European Journal of Operational Research 260 (2017) 693–705

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Innovative Applications of O.R.

A utility-based link prediction method in social networks

Yongli Li[a,*], Peng Luo[b,**], Zhi-ping Fan[a], Kun Chen[c], Jiaguo Liu[d]

[a] School of Business Administration, Northeastern University, Shenyang 110169, PR China
[b] School of Management, Harbin Institute of Technology, Harbin 150001, PR China
[c] Department of Financial Mathematics & Financial Engineering, South University of Science and Technology of China, Shenzhen 518055, PR China
[d] Transportation Management College, Dalian Maritime University, Dalian 116026, PR China

ARTICLE INFO

Article history:
Received 9 April 2016
Accepted 23 December 2016
Available online 28 December 2016

Keywords:
Networks
Link prediction
Utility analysis
EM algorithm
Latent variable

ABSTRACT

Link prediction is a fundamental task in social networks, with the goal of estimating the likelihood of a link between each node pair. It can be applied in many situations, such as friend discovery on social media platforms or co-author recommendations in collaboration networks. Compared to the numerous traditional methods, this paper introduces utility analysis to the link prediction method by considering that individual preferences are the main reason behind the decision to form links, and meanwhile it also focuses on the meeting process that is a latent variable during the process of forming links. Accordingly, the link prediction problem is formulated as a machine learning process with latent variables; therefore, an Expectation–Maximization (EM, for short) algorithm is adopted and further developed to cope with the estimation problem. The performance of the present method is tested both on synthetic networks and on real-world datasets from social media networks and collaboration networks. All of the computational results illustrate that the proposed method yields more satisfying link prediction results than the selected benchmarks, and in particular, logistic regression, as a special case of the proposed method, provides the lower boundary of the likelihood function.

1. Introduction

Social networks have become increasingly prevalent with the growth of social interactions online and offline. A fundamental task of social networks is link prediction, the goal of which is to estimate the likelihood of new link formation (Lichtenwalter, Lussier, & Chawla, 2010). The most prominent application of link prediction methods in social networks consists of friend recommendations for information sharing. For example, link prediction can help individuals to rapidly find friends on social media platforms such as Facebook or Twitter (Ballings & Van den Poel 2015; Barbieri, Bonchi, & Manco, 2014), or can benefit authors searching for promising co-authors to finish state-of-the-art research (Li, Wu, Wang, & Luo, 2014; Shibata, Kajikawa, & Sakata, 2012). In the first example, the individual who can rapidly find friends on a social media platform via a link prediction method might engage the social media platform very often and even feel loyalty towards, thus enhancing the growth and sustenance of the social media platform; in the second example, the author who is able to find ideal co-authors via a link prediction method might create far better papers, thus enriching human knowledge. From the aforementioned two examples, which will be further extended in the application section of this paper, there is no doubt that a link prediction method for social networks would be useful and meaningful. In addition, link prediction can also be seen as a foundation for studying the evolution of social networks (Hellmann & Staudigl, 2014), and it can also be used in many other fields such as opinion evolution and diffusion (Liu, Li, Tang, Ma, & Tian, 2014), online commodity recommendations (Li & Chen, 2013) and so forth. Accordingly, providing a novel link prediction method, in our opinion, would be a fundamental contribution to the field of networks (Huang & Lin 2009; Richard, Gaïffas, & Vayatis, 2012).

Link formation in a social network reflects choice driven by individual preferences. From this viewpoint, a link prediction method can be developed towards uncovering individual preferences based on the observed network structure, as well as individual attributes when available. Utility analysis is commonly considered one of the most powerful tools for revealing and depicting individual preferences (Raju, Burke, & Normand, 1990). However, to the best of our knowledge, few papers have considered combining utility analysis with the process of link formation or further link prediction method. From the perspective of utility analysis, the individuals in a network are regarded as intelligent agents who can make

[*] Corresponding author.
[**] Corresponding author.
E-mail addresses: ylli@mail.neu.edu.cn, yongli.0440004@gmail.com (Y. Li), luopeng_hit@126.com (P. Luo).

http://dx.doi.org/10.1016/j.ejor.2016.12.041
0377-2217/© 2016 Elsevier B.V. All rights reserved.

decisions to build links with others by maximizing their utility functions. Because of the benefits from utility analysis, adopting it to our link prediction method is the first motivation of this paper.

The second motivation of this paper is to consider the meeting process of individuals. Generally speaking, the meeting process is always embedded in the process of link formation, where the used “meeting” not only means face-to-face interactions but also includes all types of online, offline and any other forms of interactions. Furthermore, if two individuals have formed a link, they must have met each other; however, if two individuals have not formed a link, two different reasons exist: they have met each other but do not want to form a link, or they have not met each other. Accordingly, a pure focus on the utility analysis, or say ignoring the existed meeting processing, cannot explain the unformed links due to lack of meeting opportunity. Although considering the meeting process would be meaningful in link prediction, it is not an easy task since the meeting process cannot be observed directly in most cases. This paper will introduce the meeting process into the mentioned utility analysis so as to present a realistic prediction model with satisfying predictive accuracy.

Similar to the existing literature, such as Barbieri et al. (2014), Kashima, Kato, Yamanishi, Sugiyama, and Tsuda (2009), Nguyen and Mamitsuka (2012) and many others, the link prediction method is also seen as a data mining task in this paper. Taking the observed network structure and the available individual attributes as inputs, the proposed method will first yield the estimates of meeting probability and the preference parameters embedded in the designed utility functions and will then predict the likelihood of forming potential links based on the estimated parameter values. From this viewpoint, our method falls into the supervised learning category, with the aim of obtaining much more accurate parameter estimates (see Elkabani & Khachfeh, 2015). Thus, developing the estimation technique is the third motivation of this paper.

Following the above motivations, this paper aims to accomplish the following work: (1) to introduce utility analysis to the link prediction problem; (2) to divide the link formation process into individual meeting process and decision making process; (3) to develop a supervised learning algorithm that can be used for parameter estimation and link prediction in both undirected and directed social networks; (4) to demonstrate that logistic regression is a special case of the proposed method; and (5) to deduce the estimation lower boundary of the proposed method in terms of the maximum likelihood function. In addition, several numerical experiments are designed elaborately to validate the accuracy and properties of the proposed method, and also three applications are also illustrated.

2. Background and related work

In this section, we provide a brief overview of the EM algorithm and review the relevant research in the field of link prediction. In fact, numerous link prediction methods have been developed in recent decades. These methods can be classified into two main categories: unsupervised methods and supervised methods. In related work, we introduce several specific methods in each category based on the work of Liben-Nowell and Kleinberg (2007), Lü and Zhou (2011), and Lee et al. (2015).

2.1. Expectation–Maximization algorithm

The EM algorithm is often adopted to compute maximum-likelihood estimates when the problem includes some hidden states (Dempster, Laird, & Rubin, 1977). Generally speaking, the estimation process undergoes two alternating steps. *Expectation step* (E step, for short). Based on the obtained parameter estimates in the last iteration cycle, the probabilities of the hidden states are computed in this step. *Maximization step* (M step, for short). Based on the computed distribution over the hidden states in E step of this iteration cycle, the expected log-likelihood of the targeting parameters is maximized to obtain their estimates.

EM algorithm has been proven to be convergent because these two iterative steps monotonically increase the values of the log-likelihood function by updating these parameters. As a well-known machine learning algorithm, EM algorithm has often been creatively adopted to deal with the real world problems. For example, Fang, Hu, Li, and Tsai (2013) adopted this algorithm to predict the adoption probability of goods or opinion within the context of social networks, Dirick, Claeskens, and Baesens (2015) also developed the EM algorithm to estimate the multiple event mixture cure model, and also the many others not mentioned.

2.2. Unsupervised link prediction methods

From the perspective of machine learning, the unsupervised method is always adopted in cases where no explicit labels or evaluations can be utilized to determine or estimate the values of these targeting parameters (Hastie, Tibshirani, & Friedman, 2009). In particular, the most widely used link prediction methods in earlier periods, such as AA (Adamic & Adar, 2003), Katz method (Katz, 1953) and Preferential Attachment (Barabási et al., 2002), were all established based on the idea of unsupervised learning.

The aforementioned methods are also called scoring methods in many studies, such as Liben-Nowell and Kleinberg (2007) and Lee et al. (2015), in which the detailed calculation processes of these methods can also be found. The common factor of these methods is that a score for each link is calculated via the static representation of a given network, and the orderings of the scores matter. The difference between them arises from their different focuses on one or more aspects of network structure; for example, some methods pay attention to the common friends shared by two nodes, such as JI, whereas others focus on the node centrality, such as Katz, etc. These unsupervised link prediction methods can reflect one or several aspects of network structures concerned so that they could be useful in some cases but not all, and they could be regarded as benchmarks for comparative analysis.

2.3. Supervised link prediction methods

Unlike the unsupervised link prediction method, the supervised methods focus on parameter estimation, in most cases, based on the observed links (seen as labels in the field of supervised learning), network structures and node attributes. Numerous supervised link prediction methods have been presented in recent years. These methods include support vector machine (SVM, for short; Ben-Hur & Nobel, 2005; Bleakley, Biau, & Vert, 2007), the hierarchical structure model (HSM for short; Clauset, Moore, & Newman, 2008), random forest (RF for short; Guns & Rousseau, 2014), adaptive boosting (AB for short; Wang, Gao, Chen, Mensah, & Fu, 2015) and so forth. The method proposed in this paper belongs to this type, and furthermore it could be regarded as a novel approach parallel to the aforementioned approaches, since the mentioned existing ones do not start from the individual utilities. Similar to the unsupervised ones, these existing supervised methods are also the benchmarks for comparative analysis.

3. Utility analysis

Utility analysis, as one critical part of the whole method, focuses on costs and benefits of forming a link. In other words, these network individuals will decide whether to form a link according

(a) direct effect (b) effect from 2-path actors

Fig. 1. Graphical representation of network effects.

to the perceived utility change. In order to illustrate clearly how to conduct the mentioned utility analysis, this section presents a general form of utility function, and in particular a specific form as an example. Moreover, a detailed link formation process is also modeled in this section.

3.1. Utility function

Let $\mathbf{G}(N, L)$ represent the graph of the observed network, where N is the set of nodes with the total number of n, and L is the set of observed links. The defined $\mathbf{G}(N, L)$ is an $n \times n$ binary matrix, where its element l_{ij} (hereafter, $i, j = 1, \ldots, n$ and $i \neq j$) is defined as

$$l_{ij} = \begin{cases} 1, & \text{if a directed link exist from individual } j \text{ to } i; \\ 0, & \text{otherwise.} \end{cases}$$

Here, simple graphs that do not contain loops are adopted, and therefore l_{ii} is set to 0 for all the network nodes. Especially, when $l_{ij} = l_{ji}$ holds for all the nodes in N, the graph or say the corresponding social network becomes an undirected one. Note that some social media platforms, such as the Twitter, provide the information of directed link with the consideration that the friendship can be unidirectional; whereas, others think the friendship should be bidirectional and only provide the undirected link information, such as Facebook. The mentioned two kinds of datasets will both be adopted to apply the proposed method, so we here discuss the method in a general graph.

Without loss of generality, a utility function within social networks is designed as a function of the current network state, the individual attributes and some preference parameters. Accordingly, the utility function of individual i can be written as

$$U_i(\mathbf{G}(N, L), \mathbf{C}; \theta), \tag{1}$$

where the population attribute matrix $\mathbf{C}$ is defined by ($\mathbf{C}_1, \ldots, \mathbf{C}_i, \ldots, \mathbf{C}_n$) whose column $\mathbf{C}_i$ denotes the attributes of individual i. Namely, $\mathbf{C}_i$ may consist of age, gender, education, occupation and so on. Besides, $\boldsymbol{\theta}$ is the preference parameter vector.

According to the background of application or the targets of research, the above general utility function should be specified. For example, if we consider that direct friends and 2-path actors affect the formation of friendships (or linkages), the utility function should contain at least two items: direct effects and effects from 2-path actors. Accordingly, a possible specific utility function of individual i in this example could be

$$U_i(\mathbf{G}(N, L), \mathbf{C}; \theta) = \theta_0 \cdot \sum\nolimits_k l_{ik} + \theta_1 \sum\nolimits_k l_{ik} \|\mathbf{C}_i - \mathbf{C}_k\| + \theta_2 \sum\nolimits_{k,m} l_{ik} l_{km} \|\mathbf{C}_i - \mathbf{C}_m\|, \tag{2}$$

where $\|\mathbf{C}_i - \mathbf{C}_k\|$ measures the attribute difference between a pair of individuals. Note that the specific form of $\|\mathbf{C}_i - \mathbf{C}_k\|$ allows flexibility, namely it could be L1-norm, L2-norm or any other distance norms. Furthermore, the item $\sum_k l_{ik}$ and $\sum_{k,m} l_{ik} l_{km}$ adopted in Eq. (2) express the direct effect and the effects from 2-path actors, respectively, and their graphical representations are displayed in Fig. 1 for easy understanding.

In fact, the above specific from of utility function accords with the theory of “Three Degrees of Influence”, which is famous in the field of social networks (see Christakis & Fowler, 2009 for more information). Here, the designed utility function emphasizes the first and the second degree of influence because they can take significant effects on individual actions within social networks. Of course, different application backgrounds and research targets would lead to different specific utility functions; thus, the form of utility function is flexible, guaranteeing that the proposed method could be useful in a broad range of application coverage.

Next, based on the defined utility function, the net utility of individual i by forming a link to individual j, given no link existing between them, is

$$\Delta U_{i \to j}(\mathbf{G}(N, L), \mathbf{C}; \theta) = U_i(\mathbf{G}(N, L_{+ij}), \mathbf{C}; \theta) - U_i(\mathbf{G}(N, L), \mathbf{C}; \theta), \tag{3a}$$

where L_{+ij}: $= L \cup \{(i, j)\}$ denotes the set of observed links adding the directed link from individual i to j. Similarly, the net utility of individual j by forming a link with individual i, is

$$\Delta U_{j \to i}(\mathbf{G}(N, L), \mathbf{C}; \theta) = U_j(\mathbf{G}(N, L_{+ji}), \mathbf{C}; \theta) - U_j(\mathbf{G}(N, L), \mathbf{C}; \theta). \tag{3b}$$

In real life, some unobserved disturbances always exist during the human decision process. For example, it is not surprising that the perception deviation of net utility exists, and emotional fluctuation is common for any person. To reflect these unobserved disturbances, the random error variables ε_i and ε_j are added into the two individuals' net utilities. That is to say, in the eye of individual i, his or her perceived net utility value is not the one shown in Eq. (3a), but

$$\Delta \tilde{U}_{i \to j}(\mathbf{G}(N, L), \mathbf{C}; \theta) = \Delta U_{i \to j}(\mathbf{G}(N, L), \mathbf{C}; \theta) + \varepsilon_i, \tag{4a}$$

and similarly, individual j will make his or her decision based on the perceived net utility

$$\Delta \tilde{U}_{j \to i}(\mathbf{G}(N, L), \mathbf{C}; \theta) = \Delta U_{j \to i}(\mathbf{G}(N, L), \mathbf{C}; \theta) + \varepsilon_j. \tag{4b}$$

Following the specific form of utility function shown in Eq. (2), the net utility values of two individuals are

$$\Delta \tilde{U}_{i \to j}(\mathbf{G}(N, L), \mathbf{C}; \theta) = \theta_0 + \theta_1 \left\|\mathbf{C}_i - \mathbf{C}_j\right\| + \theta_2 \sum_{r \neq i,j} l_{jr} \|\mathbf{C}_i - \mathbf{C}_r\| + \varepsilon_i, \tag{5a}$$

$$\Delta \tilde{U}_{j \to i}(\mathbf{G}(N, L), \mathbf{C}; \theta) = \theta_0 + \theta_1 \left\|\mathbf{C}_i - \mathbf{C}_j\right\| + \theta_2 \sum_{r \neq i,j} l_{ir} \left\|\mathbf{C}_j - \mathbf{C}_r\right\| + \varepsilon_j. \tag{5b}$$

Here, the parameter vector $\boldsymbol{\theta}$ is specified as (θ_0, θ_1, θ_2). The specific forms of utility function and net utility functions will be used again for validation and application in the next parts of this paper. Note that the change of utility is of concern in this paper rather than the basic utility values, which is different with the utility analysis designed in Li, Fang, Bai, and Sheng (2016). In other words, the sign of net utility will be considered rather than its basic value, which is similar to the logistic regression with the binary dependent variable.

3.2. Link formation

Link formation is divided into a meeting process and a decision process. Here, the decision process is first discussed based on the above mentioned utility analysis. In particular, if individuals i and j have an opportunity to meet each other, they will subsequently face the problem of whether to form a link. In this paper, we assume that the individual within a social network only cares about his or her own benefit, indicating that he or she only considers his or her own utility change in the decision-making process. In some of the classic literature, such as Mele (2015), the aforementioned behavior feature is always defined to be *myopic*.

Assumption 1. Individuals are assumed to be myopic.

This assumption guarantees that the decision made by one individual is not affected by the utilities of the other individuals. Although altruistic behavior could exist during humans' decision processes, the above assumption does not destroy the main driving factors. In addition, the assumption lays a foundation for a closed form solution to our theoretical model. Based on Assumption 1, a pair of individuals will establish a link if and only if

$$\Delta\tilde{U}_{i\to j}(\mathbf{G}(N,L),\mathbf{C};\theta)\geq 0 \quad \text{and} \quad \Delta\tilde{U}_{j\to i}(\mathbf{G}(N,L),\mathbf{C};\theta)\geq 0; \tag{6}$$

otherwise, they will not establish the link.

Furthermore, because the perception deviation or the emotional fluctuation of one individual cannot affect that of the others in most cases, ε_i and ε_j can be assumed to be independent. As for all individuals, we have

Assumption 2. The random error variable ε_i (i=1, ..., n) is independent.

The assumption is similar to the assumption of independence often seen in the numerous econometric studies. Meanwhile, the assumption guarantees that when the realization of the network **G**(N, L) and the collected population attribute matrix **C** are given, $\Delta\tilde{U}_{i\to j}(\mathbf{G}(N,L),\mathbf{C};\theta)$ and $\Delta\tilde{U}_{j\to i}(\mathbf{G}(N,L),\mathbf{C};\theta)$ (Hereafter, they are shortened as $\Delta\tilde{U}_{i\to j}$and $\Delta\tilde{U}_{j\to i}$, respectively) are conditionally independent. Namely, it holds that

$$p\left(\Delta\tilde{U}_{i\to j}\geq 0,\, \Delta\tilde{U}_{j\to i}\geq 0|\mathbf{G}(N,L),\mathbf{C}\right)=p\left(\Delta\tilde{U}_{i\to j}\geq 0|\mathbf{G}(N,L),\mathbf{C}\right)\cdot p\left(\Delta\tilde{U}_{j\to i}\geq 0|\mathbf{G}(N,L),\mathbf{C}\right), \tag{7}$$

Moreover, according to Coles, Bawa, Trenner, and Dorazio (2001), the random variable that reflects non-rational factors is often assumed to meet the type-I-extreme-value distribution. We follow the common distribution assumption.

Assumption 3. The random error variable ε_i (i=1, $\cdots$, n) meets the Type-I-extreme-value distribution, *i.i.d.* among individuals.

The adopted type-I-extreme-value distribution (see Guegan & Hassani, 2014 for more information) originates from the discrete choice model in the field of economics (Hausman & Wise, 1978) in which the discrete choice model is often used to describe and explain humans' choices between two discrete alternatives, such as forming a link or not, as discussed in this paper. This assumption guarantees that the probabilities shown in formula (7) can be expressed explicitly as

$$p^l_{i\to j}:=p\left(\Delta\tilde{U}_{i\to j}\geq 0|\mathbf{G}(N,L),\mathbf{C}\right)=\frac{\exp(\Delta U_{i\to j})}{1+\exp(\Delta U_{i\to j})}, \tag{8a}$$

and,

$$p^l_{j\to i}:=p\left(\Delta\tilde{U}_{j\to i}\geq 0|\mathbf{G}(N,L),\mathbf{C}\right)=\frac{\exp(\Delta U_{j\to i})}{1+\exp(\Delta U_{j\to i})}. \tag{8b}$$

Here, $p^l_{i\to j}$ or $p^l_{j\to i}$ denotes the probability of forming a directed link during the decision-making process. With regarding to the undirected case, let p^l_{ij} denote the probability of forming a bidirectional link between individuals i and j. Then, according to the condition in formula (6) and the equation in formula (7), we further have

$$p^l_{ij}=p^l_{i\to j}\cdot p^l_{j\to i}=\frac{\exp(\Delta U_{i\to j})}{1+\exp(\Delta U_{i\to j})}\cdot\frac{\exp(\Delta U_{j\to i})}{1+\exp(\Delta U_{j\to i})}. \tag{9}$$

where the specific forms of $\Delta U_{i\to j}$ and $\Delta U_{j\to i}$ has been defined in Eqs. (5a) and (5b), respectively.

Subsequently, we discuss the meeting process. Here, we consider that the meeting likelihood between a pair of individuals is dependent on the number of their common friends; namely, more common friends always lead to a higher meeting probability (Jin, Girvan, & Newman, 2001). Accordingly, let p^m_{ij} be the meeting probability between individuals i and j, and its mathematical expression would be designed as

$$p^m_{ij}=\frac{\exp\left(b\cdot\sum_{r\neq i,j}l_{ir}l_{jr}\right)}{a+\exp\left(b\cdot\sum_{r\neq i,j}l_{ir}l_{jr}\right)}. \tag{10}$$

Here, a and b are two parameters with restrictions of $a\geq 0$ and $b\geq 0$. Eq. (10) shows that the parameter a reflects the initial meeting likelihood, and the parameter b reflects the effect intensity of the number of common friends. It is not difficult to find that the restriction of two parameters guarantees that $p^m_{ij}\in(0,1]$. Moreover, unlike the link could be directed, the meeting is considered to be undirected in this paper, because meeting should be conducted by both sides rather than by one side.

From the Eqs. (8a), (8b) and (10), we can find that the defined meeting probability between a pair of individuals does not influence their net utilities so that it does not affect the values of $p^l_{i\to j}$ and $p^l_{j\to i}$, and meanwhile the calculated $p^l_{i\to j}$ and $p^l_{j\to i}$ from the decision process are not the factors affecting the meeting probability p^m_{ij}. As a result, the meeting process and the decision process can be regarded to be independent according to this paper's setup. Then, as for the directed case, the whole probability of forming a link from individual i to j (denoted as $p_{i\to j}$) is

$$p_{i\to j}=p^m_{ij}\cdot p^l_{i\to j}, \tag{11a}$$

and similarly, the whole probability from individual j to i (denoted as $p_{j\to i}$) is

$$p_{j\to i}=p^m_{ij}\cdot p^l_{j\to i}. \tag{11b}$$

As for the undirected case, the whole probability of forming a link between individual i and j (denoted as p_{ij}) is

$$p_{ij}=p^m_{ij}\cdot p^l_{ij}. \tag{11c}$$

It is worth to note that the two right forms in Eqs. (11a)–(11c) contain the preference parameter vector $\boldsymbol{\theta}$ as well as two parameters, a and b. For easy expression, we further denote $\boldsymbol{\beta}$=($\boldsymbol{\theta}$, a, b), and once $\boldsymbol{\beta}$ is estimated, the values of $p_{j\to i}$, $p_{j\to i}$ or p_{ij} appearing in Eqs. (11a)–(11c) can be achieved immediately. Thus, the next problem concerns how to estimate $\boldsymbol{\beta}$.

4. Estimation strategy

This section aims to determine how to calibrate the preference parameters embedded in the utility function and the probabilities of potential meeting states. Simply put, the observed network **G**(N, L) and the collected attribute **C** are the inputs, and the estimates of parameters are the output. Considering that the latent meeting states are unobserved, the idea of the EM algorithm would be a proper choice because it is often adapted to estimate parameters when unobserved latent variables or missing data exist. However, the EM algorithm cannot be very effective without the explicit expressions of the posterior probability in the E-step and the objective function in the M-step. Thus, the significant work conducted in this section intends to infer the close forms of the posterior probability and the objective function via Bayesian analysis. Moreover, the lower boundary of the objective function of the proposed EM algorithm is deduced via theoretical analysis, which would explain why the proposed algorithm precedes the logistic regression. In addition, the famous logistic regression can be regarded as a special case of our method.

4.1. Posterior probability and objective function

As stated in the introduction, the traditional EM algorithm requires two alternating steps for iteratively achieving the optimal parameter estimates that maximize the likelihood function of the whole model. The Expectation step (E step, for short), as one of the two alternating steps, should infer the posterior probability of the latent states, which are necessary components of the objective function in the subsequent maximization step of this iteration. Meanwhile, the maximization step (M step, for short) should present the closed form of the objective function, based on the result of the E step in this iteration, and should then complete the goal of achieving the optimal parameter estimates, which constitute one part of the inputs in the next iteration. Without loss of generality, the following statement takes the tth iteration as the example, given that the obtained parameter estimates in the $(t-1)$th iteration.

In the E step, $m_{ij}(t)$ denotes the meeting state between individuals i and j at the tth iteration, where $m_{ij}(t)=1$ indicates that they meet each other in this iteration, and $m_{ij}(t)=0$ indicates that they do not. Then, given $\boldsymbol{\beta}(t-1)$, the observed l_{ij} and $\mathbf{C}$, the posterior probabilities of the latent meeting states can be inferred via the following Lemma 1, where $p^{l}_{ij}(t-1)$ and $p^{m}_{ij}(t-1)$ denote p^{l}_{ij} and p^{m}_{ij} of the $(t-1)$th iteration, respectively.

Lemma 1. *According to different values of $m_{ij}(t)$, l_{ij} and l_{ji}, the posterior probabilities of the latent meeting states can be divided into four cases:*

$$p(m_{ij}(t)=1|l_{ij}=1 \text{ or } l_{ji}=1;\beta(t-1))=1, \tag{12a}$$

$$p(m_{ij}(t)=0|l_{ij}=1 \text{ or } l_{ji}=1;\beta(t-1))=0, \tag{12b}$$

$$\begin{aligned} &p(m_{ij}(t)=1|l_{ij}=0 \text{ and } l_{ji}=0;\\ &\beta(t-1))=\frac{p^{m}_{ij}(t-1)\cdot(1-p^{l}_{ij}(t-1))}{1-p^{m}_{ij}(t-1)\cdot p^{l}_{ij}(t-1)}, \end{aligned} \tag{12c}$$

$$\begin{aligned} &p(m_{ij}(t)=0|l_{ij}=0 \text{ and } l_{ji}=0;\\ &\beta(t-1))=\frac{1-p^{m}_{ij}(t-1)}{1-p^{m}_{ij}(t-1)\cdot p^{l}_{ij}(t-1)}. \end{aligned} \tag{12d}$$

Proof. of Lemma 1 is provided in Appendix A. ■

Now, we explain how to calculate the $p^{l}_{ij}(t-1)$ and $p^{m}_{ij}(t-1)$ appearing in Lemma 1. The $p^{l}_{ij}(t-1)$ can be calculated via two steps. In the first step, the obtained estimates $\boldsymbol{\theta}(t-1)$ are fed into Eqs. (5a) and (5b) to simultaneously obtain the values of $\Delta U_{i\to j}(t-1)$ and $\Delta U_{j\to i}(t-1)$, and in the second step, $p^{l}_{ij}(t-1)$ can be immediately achieved by substituting $\Delta U_{i\to j}(t-1)$ and $\Delta U_{j\to i}(t-1)$ into Eq. (9). On the other hand, $p^{m}_{ij}(t-1)$ can be calculated based on the estimated $a(t-1)$and $b(t-1)$ via Eq. (10). Besides, Lemma 1 shares the same form no matter in the directed case or in the undirected one.

In the M step, the objective function $l_t(\boldsymbol{\beta})$ of the tth iteration has the following form, according to the classic literature, such as Ng, Krishnan, and McLachlan (2012), that is,

$$l_t(\beta)=\sum_{i\neq j} l_{t,ij}(\beta), \tag{13a}$$

such as

$$l_{t,ij}(\beta)=\sum_{l_{ij}=0}^{1}\sum_{m_{ij}(t)=0}^{1} p(m_{ij}(t)|l_{ij},\beta(t-1))$$

$$\cdot\log\frac{p(m_{ij}(t),l_{ij};\beta)}{p(m_{ij}(t)|l_{ij},\beta(t-1))}, \tag{13b}$$

Especially, when $m_{ij}(t)=0$ and $l_{ij}=1$, we set

$$p(m_{ij}(t)|l_{ij},\beta(t-1))\cdot\log\frac{p(m_{ij}(t),l_{ij};\beta)}{p(m_{ij}(t)|l_{ij},\beta(t-1))}=0. \tag{13c}$$

Generally speaking, an explicit form of objective function would facilitate solving its maximal value. To this end, the explicit forms of $p(m_{ij}(t), l_{ij}; \boldsymbol{\beta})$ are deduced in the directed case and in the undirected case, respectively. Lemma 2 shows the results.

Lemma 2. *As for the directed case, the explicit forms of $p(m_{ij}(t), l_{ij}; \boldsymbol{\beta})$, according to the different values of $m_{ij}(t)$ and l_{ij}, include*

$$p(m_{ij}(t)=0,l_{ij}=0;\beta)=1-p^{m}_{ij}, \tag{14a}$$

$$p(m_{ij}(t)=0,l_{ij}=1;\beta)=0, \tag{14b}$$

$$p(m_{ij}(t)=1,l_{ij}=0;\beta)=p^{m}_{ij}\cdot(1-p^{l}_{j\to i}), \tag{14c}$$

$$p(m_{ij}(t)=1,l_{ij}=1;\beta)=p^{m}_{ij}\cdot p^{l}_{j\to i}. \tag{14d}$$

As for the undirected case, its forms include

$$p(m_{ij}(t)=0,l_{ij}=0;\beta)=1-p^{m}_{ij}, \tag{14e}$$

$$p(m_{ij}(t)=0,l_{ij}=1;\beta)=0, \tag{14f}$$

$$p(m_{ij}(t)=1,l_{ij}=0;\beta)=p^{m}_{ij}\cdot(1-p^{l}_{ij}), \tag{14g}$$

$$p(m_{ij}(t)=1,l_{ij}=1;\beta)=p^{m}_{ij}\cdot p^{l}_{ij}. \tag{14h}$$

Proof. of Lemma 2 is provided in Appendix A. ■

Note that $p^{l}_{j\to i}$ (or p^{l}_{ij}) appearing in Lemma 2 has been defined in Eq. (8b) (or Eq. (9)), so it is the function of $\boldsymbol{\theta}$, considering the expressions of net utility functions shown in Eqs. (5a) and (5b). Until now, the general form of likelihood function shown in Eqs. (13a)–(13c) can be explicitly expressed in different cases based on Lemmas 1 and 2.

Subsequently, the parameter estimates could be achieved by maximizing the explicit objective function. Particularly, the M step of the tth iteration faces the following optimal problem:

$$\beta(t):=\arg\max_{\beta} l_t(\beta). \tag{15}$$

Here, the achieved optimal solution $\boldsymbol{\beta}(t)$ is the results of the tth iteration and meanwhile the input of the $(t+1)$th iteration. By repeating the above mentioned E step and M step until convergence, the obtained optimal estimates $\boldsymbol{\beta}^{*}$ will maximize the likelihood function of the whole model (see Dempster et al., 1977).

4.2. Algorithm procedures

This section will firstly display the algorithm procedure of parameter estimate based on the above lemmas and deduced formulas, and then present the detailed process of link prediction given the estimated parameters. With regarding to the first procedure, the optimal problem displayed in formula (13a)–(13c) can be coped with by the classical gradient descent algorithm when the closed form of the objective function is achieved. Besides, it is possible to compute the mentioned gradient descent in a multicore setting, and thus we here design a parallel computation process for parameter estimation so as to save computation time. In details, the following algorithm procedure realizes the distributed computation

Table 1
Algorithm procedure of parameter estimation.

```
1. Input the network adjacent matrix $\mathbf{G}$, the individual attribute matrix $\mathbf{C}$, the core number $k$, the convergence rate $\eta$ and the maximal iteration number $T$.
2. Initialize parameter vector $\boldsymbol{\beta}_0$, $t=0$, and partition all the individuals into the cores on average.
3. While $\eta(t) > \eta$ or $t < T$ do
4.     $t=t+1$ and $\tau(t) = 1/\sqrt{t}$
5.     For all $ii \in \{1,\ldots, k\}$ parallel do
6.         For all pairs of individuals $(p,q)$ in core $ii$ do
7.             Compute the gradient $\nabla_{\boldsymbol{\beta}} l_{t,\, pq}(\boldsymbol{\beta}(t-1))$
8.         End For
9.         Compute the gradient value local on core $ii$, namely $\nabla^{ii}_{\beta} l_t(\beta(t-1)) = \sum_{p\neq q} \nabla_{\beta} l_{t,pq}(\beta(t-1))$
10.    End For
11.    Compute $\beta(t) = \beta(t-1) - \tau(t) \cdot \sum_{ii=1}^{k} \nabla^{ii}_{\beta} l_t(\beta(t-1))$ and $\eta(t) = \|\beta(t) - \beta(t-1)\| / \|\beta(t-1)\|$
12. End While
13. Return the optimal estimates $\boldsymbol{\beta}^*$.
```

Note: Step 7 is the E step, and Step 11 is the M step. The initial values of parameters $(\theta_0, \theta_1, \theta_2)$ within the vector $\boldsymbol{\beta}$ originate from the parameter estimates of logistic regression (LR), and the reasons for which can be found in Section 4.3.

Table 2
Algorithm procedure of link prediction.

1. **Input** the parameter estimates $\boldsymbol{\beta}^*$, the number of common friends $\sum_{r\neq i,j} l_{ir} l_{jr}$, the attributes of two individuals $\mathbf{C}_i$ and $\mathbf{C}_j$, and the attributes of their friends $\mathbf{C}_r$.
2. **Infer** the pair's meeting probability p^m_{ij} according to Eq. (10).
3. **Calculate** the probability $p^l_{i\to j}$ and $p^l_{j\to i}$ according to Eqs. (8a), (8b), (5a) and (5b) in the directed case, or the probability p^l_{ij} according to Eqs. (9), (5a) and (5b) in the undirected case.
4. **Obtain** the whole probability of directed link formation $p_{i\to j}$ and $p_{j\to i}$ via Eqs. (11a) and (11b) in the directed case, or the whole probability p_{ij} via Eqs. (11c) in the directed case.
5. **Output** the results of link prediction according to the achieved whole probabilities under different cases.

of gradients on part of data loaded locally on each core, and then conduct a global update step via aggregation of the gradients on each core. To make it clear, the whole algorithm procedure is listed in Table 1.

Next, once these parameter estimates are achieved via the procedure in Table 1, the likelihood of forming a link between any pair of individuals can be achieved immediately by considering the opportunity of meeting and the decision result via utility analysis. Without loss of generality, also taking individual i and j as the example, Table 2 summarizes the procedure of link prediction for this pair.

4.3. Algorithm property

The first property is to uncover the relationship between the proposed method (UbLP for short, hereafter) and logistic regression (LR for short, hereafter), especially in the field of link prediction within social networks. First, LR is a popular method when the dependent variable is categorical. However, in contrast with UbLP, LR does not consider the latent meeting states; in other words, the parameter estimation of LR is based on the hidden assumption that all pairs of network individuals have met each other. That is to say, LR does not infer the posterior probability of latent meeting states since LR deems that $m_{ij}=1$ and $p^m_{ij} = 1$ holds in all cases. Meanwhile, in the frame of UbLP, if all pairs of individuals have met, the log-likelihood is simplified as

$$\sum_{i\neq j} \log p(m_{ij}(t) = 1, l_{ij}; \theta, a = 0, b). \tag{16}$$

As for the directed case, Eqs. (14a)–(14d) guarantee that formula (16) can be further expressed as

$$\sum_{i\neq j} \left[l_{ij} \log p^l_{j\to i} + (1 - l_{ij}) \log(1 - p^l_{j\to i}) \right] \tag{17a}$$

As for the undirected case, Eqs. (14e)–(14h) guarantee that formula (16) can be further expressed as

$$\sum_{i\neq j} \left[l_{ij} \log p^l_{ij} + (1 - l_{ij}) \log(1 - p^l_{ij}) \right]. \tag{17b}$$

No matter in which case, Eq. (17a) or (17b) is just the log-likelihood function of LR. As a result, LR can be regarded as a special case of UbLP in which the meeting process is completely ignored. The above reasoning leads to the results summarized in Property 1.

Property 1. LR is a special case of UbLP. In details, when all pairs of individuals are assumed to have met in UbLP, UbLP is simplified to be LR.

Second, because LR is a popular method for measuring the relationships between the categorical dependent variable and the selected independent variables, it would be meaningful to compare LR with UbLP in terms of estimation accuracy. By way of solid theoretical analysis, the result is summarized in the following Property 2.

Property 2. The log-likelihood function of UbLP is lower bounded as that of LR, so that UbLP precede LR in terms of the optimal value of likelihood function.

Proof. of Property 2 is provided in Appendix B. ■

By taking the optimal value of likelihood function as the criterion for estimation accuracy, Property 2 indicates that UbLP occurs prior to LR, which would be a theoretical foundation for explaining why the performance of link prediction via UbLP would generally precede that of LR. In contrast, LR could be seen as a representative method that ignores the unobservable meeting processes. Accordingly, the comparison between LR and UbLP can uncover, to some extent, the significance of considering the meeting states. Additionally, beyond the qualitative result given in Property 2, the following numerical experiments illustrate how much the prediction performance has improved from LR to UbLP, as well as to the other selected benchmarks.

5. Numerical experiments

Two types of numerical experiments are designed to validate the above-proposed method based on simulated social networks.

The first experiment aims to examine the performance of the proposed method and to compare it with the other classic baselines, introduced mainly in Liben-Nowell and Kleinberg (2007). The second one plans to uncover why the concerned latent meeting process is important. Before displaying the results, the shared parts of two experiments are first introduced: the appointment of individual attributes and the generation of artificial networks that consists of the meeting process and the decision process.

Random appointment of individual attributes. Each individual is randomly assigned a number from the interval [0,1] as his or her attribute. To clarify the simulation, the attribute vector reduces to one dimension, although it would include several dimensions, such as gender, occupation, education, and background, in real life. However, the simplification of the attribute vector does not greatly affect the main result because these attribute values are only used to measure the difference between individuals.

Progressive generation of artificial networks. The generated network starts from a null network with 200 nodes. Here, an artificial network must undergo several evolution steps until it comes into a state of stable distribution. In this paper, the meeting process and the decision process are two necessary ingredients at each step of network formation.

Meeting process. Let $p_{ij}^{m}(k)$ denote the meeting probability between individuals i and j at the kth step of network generation. According to Eq. (10), it has the form

$$p_{ij}^{m}(k) = \frac{\exp\left(b \cdot \sum_{r \neq i,j} l_{ir}(k) l_{jr}(k)\right)}{a + \exp\left(b \cdot \sum_{r \neq i,j} l_{ir}(k) l_{jr}(k)\right)}, \tag{18}$$

where a and b are two controllable parameters, and their different values reflect different communication efficiencies in various social media platforms because some platforms allow for high meeting opportunity, while others do not. Note that some pairs have no opportunity to meet in the kth step so that their $p_{ij}^{m}(k)$ will be zero in this step.

Decision process. The undirected friendship is adopted here, so that the generated social networks are symmetrical. To this end, based on the specific net utility functions in Eqs. (5a) and (5b), the values of $\Delta U_{i \to j}$ and $\Delta U_{j \to i}$ can be obtained, given the preference parameters $\boldsymbol{\theta}$. As a result, the probability of linkage formation between individual i and j can further be calculated by formula (9). Then, two cases could occur: if a pair of individuals has not formed a link, the pair will form a link with the probability p_{ij}^{l}; otherwise, the probability of linkage disconnect will be $1 - p_{ij}^{l}$. Accordingly, the change of linkage between any pair could be simulated.

Undergoing sufficient evolving steps as the above stated two processes, the simulated network will come to a state of stable distribution because the evolution process is a Markov process, and accordingly, a stable distribution can be achieved (Pin & Rogers, 2016). Then, we can sample a specific number (here, the number is 500) of networks from the series of generated artificial networks, which is regarded as running a long MCMC sample from a stable contribution. To make it clear, all of the required parameters in the generation process are summarized in Table 3.

5.1. Performance

Taking each of the artificial networks as the input, parameter estimates and the result of link prediction can be achieved in succession via the algorithms in Tables 1 and 2. Then, by comparing the estimated links with the true ones within the artificial networks, we can evaluate the performance of the proposed method (UbLP for short, hereafter) in the context of our designed experiments. Generally speaking, two common metrics are often adopted to measure the performance of link prediction: AUC and precision. AUC means the area under the curve of the receiver operating characteristic curve (Bradley, 1997), and precision indicates the proportion of the true positives among all of the positive results (Davis & Goadrich, 2006). Furthermore, AUC could be regarded as a global index because it can represent a complete evaluation of the predicted links, whereas precision could be seen as a local index because it can be evaluated at a given cut-off rank or indicate that it considers only the topmost results.

Table 3
Parameter values in generation processes.

Individual	Attribute	Random number from [0,1]	
	Preference parameters (utility function)	θ_0	1.50
		θ_1	−1.00
		θ_2	−0.05
Network	Size	200 nodes	
	Sample number	500	
Controllable parameters	a	4, 2, 1, 0.25 or 0	
	b	10	

Note: After approximately 300 evolving steps starting from a null network, the artificial network is sampled at every 20 evolving steps until 500 networks are obtained.

Table 4
Performance with the varying parameter a.

a	AUC	Precision		
		1%	5%	10%
4	0.75	0.17	0.16	0.18
2	0.69	0.24	0.22	0.21
1	0.56	0.48	0.44	0.42
0.25	0.76	0.66	0.65	0.67
0	0.91	0.88	0.86	0.84

Note: The results are achieved by averaging 500 artificial networks. The initial values of preference parameters come from the parameter estimates of LR.

First, the proposed method is tested on the varying values of parameter a. Recalling Eq. (18), the controllable parameter a determines the initial likelihood of meeting. Therefore, different values of parameter a reflect different situations of social networks in the real world; in other words, it is easier for individuals in some social networks to “meet” than in other ones. The implication of this test is to uncover the capacity of the proposed method on a large range of social networks with diverse meeting likelihoods. Table 4 shows the results.

The AUCs in Table 4 demonstrate that the results are better than random guessing in all cases. Another discovery is that the AUC first decreases and then increases as parameter a rises, partly because, in our opinion, the entropy of the system increases first and then decreases with the increase of parameter a. Moreover, precision also performs better than random guessing whose baseline is dependent on the given parameter a. For example, when $a=4$, the initial meeting probability p_{ij}^{m} is 0.20 on average, indicating that the probability of forming a link (or say baseline) is approximately 0.10, given that the randomly guessed p_{ij}^{l} equals 0.5. In this example, if a method allows us to provide a better performance than 0.10 for the set of truly formed links, the tested method could be regarded as better than random guessing. In addition, the other finding is that the performance of the proposed method is dependent on the controllable parameter a, and the precision metrics decrease gradually as parameter a increases. The series of discoveries are summarized in Finding 1.

Finding 1. UbLP occurs prior to random guessing in terms of AUC and precision, although the controllable parameter a varies in different cases. Furthermore, the controllable parameter a will affect

Table 5
The benchmarks for comparison.

Method	Abbreviation	Note
Utility-based link prediction	UbLP	Proposed method
Logistic regression	LR	Special benchmark
Jaccard index	JI	Unsupervised benchmark
Adamic–Adar index	AA	Unsupervised benchmark
Resource allocation index	RA	Unsupervised benchmark
Support vector machine	SVM	Supervised benchmark
Hierarchical structure model	HSM	Supervised benchmark
Random forest	RF	Supervised benchmark
Adaptive boosting (Naive Bayes as the weak leaner)	AB	Supervised benchmark

Note: See Liben-Nowell and Kleinberg (2007) for detailed descriptions of JI, AA, SR and RA, and they are all unsupervised benchmarks. Because LR is a special case of UbLP, it is regarded as a special benchmark. Moreover, SVM (see Bleakley et al., 2007 for detailed information), HSM (see Clauset et al., 2008 for detailed information), RF (see Guns & Rousseau, 2014 for detailed information) and AB (see Wang et al., 2015; Freund & Schapire, 1995 for detailed information) are all supervised methods. Besides, it is worth to note that the adopted weak learner of AB is Naive Bayes in this paper.

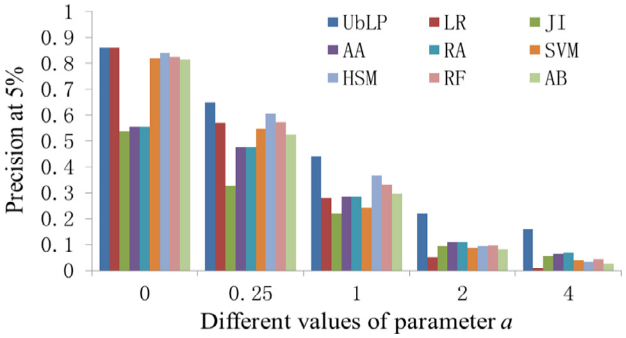

Fig. 2. Precision of UbLP and the selected benchmarks.

the performance of UbLP, and interestingly the trends of AUC and Precision are different with varying values of parameter *a*.

On the other hand, to uncover further the predictive performance of UbLP, some selected methods including unsupervised and supervised ones are benchmarked in this paper. Table 5 summarizes the benchmark methods and Fig. 2 displays the comparison results based on the above mentioned performance metrics. The charts in Fig. 2 reveal that precision at 5% of all the methods drops dramatically with the parameter *a* rising. However, UbLP reports the best performance compared to the other benchmarks in all the cases. Furthermore, when the parameter *a* is small, the supervised benchmarks perform better than the selected unsupervised ones, but when *a* becomes bigger, the supervised ones perform worse compared to the unsupervised ones. In addition, UbLP shares the same value of precision at 5% with LR when $a = 0$, illustrating Property 1 in Section 4.3. The following Finding 2 summarizes these results.

Finding 2. UbLP performs better than the selected unsupervised and supervised benchmarks in our settings. Additionally, with the rise of parameter *a*, the performance of the selected supervised methods becomes worse.

Finding 2 implies that, in the context of our simulations, the unsupervised method would be a good choice for link prediction in such a social network where individuals have many more opportunities to meet (recalling the implication of the parameter *a* in Eq. (10)), and regardless of the types of social networks, the proposed UbLP would be a better choice in most cases.

Table 6
Parameter estimate bias of UbLP and LP.

a	LR			UbLP		
	θ_0	θ_1	θ_2	θ_0	θ_1	θ_2
4	−1.41	−0.50	−0.56	0.95	−0.99	−0.09
	(0.08)	(0.05)	(0.07)	(0.09)	(0.07)	(0.10)
2	0.00	−0.87	−0.68	0.99	−1.04	−0.11
	(0.04)	(0.05)	(0.06)	(0.07)	(0.06)	(0.06)
1	1.22	−1.07	−0.15	1.34	−1.10	−0.06
	(0.04)	(0.05)	(0.06)	(0.05)	(0.05)	(0.02)
0.25	1.47	−1.15	−0.11	1.49	−1.06	−0.05
	(0.04)	(0.05)	(0.06)	(0.04)	(0.03)	(0.02)
0.00	1.51	−1.05	−0.05	1.51	−1.05	−0.05
	(0.04)	(0.03)	(0.01)	(0.04)	(0.03)	(0.01)

Note: The true values of the three parameters are 1.50, −1.00 and −0.05 from left to right, according to Table 3.

5.2. Latent variable

This section aims to demonstrate why the considered latent variable should not be ignored in the problem of link prediction. To achieve this aim, we compare UbLP with LR which is a typical method without considering the latent variable. Because the two methods share the same process of parameter estimation, we choose the closeness between parameter estimates and their true values as the criteria. By running the two methods on the 500 artificial networks, the average values of the estimates and their corresponding standard deviations are listed in Table 6.

Table 6 demonstrates that UbLP is superior to LR regarding the bias of parameter estimates when $a > 0$; however, the two methods are identical when $a=0$ (namely $p_{ij}^m = 1$), illustrating Property 2 that LR is a special case of UbLP. In addition, the results of LR become increasingly worse when parameter *a* becomes increasingly larger so that the sign of parameter θ_0 is even wrong when $a \geq 2$, whereas the signs of the parameters are correct via UbLP in all the cases. All of the discoveries are summarized in the Finding 3.

Finding 3. UbLP performs better than LR in terms of parameter estimate bias in our settings, which illustrates the importance of considering the latent variables, namely the unobservable meeting states in the problem of link prediction.

6. Applications

Three different real-world networks are selected for the application study in this section. The first one is the collaboration network of scientists posting preprints on condensed matter at www.arxiv.org (Newman, 2001). The dataset contains two undirected collaboration networks: one includes all preprints posted between 1 January 1995 and 30 June 2003, and the other is between 1 January 1995 and 31 March 2005, indicating that the second one completely contains the first one. The difference between the two networks reflects the network evolution over time. The earlier network is used to train our model and to obtain the parameter estimates; then, the later network is used to test the proposed method. Hereafter, we use the collaboration network to denote this dataset, and the collaboration network is undirected in nature.

The second dataset, available at http://snap.stanford.edu/data/egonets-Facebook.html, is a Facebook friend network from a survey organized by McAuley (2012) and Leskovec (2014) who used the *Facebook app*. The collected friendship is undirected in Facebook, which means the friend network is symmetrical. The whole dataset includes 10 ego-networks, and each ego-network dataset contains individual IDs (nodes), their friendships (undirected edges) and their attributes, such as gender, educational background, occupation, etc. Taking one as the training set, when two ego-networks are selected, we can use the other to test the prediction performance.

The third one is a Twitter friend network whose data is available at http://snap.stanford.edu/data/egonets-twitter.html. The friendship collected in Twitter is directed, which is different with Facebook. The whole dataset includes 973 ego-networks, and each ego-network dataset also contains individual IDs (nodes), their friendships (directed edges) and their attributes similar to the dataset from Facebook. Following the idea of test used in Facebook friend network, we randomly choose one ego-network as the training set, and then test the performance of link prediction in another randomly chosen ego-network.

In all, the selected three datasets are consistent with the two examples mentioned at the beginning of this paper. Furthermore, the elaborately selected three datasets have different features, which would soundly illustrate the applicability of the proposed method in our opinion. In details, the first dataset enables to display the application of the proposed method in one network with different periods. However, the remaining two datasets contain several different ego-networks collected almost during the same period so that they can test the method in different networks. Besides, the remaining two datasets are also different: one is directed and the other is undirected, although they are both friend network.

6.1. Collaboration network

The collaboration network is from preprints on the condensed matter archive at www.arxiv.org (www-personal.umich.edu/~mejn/netdata). The first part of the dataset contains 31,163 authors with 120,029 co-author relationships during the period from 1 January 1995 to 30 June 2003, and the second part contains more authors and many more co-author relationships because of the additional period from 1 July 2003 to 31 March 2005. Intuitively, the degree distribution of the first part is shown in Fig. 3, where the 50 authors who have published at least 100 papers are paid special attention to. Then, Fig. 4 displays the co-author networks of the selected 50 authors in two periods. Specifically, the blue lines show the co-author links established during the period from 1 January 1995 to 30 June, 2003, and the red line shows the added co-author linkages from 1 July, 2003, to 31 March, 2005. Then, the network of the first period is used for training and that of the second period is used for testing.

Recalling the designed utility function shown in formula (1), the attribute $\mathbf{C}_i$ should be calculated in the first step. In this application, we use the keywords as the attribute of the author. That is to say, each author is assigned a vector consisting of the keywords that appeared in his or her posted preprints, as well as the frequency of each word via the text mining technique in R software (version R-x64-3.2.1). For example, the keywords information of Authors No. 1 and No. 2 are shown in Tables 7 and 8, respectively.

Table 7
Keyword information of Author No. 1.

Keywords	Occurrence	Frequency
Hole-doped cuprates	**4**	**0.4**
Pseudogap	**3**	**0.3**
Bond-density wave	**2**	**0.2**
Statistical physics	1	0.1

Note: "Occurrence" indicates the times that the corresponding keyword appears in all of the preprints posted by the author.

Table 8
Keyword information of Author No. 2.

Keywords	Occurrence	Frequency
High temperature	2	0.2
Hole-doped cuprates	**2**	**0.2**
Pseudogap	**2**	**0.2**
Bond-density wave	**2**	**0.2**
Charge density wave	1	0.1

Note: Bold type indicates the shared keywords.

Table 9
AUC of UbLP and its benchmarks.

Method	AUC	Method	AUC
UbLP	0.930	SVM	0.826
LR	0.612	HSM	0.835
JI	0.723	RF	0.871
AA	0.807	AB	0.897
RA	0.881	Diagonal	0.500

Based on the mined keywords information, the attribute distance between two individuals is defined as

$$\left\|\mathbf{C}_i - \mathbf{C}_j\right\| := 1 - \frac{\sum_{sw} f_i(sw) + f_j(sw)}{2}, \tag{19}$$

where "*sw*" indicates the keywords shared by the two individuals, and $f_i(sw)$ (or $f_j(sw)$) denotes the occurrence frequency of the relevant keywords for individual i (or j). According to the information in Tables 7 and 8, the attribute distance of the two authors is 0.25.

UbLP is conducted following the procedures for an undirected network in Tables 1 and 2. First of all, the initial values of parameters (θ_0, θ_1, θ_2) are reached via LR and the result is (4.44, −0.82, −0.10). Then, the other two parameters a and b are initially set as 0.5 and 1.0, respectively. Next, the algorithm in Table 1 guarantees that the optimal estimate of parameter vector $\boldsymbol{\beta}$ is (5.88, −2.74, −0.97, 0.21, 0.92). Then, the optimal parameter estimate and the algorithm in Table 2 achieve the predicted probability of link formation between each pair of 50 co-authors in the second period. As a result, Fig. 5 displays the ROC curve of UbLP as well as its benchmarks, and Table 9 shows their AUCs.

The plots in Fig. 5 clearly show that UbLP can provide much more accurate prediction results than its benchmarks, while all of the methods perform better than random guessing, indicating that these methods are effective for this real dataset. In addition, UbLP occurs prior to LR, supporting Property 2. Besides, the four supervised methods listed in the right part of Table 9 provide similar performances, whereas the three unsupervised methods perform differently. The above finding would confirm that the feature of networks will dramatically influence the performance of unsupervised methods. In addition, LR performs worst partly because it will result biased estimates if too many pairs have no meeting

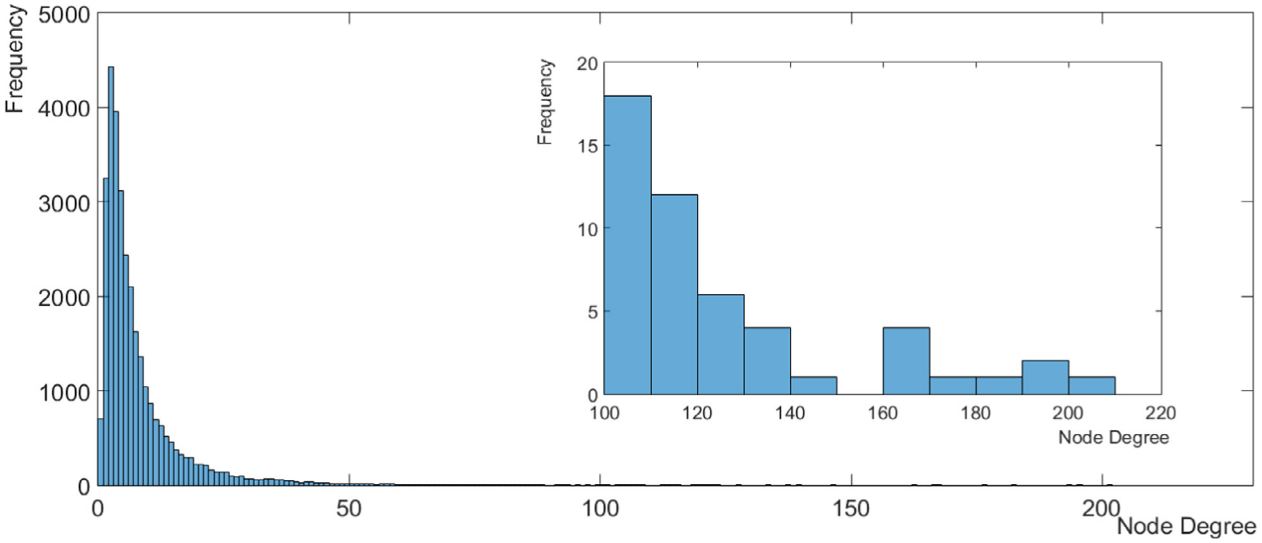

Fig. 3. Frequency histogram of node degree.

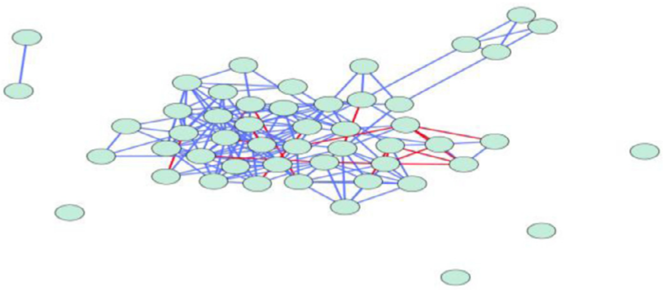

Fig. 4. Co-author network in two periods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

opportunities to form links. Thus, it will be very likely that LR cannot perform well in a sparse network, for example, the co-author network adopted here.

6.2. Facebook friend network

The selected Facebook friend network belongs to the undirected ego-network, for which an ego-network consists of an ego node and the nodes to which the ego is directly connected, as well as the links among these nodes. It is interesting to note that the chosen ego-network is a good target for analysis: on the one hand, it is impossible to analyze the whole friend network in Facebook due to its large size, and it is also difficult to find an isolated network with a proper size in Facebook; on the other hand, it is not appropriate to sample a network randomly from Facebook for link prediction because the sample will omit many important nodes and linkages as well as much structure information.

Two ego-networks are selected for this application: the first one for training and the second one for testing, where the ego-network No. 1 containing 59 nodes and 146 links and ego-network No. 2 containing 66 nodes and 270 links. Unlike the collaboration network adopted in the last application, individual attributes are included in the dataset, which facilitate calculating the attribute difference. Also different from the definition of the attribute difference in the last application, its definition here is

$$\|\mathbf{C}_i - \mathbf{C}_k\| := \sum\nolimits_m |c_{im} - c_{km}|, \tag{20}$$

where c_{im} and c_{jm} are the mth items of attribute vectors $\mathbf{C}_i$ and $\mathbf{C}_j$, respectively.

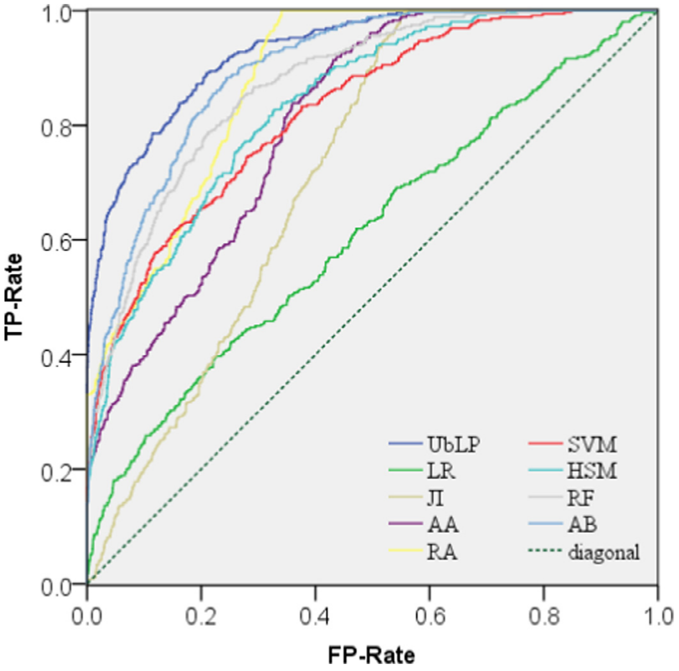

Fig. 5. ROC curves of UbLP and its benchmarks.

Table 10
AUCs of UbLP and its benchmarks.

Method	AUC	Method	AUC
UbLP	**0.905**	SVM	0.811
LR	0.620	HSM	0.817
JI	0.824	RF	0.834
AA	0.641	AB	0.839
RA	0.667	Diagonal	0.500

Following the process of UbLP, the achieved initial values of (θ_0, θ_1, θ_2) are estimated as (−1.53, 0.02, −0.07) via LR in the first step, and meanwhile parameters a and b are initially set as 0.5 and 1.0. Then, the optimal estimate of $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{a}, \hat{b})$ is (5.54, −2.01, −0.85, 3.50, 2.10) via the algorithm in Table 1. As a result, the ROC curves of UbLP and its benchmarks are achieved in Fig. 6, and their AUCs are listed in Table 10.

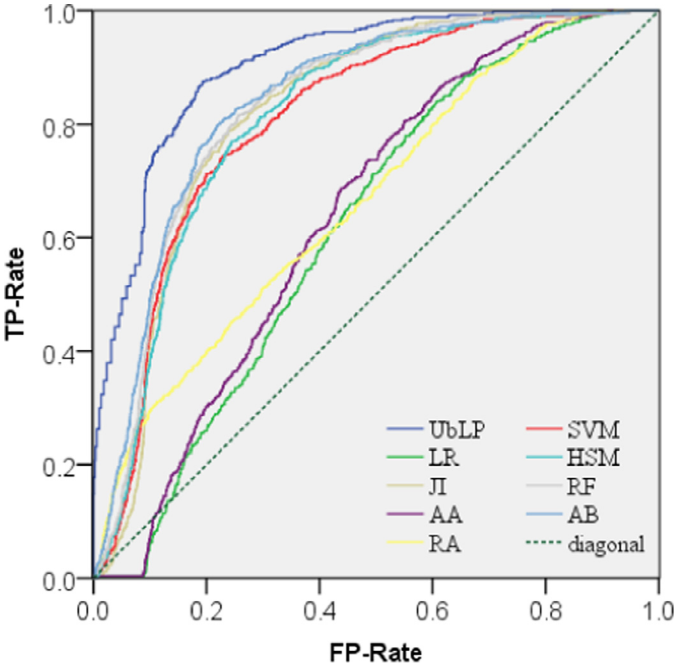

Fig. 6. ROC curves of UbLP and its benchmarks.

Similar to the results in Table 9, UbLP reports the best AUC value compared to the selected benchmarks. The reason that UbLP precedes LR has been explained in the first application. Here, we pay attention to another finding that may be interesting. Namely, JI reports a good result that is close to the best one. As we have mentioned in Section 2.2, different unsupervised methods can reflect different features of network structures, so it is not surprising that one unsupervised method will perform better in one kind of network than in other kinds. However, why JI performs well in the selected Facebook friend network and even better than two of the supervised benchmarks? On the one hand, JI emphasizes the effect of common friends (see Liben-Nowell & Kleinberg, 2007) so that it will perform well in such a network that common friends take great effect on link formation. On the other hand, the optimal estimate of parameter a and b is 3.50 and 2.10, which means that the initial meeting likelihood is tiny, but the number of common friends greatly influence the meeting opportunities, by recalling the parameter implications shown in Eq. (10). Thus, the estimation results of the meeting process implies that the JI would be a good unsupervised method for this dataset.

6.3. Twitter friend network

Similar to the Facebook friend network, the dataset collected in Twitter also consists of ego-networks, which would be proper for the application as explained in Section 6.2. However, two things are different between the Facebook friend network and the Twitter friend network. First, compared to the undirected friend network in Facebook, the Twitter friend network is directed, which means the friendship is not necessary to be bidirectional in Twitter. Second, different ego-networks in Twitter are surveyed different attributes; for example, gender appears in Twitter ego-network 1 as the attribute, but it may not appear in Twitter ego-network 2. Thus, contrast to Facebook, a new definition of attribute distance in Twitter should be provided in order to balance the different numbers of attributes contained in different ego-networks. Thus, we define

$$\|\mathbf{C}_i - \mathbf{C}_k\| := \frac{\sum_m |c_{im} - c_{km}|}{M_{ego-network\ No.}}, \tag{21}$$

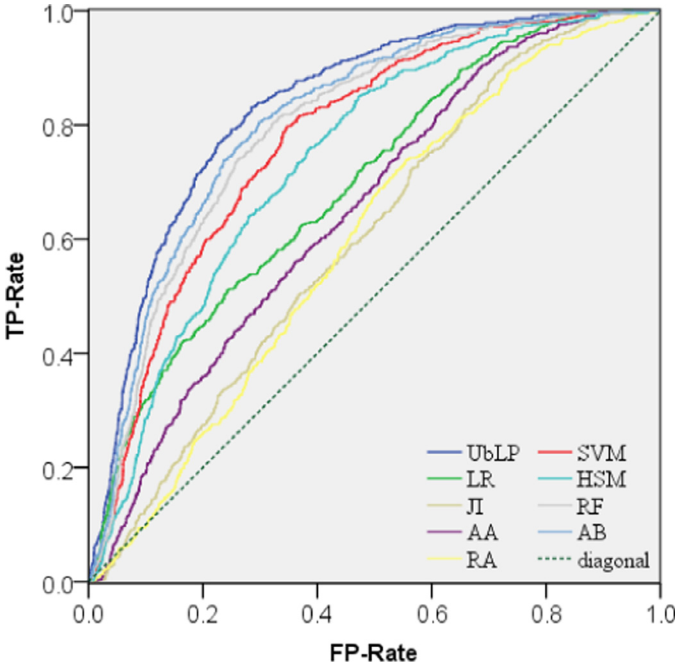

Fig. 7. ROC curves of UbLP and its benchmarks.

Table 11
AUCs of UbLP and its benchmarks.

Method	AUC	Method	AUC
UbLP	0.835	SVM	0.776
LR	0.697	HSM	0.740
JI	0.602	RF	0.797
AA	0.650	AB	0.812
RA	0.595	Diagonal	0.500

where $M_{ego-network\ No.}$ denotes the total number of attributes contained in the ego-network *No.* For example, if ego-network 7 is chosen, $M_{ego-network\ 7}$ is the total number of attributes in the ego-network 7. Here, the other symbols are identical with Eq. (20).

Rather than a pair of deterministic ego-networks selected for training and test in Facebook, the pair of ego-networks are randomly selected in Twitter. By averaging the results in these random selected pairs of ego-networks, the results of link prediction can be achieved immediately. In details, the part of algorithm for directed networks are adopted (recalling Tables 1 and 2) and the pairs of ego-networks are randomly selected 25 times from the 973 ego-networks in the dataset, then the ROC curves of UbLP and its benchmarks are achieved in Fig. 7**,** and their AUCs are listed in Table 11.

Parallel to the former two applications on the undirected network, the Twitter friend network witnessed the applicability of UbLP on directed network. Here, the three unsupervised methods, namely JI, AA and RA, cannot distinguish the directed and undirected networks so that they all take the directed friendship as the undirected one. The results in Fig. 7 and Table 11 again confirm that UbLP performs robustly and well in the problem of link prediction.

7. Conclusion and future work

This paper proposes a novel approach for coping with the problem of link prediction by dividing the process of link formation into the unobservable meeting process and the inner decision-making process. On the one hand, Bayesian inference is used for the meeting process to estimate the latent meeting states and the relevant parameters. On the other hand, utility analysis is introduced into the decision-making process for uncovering the inner behavior factors that influence the link formation. Subsequently, EM algorithm is adopted and further developed to estimate the parameters embedded in the two mentioned processes. Thus, the proposed approach can be regarded as a supervised machine learning method with latent variables. Although numerous traditional methods have achieved a certain degree of success in forecasting the missing links in social networks, the proposed method features the elaborate analysis of link formation process and also theoretically precedes logistic regression so that it can become a new member and also a contributor in the large family of link prediction methods.

The main findings, through proofs, simulation validation and application analysis, are summarized as follows: (1) the proposed method can almost correctly estimate the parameters embedded in the designed utility function and the meeting process; (2) the proposed method yields satisfying link prediction results in these synthetic networks; (3) it would benefit improving the performance of link prediction that dividing the process of link formation into the unobservable meeting process and the inner decision-making process; (4) logistic regression, as a special case of the proposed method, provides the lower boundary of the likelihood function; and (5) the applications on the three real datasets illustrate that our approach could yield more satisfying and more robust link prediction results than the selected benchmarks.

Future work could consider the development of our method towards the analysis of dynamic networks, with which it will face the new problem of how to estimate parameters through a series of networks based on multi- stage observations. Apart from our design of utility function, another further work would be to design numerous other forms of utility functions under different application backgrounds and for diverse research goals. Last but not least, any potential comparisons and applications related to the proposed method would be welcome.

Acknowledgments

The authors greatly thank three anonymous referees and the corresponding editor for their useful comments and suggestions. Of course, all remaining errors are our own. This study is funded by National Natural Science Foundation of China (No. 71501034), China Postdoctoral Science Foundation (No. 2016M590230), and Social Science Fund of Liaoning Province (No. L15CGL014).

Appendix A. Proof of Lemmas 1 and 2

Lemma 2 is proved first. As for the undirected case, by following the denotations in the main text, it is clear that,

$$p(m_{ij}(t) = 0, l_{ij} = 1; \beta) = 0, \tag{A1}$$

because a link cannot be formed without meeting; in other words, it is impossible that $m_{ij}(t)=0$ and $l_{ij}=1$ occur at the same time.

Next, because the meeting process does not affect the result of the decision-making process, namely the independence assumption shown in Eq. (11a)–(11c), it holds immediately that

$$p(m_{ij}(t) = 1, l_{ij} = 1; \beta) = p^m_{ij} \cdot p^l_{ij}, \tag{A2}$$

$$\text{and} \quad p(m_{ij}(t) = 1, l_{ij} = 0; \beta) = p^m_{ij} \cdot (1 - p^l_{ij}). \tag{A3}$$

In contrast, the sum of the probabilities of all of the cases equals 1; thus,

$$p(m_{ij}(t) = 0, l_{ij} = 0; \beta) = 1 - p^m_{ij}. \tag{A4}$$

Similarly, the results in the directed case can also be achieved by following the above reasoning process. In all, Lemma 2 has been proved.

Subsequently, we prove Lemma 1. According to Bayes' theorem, it holds that

$$p(m_{ij}(t)|l_{ij}; \beta(t-1)) = \frac{p(m_{ij}(t), l_{ij}|\beta(t-1))}{\sum_{l_{ij}=0}^{1} p(m_{ij}(t), l_{ij}|\beta(t-1))}, \tag{A5}$$

in all cases. Then, it can be further divided into four cases according to different values of $m_{ij}(t)$ and l_{ij}. Based on Lemma 2, the results in Lemma 1 can be obtained immediately in both undirected and directed networks. □

Appendix B. Proof of Property 2

Firstly, the log-likelihood function of UbLP at the tth iteration is defined as

$$ll_t(\beta(t)) = \sum_{i \neq j} \sum_{l_{ij}=0}^{1} \log p(l_{ij}; \beta(t)), \tag{B1}$$

where $p(m_{ij}(t), l_{ij}; \boldsymbol{\beta}(t))$ follows the definition given in Eqs. (14a)–(14d). Next, regardless of the value l_{ij}, it holds that

$$\sum_{m_{ij}(t)=0}^{1} p(m_{ij}(t), \beta|l_{ij}) = 1, \tag{B2}$$

according to Eqs. (12a)–(12d) shown in Lemma 1. Then, Jensen inequality grants that

$$\begin{aligned}&\sum_{m_{ij}(t)=0}^{1} p(m_{ij}(t)|l_{ij}, \beta(t-1)) \cdot \log \frac{p(m_{ij}(t), l_{ij}; \beta)}{p(m_{ij}(t)|l_{ij}, \beta(t-1))} \\ &\le \sum_{m_{ij}(t)=0}^{1} p(m_{ij}(t)|l_{ij}, \beta(t-1)) \cdot \log \frac{p(l_{ij}; \beta)}{p(m_{ij}(t)|l_{ij}, \beta(t-1))} \\ &\le \log p(l_{ij}; \beta)\end{aligned}$$

and thus we have

$$ll_t(\beta(t)) \ge l_t(\beta(t)), \tag{B3}$$

where the definition of $l_t(\boldsymbol{\beta}(t))$ has been given in Eqs. (13a)–(13c). Furthermore, Dempster et al. (1977) has proved the monotonicity of the EM algorithm; namely, we have

$$l_t(\beta(t)) \ge l_{t-1}(\beta(t-1)), \tag{B4}$$

and then it holds that

$$ll(\beta*) \ge ll_t(\beta(t)) \ge l_t(\beta(t)) \ge l_0(\beta(0)), \tag{B5}$$

where $\boldsymbol{\beta}*$ is the optimal estimates of UbLP, and $\boldsymbol{\beta}(t)$ is the optimal solution at the t-th iteration (see formula (15)). Regarding the initial parameter values $\boldsymbol{\beta}(0)$, we can set the parameter a as 0 so that $l_0(\boldsymbol{\beta}(0))$ becomes the log-likelihood function of LR according to formula (16). Subsequently, let $\boldsymbol{\beta}^*(0)$ maximize l_0; in other words, $l_0(\boldsymbol{\beta}^*(0))$ is the maximal value of the log-likelihood function of LR. As a result, Inequality (B5) grants that

$$ll(\beta*) \ge l_0(\beta * (0)), \tag{B6}$$

which indicates that the maximal value of the log- likelihood function of UbLP is no less than that of LR. Thus, Property 2 holds. □

References

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

Ballings, M., & Van den Poel, D. (2015). CRM in social media: Predicting increases in Facebook usage frequency. *European Journal of Operational Research, 244*(1), 248–260.

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

Barbieri, N., Bonchi, F., & Manco, G. (2014). Who to follow and why: Link prediction with explanation. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1266–1275).

Ben-Hur, A., & Noble, W. S. (2005). Kernel methods for predicting proteinprotein interactions. *Bioinformatics, 21*(Suppl. 1), i38–i46.

Bleakley, K., Biau, G., & Vert, J.-P. (2007). Supervised reconstruction of biological networks with local models. *Bioinformatics, 23*, i57–i65.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Christakis, N. A., & Fowler, J. H. (2009). *Connected: The surprising power of our social networks and how they shape our lives*. Little: Brown.

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

Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*: 208. London: Springer.

Davis, J., & Goadrich, M. (June 2006). The relationship between Precision–Recall and ROC curves. In *Proceedings of the 23rd international conference on machine learning* (pp. 233–240).

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1), 1–38.

Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research, 241*(2), 449–457.

Elkabani, I., & Khachfeh, R. A. A. (2015). Homophily-based link prediction in the facebook online social network: A rough sets approach. *Journal of Intelligent Systems, 24*(4), 491–503.

Fang, X., Hu, P. J. H., Li, Z., & Tsai, W. (2013). Predicting adoption probabilities in social networks. *Information Systems Research, 24*(1), 128–145.

Freund, Y., & Schapire, R. E. (March 1995). A desicion-theoretic generalization of on–line learning and an application to boosting. *Proceedings of the 1995 European conference on computational learning theory* (pp. 23–37). Springer Berlin Heidelberg.

Guegan, D., & Hassani, B. (2014). A mathematical resurgence of risk management: An extreme modeling of expert opinions. *Frontiers in Finance and Economics, 11*(1), 25–45.

Guns, R., & Rousseau, R. (2014). Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics, 101*(2), 1461–1473.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *Unsupervised learning* (pp. 485–585). New York: Springer.

Hausman, J. A., & Wise, D. A. (1978). A conditional probit model for qualitative choice: Discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica, 46*(2), 403–426.

Hellmann, T., & Staudigl, M. (2014). Evolution of social networks. *European Journal of Operational Research, 234*(3), 583–596.

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

Jin, E. M., Girvan, M., & Newman, M. E. (2001). Structure of growing social networks. *Physical Review E, 64*(4), 046132.

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

Kashima, H., Kato, T., Yamanishi, Y., Sugiyama, M., & Tsuda, K. (2009). Link propagation: A fast semi-supervised learning algorithm for link prediction. In *Proceedings of SIAM international conference on data mining* (pp. 1099–1110).

Lee, C., Pham, M., Jeong, M. K., Kim, D., Lin, D. K., & Chavalitwongse, W. A. (2015). A network structural approach to the link prediction problem. *INFORMS Journal on Computing, 27*(2), 249–267.

Leskovec, J., & McAuley, J. J. (2012). Learning to discover social circles in ego networks. *In Neural Information Processing Systems* (pp. 539–547).

Li, X., & Chen, H. (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems, 54*(2), 880–890.

Li, Y., Wu, C., Wang, X., & Luo, P. (2014). A network-based and multi-parameter model for finding influential authors. *Journal of Informetrics, 8*, 791–799.

Li, Z., Fang, X., Bai, X., & Sheng, O. R. L. (2016). Utility-based link recommendation for online social networks. *Management Science*. http://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2446.

Liu, Y., Li, Q., Tang, X., Ma, N., & Tian, R. (2014). Superedge prediction: What opinions will be mined based on an opinion supernetwork model? *Decision Support Systems, 64*, 118–129.

Liben-Nowell, D., & Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology, 58*(7), 1019–1031.

Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 243–252).

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

McAuley, J., & Leskovec, J. (2014). Discovering social circles in ego networks. *ACM Transactions on Knowledge Discovery from Data (TKDD), 8*(1), 4.

Mele, A. (2015). A structural model of segregation in social networks. *Available at SSRN 2294957*.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences United States of America, 98*(2), 404–409.

Ng, S. K., Krishnan, T., & McLachlan, G. J. (2012). The EM algorithm. *Handbook of computational statistics* (pp. 139–172). Berlin Heidelberg: Springer.

Nguyen, C. H., & Mamitsuka, H. (2012). Latent feature kernels for link prediction on sparse graphs. *IEEE Transactions on Neural Networks and Learning Systems, 23*(11), 1793–1804.

Pin, P., & Rogers, B. (2016). Stochastic network formation and homophily. In Y. Bramoullé, B. Rogers, & A. Galeotti (Eds.), *Oxford handbook on the economics of networks* (pp. 138–166). Oxford University Press.

Raju, N. S., Burke, M. J., & Normand, J. (1990). A new approach for utility analysis. *Journal of Applied Psychology, 75*(1), 3–12.

Richard, E., Gaïffas, S., & Vayatis, N. (2012). Link prediction in graphs with autoregressive features. *Journal of Machine Learning Research, 15*(1), 565–593.

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the Association for Information Science & Technology, 63*(1), 78–85.

Wang, G. N., Gao, H., Chen, L., Mensah, D. N., & Fu, Y. (2015). Predicting positive and negative relationships in large social networks. *PloS One, 10*(6), e0129530.