theories underlying link recommendation methods and uncover the missing link between them. We further suggest several extensions to existing link recommendation studies, including utility-based link recommendation, diversity of link recommendation, link recommendation from incomplete data, and experimental study of link recommendation. We hope this survey serves as a useful summary of what has been done as well as a stimulus to the advancement of link recommendation research.

REFERENCES

E. Acar, D. M. Dunlavy, and T. G. Kolda. 2009. Link prediction on evolving data using matrix and tensor factorizations. In *Proceedings of IEEE International Conference on Data Mining Workshops*. 262–269.

S. Adali, F. Sisenda, and M. Magdon-Ismail. 2012. Actions speak as loud as words: Predicting relationships from social behavior data. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)* 689–698.

L. A. Adamic and E. Adar. 2003. Friends and neighbors on the web. *Soc. Netw.* 25, 3 (2003), 211–230.

C. Aggarwal. 2007. *Data Streams: Models and Algorithms*. Springer, New York, NY.

G. A. Akerlof. 1997. Social distance and social decisions. *Econometrica* 65, 5 (1997), 1005–1027.

R. Albert, H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature* 406, 6794 (2000), 378–382.

S. Aral, L. Muchnik, and A. Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. U.S.A.* 106, 51 (2009), 21544–21549.

S. Aral and D. Walker. 2012. Identifying influential and susceptible members of social networks. *Science* 337, 6092 (2012), 337–341.

J. Akehurst, I. Koprinska, K. Yacef, L. Pizzato, J. Kay, and T. Rej. 2011. A hybrid content-collaborative reciprocal recommender for online dating. In *Proceedings of International Joint Conference on Artificial Intelligence.*

S. B. Bacharach. 1989. Organizational theories: Some criteria for evaluation. *Acad. Manage. Rev.* 14, 4 (1989), 496–515.

L. Backstrom and J. Leskovec. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. 635–644.

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

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

N. Barbieri, F. Bonchi, and G. Manco. 2014. Who to follow and why: Link prediction with explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 1266–1275.

G. Becker. 1974. A Theory of social interactions. *J. Polit. Econ.* 82 (1974), 1063–1093.

N. Benchettara, R. Kanawati, and C. Rouveirol. 2010. A supervised machine learning link prediction approach for academic collaboration recommendation. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 253–256.

C. R. Berger and R. J. Calabrese. 1975. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Hum. Commun. Res.* 1, 2 (1975), 99–112.

C. A. Bliss, M. R. Frank, C. M. Danforth, and P. S. Dodds. 2014. An evolutionary algorithm approach to link prediction in dynamic social networks. *J. Comput. Sci.* 5, 5 (2014), 750–764.

M. Bilgic, G. M. Namata, and L. Getoor. 2007. Combining collective classification and link prediction. In *Proceedings of 7th IEEE International Conference on Data.* 381–386.

M. A. Brandão, M. M. Moro, G. R. Lopes, and J. P. M. Oliveira. 2013. Using link semantics to recommend collaborations in academic social networks. In *Proceedings of the 22nd International Conference on World Wide Web Companion (WWW'13).* 833–840.

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

R. Burt. 1992. *Structural Holes: The Social Structure of Competition.* Harvard University Press, Cambridge, MA.

X. Cai, M. Bain, A. Krzywicki, W. Wobcke, Y. Kim, P. Compton, and A. Mahidadia. 2012. Reciprocal and heterogeneous link prediction in social networks. *Adv. Knowl. Discov. Data Min.* (2012) 193–204.

J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. 2009. Make new friends, but keep the old: Recommending people on social networking sites. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems.* 201–210.

J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg. 2011. Predicting reciprocity in social networks. In *Proceedings of the IEEE 3rd International Conference on Privacy, Security, Risk and Trust and the IEEE 3rd International Conference on Social Computing.* 49–56.

A. Clauset, C. Moore, and M. E. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 7191 (2008), 98–101.

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

T. Davenport and D. J. Patil. 2012. Data scientist: the sexist job of the 21st century. *Harv. Bus. Rev.* 90, 10 (2012), 70–76.

J. A. Davis and S. Leinhardt. 1971. The structure of positive interpersonal relations in small groups. In *Sociological Theories in Progress* (2nd ed.). Houghton-Mifflin, Boston.

S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer. 1990. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* 41, 6 (1990), 391–407.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Stat. Soc. Ser. B* 39, 1, 1–38.

J. R. Doppa, J. Yu, P. Tadepalli, and L. Getoor. 2010. Learning algorithms for link prediction based on chance constraints. *Machine Learning and Knowledge Discovery in Databases.* Springer, Berlin, 344–360.

Y. Dong, J. Tang, S. Wu, J. Tian, N. V. Chawla, J. Rao, and H. Cao. 2012. Link prediction and recommendation across heterogeneous social networks. In *Proceedings of the 12th International Conference on Data Mining.* 181–190.

Y. Dong, J. Zhang, J. Tang, N. V. Chawla, and B. Wang. 2015. CoupledLP: Link prediction in coupled networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15).* 199–208.

D. M. Dunlavy, T. G. Kolda, and E. Acar. 2011. Temporal link prediction using matrix and tensor factorizations. *ACM Trans. Knowl. Discov. Data* 5, 2 (2011), 1–27.

N. Eagle, M. Macy, and R. Claxton. 2010. Network diversity and economic development. *Science* 328, 5981 (2010), 1029–1031.

Facebook 10-K 2013. Form 10-K of Facebook Inc. for the Fiscal Year Ended December 31, 2013. Securities and Exchange Commission's Edgar Website. Retrieved on January 7, 2016 from https://www.sec.gov/edgar.shtml.

X. Fang, P. Hu, Z. Li, and W. Tsai. 2013a. Predicting adoption probabilities in social networks. *Inf. Syst. Res.* 24, 1 (2013), 128–145.

X. Fang, O. R. Liu Sheng, and P. Goes. 2013b. When is the right time to refresh knowledge discovered from data? *Operat. Res.* 61, 1 (2013), 32–44.

X. Fang. 2013. Inference-based naïve bayes: Turning naïve bayes cost-sensitive. *IEEE Trans. Knowl. Data Eng.* 25, 10 (2013), 2302–2313.

Scott L. Feld. 1981. The focused organization of social ties. *Am. J. Sociol.* 86, 5 (1981), 1015–1035.

M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. 2011. Link prediction in social networks using computationally efficient topological features. In *Proceedings of Privacy, Security, Risk and Trust and IEEE 3rd International Conference on Social Computing.* 73–80.

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

N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. 1999. Learning probabilistic relational models. In *Proceedings of International Joint Conferences on Artificial Intelligence* 1300–1309.

L. Getoor, N. Friedman, D. Koller, and B. Taskar. 2001. Learning probabilistic models of relational structure. In *Proceedings of International Conference on Machine Learning (ICML'01).* 170–177.

L. Getoor, N. Friedman, D. Koller, and B. Taskar. 2003. Learning probabilistic models of link structure. *J. Mach. Learn. Res.* 3 (2003), 679–707.

L. Getoor and B. Taskar. 2007. *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA.

N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. R. Shhi, and D. Song. 2014. Jointly predicting links and inferring attributes using a social-attribute network. *ACM Trans. Intell. Syst. Technol.* 5, 2 (2014), 1–20.

R. Guimerà and S. Marta. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci U.S.A.* 106, 52 (2009), 22073–22078.

R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. 2004. Propagation of trust and distrust. In *Proceedings of the 23rd International Conference on World Wide Web*. 403–412.

M. Hasan, V. Chaoji, S. Salem, and M. Zaki. 2006. Link prediction using supervised learning. In *SIAM International Conference on Data Mining Workshop on Link Analysis, Counter-terrorism and Security*.

M. Hasan and M. Zaki. 2009. A survey of link prediction in social networks. In *Social Network Data Analytics*. Springer, New York, NY.

T. H. Haveliwala. 2002. Topic-sensitive Pagerank. In *Proceedings of the 11th International Conference on World Wide Web (WWW'02)*. 517–526.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. 2001. Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.* 1 (2001), 49–75.

D. Heckerman, C. Meek, and D. Koller. 2004. Probabilistic models for relational data. Technical Report MSR-TR-2004-30. Microsoft Research.

F. Heider. 1958. *The Psychology of Interpersonal Relations*. Wiley, New York, NY.

T. Hofmann. 2004. Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.* 22, 1 (2004), 89–115.

G. C. Homans. 1950. *The Human Group*. Harcourt, Brace, and World, New York, NY.

J. Hopcroft, T. Lou, and J. Tang. 2011. Who will follow you back?: Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. 1137–1146.

Z. Huang. 2010. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. Retrieved from http://dx.doi.org/10.2139/ssrn.1634014.

Z. Huang and D. K. J. Lin. 2009. The time-series link prediction problem with applications in communication surveillance. *INFORMS J. Comput.* 21, 2 (2009), 286–303.

G. Jeh and J. Widom. 2002. SimRank: A measure of structural-context similarity. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*. 538–543.

G. Jeh and J. Widom. 2003. Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*. 271–279.

B. Jiang, Z. Zhang, and D. Towsley. 2015. Reciprocity in social networks with capacity constraints. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15)*. 457–466.

I. Kahanda and J. Neville. 2009. Using transactional information to predict link strength in online social networks. In *Proceedings of the 3rd International Conference on Web and Social Media (ICWSM'09)*. 74–81.

H. Kashima and N. Abe. 2006. A parameterized probabilistic model of network evolution for supervised link prediction. In *Proceedings of the 6th International Conference on Data Mining (ICDM'06)*. 340–349.

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

H. H. Kelley. 1971. Attribution in social interaction. *General Learning Press*, New York, NY.

J. G. Kemeny and J. L. Snell. 1976. *Finite Markov Chains 210*. Springer-Verlag, New York, NY.

M. Kim and J. Leskovec. 2011. The network completion problem: inferring missing nodes and edges in networks. In *Proceedings of SIAM International Conference on Data Mining*. 47–58.

J. M. Kleinberg. 2000. Navigation in a small world. *Nature* 406, 6798 (2000), 845–845.

Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.

J. Kunegis and A. Lommatzsch. 2009. Learning spectral graph transformations for link prediction. In *Proceedings of the 29th International Conference on Machine Learning (ICML'09)*. 561–568.

J. Kunegis, A. Lommatzsch, and C. Bauckhage. 2009. The slashdot zoo: Mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*. 741–750.

T. T. Kuo, R. Yan, Y. Y. Huang, P. H. Kung, and S. D. Lin. 2013. Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*. 775–783.

E. A. Leicht, P. Holme, and M. E. J. Newman. 2006. Vertex similarity in networks. *Phys. Rev. E* 73, 2 (2006), 026120.

V. Leroy, B. B. Cambazoglu, and F. Bonchi. 2010. Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 393–402.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. 2010a. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1361–1370.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. 2010b. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*. 641–650.

Z. Li, X. Fang, X. Bai, and O. R. Liu Sheng. 2017. Utility-based link recommendation for online social networks. *Manage. Sci.* 63, 6 (2017), 1938–1952.

D. Liben-Nowell and J. Kleinberg. 2007. The link prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* 58, 7 (2007), 1019–1031.

R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*. 243–252.

W. Liu and L. Lü. 2010. Link prediction based on local random walk. *Europhys. Lett.* 89, 5 (2010), 58007.

T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. 2013. Learning to predict reciprocity and triadic closure in social networks. *ACM Trans. Knowl. Discov. Data*, 7, 2 (2013), 5.

Z. Lu, B. Savas, W. Tang, and I. S. Dhillon. 2010. Supervised link prediction using multiple sources. In *Proceedings of 10th International Conference on Data Mining (ICDM'10)*. 923–928.

L. Lü, C.-H. Jin, and T. Zhou. 2009. Similarity index based on local paths for link prediction of complex networks. *Phys. Rev. E* 80, 4, 046122.

L. Lü and T. Zhou. 2011. Link prediction in complex networks: A survey. *Physica A* 390, 6 (2011), 1150–1170.

F. Maccheroni, M. Marinacci, and A. Rustichini. 2012. Social decision theory: Choosing within and between groups. *Rev. Econ. Stud.* 79, 4 (2012), 1591–1636.

M. Makrehchi. 2011. Social link recommendation by learning hidden topics. In *Proceedings of the 5th ACM Conference on Recommender Systems*. 189–196.

J. M. McPherson. 1983. Ecology of affiliation. *Am. Sociol. Rev.* 48 (1983), 519–532.

J. M. McPherson and J. R. Ranger-Moore. 1991. Evolution on a dancing landscape: Organizations and networks in dynamic blau space. *Soc. Forces* 70 (1991), 19–42.

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

A. K. Menon and C. Elkan. 2011. Link prediction via matrix factorization. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*. 437–452.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21 (1953), 1087.

L. Mihalkova, W. Moustafa, and L. Getoor. 2011. Learning to predict web collaborations. In *Workshop on User Modeling for Web Applications*.

M. E. Newman. 2001. Clustering and preferential attachment in growing networks. *Phys. Rev. E* 64, 2 (2001), 025102.

M. E. Newman, S. H. Strogatz, and D. J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* 64, 2 (2001), 026118.

M. E. Newman and J. Park. 2003. Why social networks are different from other types of networks. *Phys. Rev. E* 68, 3 (2003), 036122.

M. Nowak and K. Sigmund. 2005. Evolution of indirect reciprocity. *Nature* 437 (2005), 1291–1298.

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

J. Pan and Q. Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2010), 1345–1359.

A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. 2012. Fast and accurate link prediction in social networking systems. *J. Syst. Softw.* 85, 9 (2012), 2119–2132.

T. Parsons. 1938. The role of theory in social research. *Am. Sociol. Rev.* 3, 1 (1938), 13–20.

J. Pfeffer, G. R. Salancik, and H. Leblebici. 1976. The effect of uncertainty on the use of social influence in organizational decision making. *Admin. Sci. Quart.* 21, 2 (1976), 227–245.

A. Popescul and L. Ungar. 2003. Statistical relational learning for link prediction. In *Proceedings of International Joint Conferences on Artificial Intelligence Workshop on Learning Statistical Models from Relational Data*. 81–90.

D. Quercia and L. Capra. 2009. Friendsensing: Recommending friends using mobile phones. In *Proceedings of the 3rd ACM conference on Recommender Systems*. 273–276.

M. J. Rattigan and D. Jensen. 2005. The case for anomalous link discovery. *SIGKDD Explor. Newslett.* 7, 2 (2005), 41–47.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. Barabási. 2002. Hierarchical organization of modularity in metabolic networks. *Science* 297, 5586 (2002), 1551–1555.

J. Rennie and N. Srebro. 2005. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd International Conference on Machine Learning*. 713–719.

G. R. Salancik and J. Pfeffer. 1978. A social information processing approach to job attitudes and task design. *Admin. Sci. Quart.* 224–253.

G. Salton and M. J. McGill. 1986. *Introduction to Modern Information Retrieval.* McGraw-Hill, New York, NY.

G. Salton. 1989. *Automatic Text Processing: the Transformation Analysis, and Retrieval of Information by Computer.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

S. Scellato, A. Noulas, and C. Mascolo. 2011. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11).* 1046–1054.

R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. 2010. Folks in folksonomies: Social link prediction from shared metadata. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10).* 271–280.

D. Shen, J. Sun, Q. Yang, and Z. Chen. 2006. Latent friend mining from blog data. In *Proceedings of IEEE 6th International Conference on Data Mining (ICDM'06).* 552–561.

D. Shin, S. Si, and I. S. Dhillon. 2012. Multi-scale link prediction. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management.* 215–224.

A. Singhal, K. Subbian, J. Srivastava, T. Kolda, and A. Pinar. 2013. Dynamics of trust reciprocation in multi-relational networks. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13).* 661–665.

D. Song, A. Meyer, and D. Tao. 2015. Efficient latent link recommendation in signed networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15).* 1105–1114.

H. H. Song, T. W. Cho, V. Dave, Y. Zhang, and L. Qiu. 2009. Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference.* 322–335.

D. Song, D. A. Meyer, and D. Tao. 2015. Efficient latent link recommendation in signed networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'15).* 1105–1114.

T. Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol. Skr.* 5 (1948), 1–34.

P. Symeonidis, E. Tiakas, and Y. Manolopoulos. 2010. Transitive node similarity for link prediction in social networks with positive and negative links. In *Proceedings of the 4th ACM Conference on Recommender Systems* 183–190.

J. Tang, S. Chang, C. Aggarwal, and H. Liu. 2015. Negative link prediction in social media. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM'15).* 87–96.

J. Tang, T. Lou, J. Kleinberg, and S. Wu. 2015a. Transfer learning to infer social ties across heterogeneous networks. *ACM Trans. Inf. Syst.* 34, 3, Article 1 (December 2015).

T. Tanizawa, G. Paul, R. Cohen, S. Havlin, and H. E. Stanley. 2005. Optimization of network robustness to waves of targeted and random attacks. *Phys. Rev. E* 71, 4 (2005), 047101.

T. Taskar, M. F. Wong, P. Abbeel, and D. Koller. 2003. Link prediction in relational data. In *Proceedings of Advances in Neural Information Processing Systems.* 659–666.

H. Tong, C. Faloutsos, and J. Y. Pan. 2006. Fast random walk with restart and its applications. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM'06).* 613–622.

J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. 2012. Structural diversity in social contagion. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16 (2012), 5962–5966.

S. Vargas and P. Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems.* 109–116.

D. J. Watts and S. H. Strogatz. 1998. Collective dynamics of 'small world' networks. *Nature* 393, 6684 (1998), 440–442.

S. Wasserman and K. Faust. 1994. *Social Network Analysis: Methods and Applications.* Cambridge University Press, Cambridge, UK.

C. Wang, V. Satuluri, and S. Parthasarathy. 2007. Local probabilistic models for link prediction. In *Proceedings of 7th IEEE International Conference on Data Mining (ICDM'07).* 322–331.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11).* 1100–1108.

R. Xiang, J. Neville, and M. Rogati. 2010. Modeling relationship strength in online social networks. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).* 981–990.

Y. Xu and D. Rockmore. 2012. Feature selection for link prediction. In *Proceedings of the 5th Ph.D. ACM Workshop on Information and Knowledge.* 25–32.

Y. Xu, N. Chen, A. Fernandez, O. Sinno, and A. Bhasin. 2015. From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2227–2236.

S. H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. 2011. Like like alike: Joint friendship and interest propagation in social networks. In *Proceedings of the 20th International Conference on World Wide Web (WWW'11)*. 537–546.

S. Yang, A. Smola, B. Long, H. Zha, and Y. Chang. 2012. Friend or frenemy? predicting signed ties in social networks. In *Proceedings of the 35th ACM SIGIR Conference on Research and Development in Information Retrieval*. 555–564.

J. Ye, H. Cheng, Z. Zhu, and M. Chen. 2013. Predicting positive and negative links in signed social networks by transfer learning. In *Proceedings of the 22nd International Conference on World Wide Web*. 1477–1488.

Z. Yin, M. Gupta, T. Weninger, and J. Han. 2010. A unified framework for link recommendation using random walks. In *Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining*. 152–159.

K. Yu, W. Chu, S. Yu, T. Volker, and Z. Xu. 2006. Stochastic relational models for discriminative link prediction. *Adv. Neur. Inf. Process. Syst.* P. B. Schölkopf, J. C. Platt, and T. Hoffman (Eds.). MIT Press. 333–340.

G. Yuan, P. K. Murukannaiah, Z. Zhang, and M. P. Singh. 2014. Exploiting sentiment homophily for link prediction. In *Proceedings of the 8th ACM Conference on Recommender Systems*. 17–24.

M. Zhang and N. Hurley. 2008. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of ACM Conference on Recommender Systems*. 123–130.

J. Zhang, P. S. Yu, and Z. H. Zhou. 2014. Meta-path based multi-network collective link prediction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*. 1286–1295.

K. Zhao, X. Wang, M. Yu, and B. Gao. 2014. User recommendations in reciprocal and bipartite social networks: An online dating case study. *IEEE Intell. Syst.* 29, 2 (2014), 27–35.

E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. 2008. Using friendship ties and family circles for link prediction. In *Proceedings of the 2nd International Conference on Advances in Social Network Mining and Analysis*. Springer-Verlag, Berlin, 97–113.

T. Zhou, L. Lü, and Y. Zhang. 2009. Predicting missing links via local information. *Eur. Phys. J B 71*, 4 (2009), 623–630.

Received February 2016; revised April 2017; accepted August 2017