J Inf Process Syst, Vol.13, No.4, pp.914~927, August 2017
https://doi.org/10.3745/JIPS.04.0040

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Weighted Local Naive Bayes Link Prediction

JieHua Wu*, GuoJi Zhang**, YaZhou Ren***, XiaYan Zhang**, and Qiao Yang**

Abstract

Weighted network link prediction is a challenge issue in complex network analysis. Unsupervised methods based on local structure are widely used to handle the predictive task. However, the results are still far from satisfied as major literatures neglect two important points: common neighbors produce different influence on potential links; weighted values associated with links in local structure are also different. In this paper, we adapt an effective link prediction model—local naive Bayes model into a weighted scenario to address this issue. Correspondingly, we propose a weighted local naive Bayes (WLNB) probabilistic link prediction framework. The main contribution here is that a weighted cluster coefficient has been incorporated, allowing our model to inference the weighted contribution in the predicting stage. In addition, WLNB can extensively be applied to several classic similarity metrics. We evaluate WLNB on different kinds of real-world weighted datasets. Experimental results show that our proposed approach performs better (by AUC and Prec) than several alternative methods for link prediction in weighted complex networks.

Keywords

Complex Network, Link Prediction, Naive Bayes Model, Weighted Network

1. Introduction

Link [1] is a bridge which connects nodes in complex network. The rich information of links plays a key role in complex network analysis and benefits a wide range of applications. Take some real-world networks for example. In social networks, people engage in social interactions by commenting, liking, mentioning and following each other. They share thoughts, beliefs, opinions, news, and even check-ins through their social relationships [2]. In protein-protein interaction networks [3], protein molecular is node, and mining interactions between nodes is helpful to reveal the protein function and determine biological mechanism. Moreover, in bibliographic networks, the heterogeneous type of collaborative relations is a useful tool to detect the scientist community (circle) [4] and model topic diffusion [5].

Therefore, the study of link information is of crucial importance to complex network analysis. One core task is usually called link prediction [6], which leverages the given network structure to infer the potential links (missing links). To achieve this goal, each potential link is assigned a similarity score—

Manuscript received August 18, 2016; first revision February 10, 2017; accepted March 4, 2017.
Corresponding Author: JieHua Wu (wu.jiehua@mail.scut.edu.cn)
* School of Computer Science and Engineering, Guangdong Polytechnic of Industry and Commerce, Guangzhou, China (wu.jiehua@mail.scut.edu.cn)
** School of Computer Science and Engineering, South China University of Technology, Guangzhou, China (magjzh@scut.edu.cn, scutzhangxiayan@163.com, 564053731@qq.com)
*** Big Data Research Center, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China (yazhou.ren@uestc.edu.cn)

the more similar the pair of nodes are, the more likely they are linked. The challenge is how to construct a robust similarity function. One effective way is to build local topological measures, such as the Common Neighbors (CN) [6], Adamic and Adar (AA) [7], the Jaccard metric [8], etc.

However, in realistic situation, there exists some networks of which the links strength are unequally treated. These networks are called the weighted networks [9]. In the context of the aforementioned bibliographic networks, the link weight indicates the number of published papers. It is obvious that two scientists are more likely to be in the same research circle if they share a higher "weight". Thus, some other literatures [10,11] tried to solve such problem and proposed corresponding weighted formulation of aforementioned similarity measures. Such solution only aimed to find the weights associated with links but ignored the difference among their common neighbors. In other words, each common neighbor is treated equally to the linked likelihood—the node weight is missing. Fig. 1 gives a brief illustration and presents the problem addressed in this paper. The left figure shows the example in which the potential link (red node) with a local structure (common neighbors, green nodes) needed to be predicted by CN measure. In CN measure, the number of common neighbor is the similarity function score. In the middle figure, the shadow associated with green nodes denotes their different properties (e.g., degree, etc.). If we use CN measure, all the three toy sub-figures have the equal score 3, indicating that they have the same link formation possibility. Naturally, it is not conforming to our basic intuition because different weighted information of green nodes is not considered.

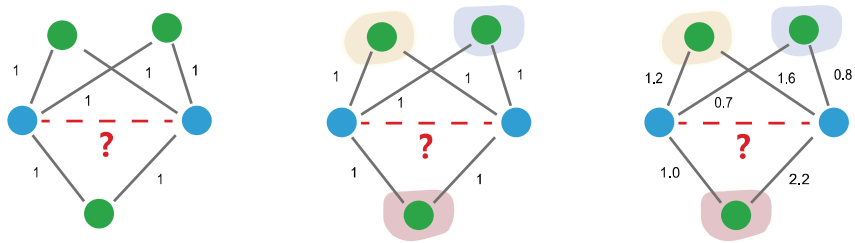


Fig. 1. Illustrative toy example of link prediction in different types of networks.

Consequently, a local naive Bayes (LNB) model [12] has recently been introduced as a probabilistic model for link prediction. LNB assumes the link probability is conditionally depended on the weight of local common neighbors. In LNB, two potential links with the same scale of common neighbors may have different link likelihoods. However, as the right figure of Fig. 1 shows, LNB has the drawback of not taking the link weighted information into account. Consequently, the question we investigate in this research is whether the LNB algorithm can be adapted to a weighted network scenario to overcome the limitation as discussed above. Specifically, we precisely define the weighted network link prediction problem and build a new connection between such problem and LNB model, i.e., a weighted local naive Bayes (WLNB) model has been proposed. WLNB not only takes advantage of LNB model to measure different roles of common neighbors (in node aspect), but also is capable of incorporating diverse link weighted information into a local similarity framework (in link aspect). Our major contributions are summarized as follows. (a) We formulate a novel problem of LNB learning method for weighted network link prediction. (b) We incorporate the weighted cluster coefficient into the proposed model. (c) We extend WLNB to other similarity baselines. (d) We empirically evaluate the effectiveness and efficiency of WLNB on several weighted real-world datasets. To the best of our knowledge, we are the

first to study how to extend LNB theory into a weighted network link prediction problem.

We first introduce the related work in Section 2. Then, Section 3 defines the link prediction task as an unsupervised learning problem and summarizes the notation used through this paper. In Section 4, we start with the local naive Bayes link prediction model, and then propose WLNB to capture the weighted information and extend it to other similarity-based methods. After that, we conduct experiments in Section 5. We finally conclude this paper in Section 6.

2. Related Work

Link prediction is one of the core techniques of complex network analysis, which has been widely used in many applications [5,13,14]. Related literatures on link prediction can be classified into two categories: unsupervised link prediction [15] and supervised link prediction [16]. General unsupervised link prediction approaches are based on similarity or proximity between nodes. Liben-Nowell and Kleinberg [6] did a great ground breaking survey. They introduced various similarity measures, such as Common Neighbors (CN), Katz, etc. Afterwards, Zhou et al. [2,12,17] improved such baselines by developing some new predictors including a local naive Bayes link prediction model. Lichtenwalter et al. [18] also designed a flow based model which has significant improvement over those baselines. Most of the above works focused on estimating similarity in unweighted networks. However, they did not consider that weights are always associated with links in real complex networks [19].

Further studies have been done to capture the weighted information for prediction. Yang et al. [20] improved the existing similarity methods and made them more suitable for weighted networks. Wind and Morup [4] proposed a Poisson-based model to infer the missing links. Murata and Moriyasu [11] introduced an improved method for predicting links based on weighted proximity measures of social networks. The major difference between our proposed WLNB and the aforementioned methods is that we consider common neighbors' contribution with weighted information. On the other hand, De Sa and Prudencio [21] investigated the relevance of using weight to improve supervised link prediction. However, this algorithm mainly focuses on classification problem, while our effort is trying to build an unsupervised framework.

3. Preliminary

In this section, we first give several necessary definitions and then present a formal definition of the problem. We use G=(V, E, W) to denote the structure of complex network, where V is a set of nodes (vertex) and $E \subseteq V \times V$ is a set of links (edges) between nodes. Each link $e_{ij}=<v_i,v_j> \in E$ represents a relationship between node v_i and v_j. Let W be a weighted matrix, where each element w_{v_i,v_j} represents the weight value associated with link e_{ij}. We use $N(v_i)$ to indicate the set of neighboring nodes of v_i, then the common neighbor set of v_i and v_j is defined as $CN(v_i,v_j)$. Let $T(v_i)$ denote the number of the triangles of node v_i, which is defined as the element count of $\{< v_j, v_k > \in E | < v_i, v_k > \in E, < v_i, v_j > \in E\}$. Specifically, in real complex networks, links could be self-links, undirected or directed. We focus on undirected networks in this paper.

Predicting tasks. We adopt the ratio r to partition G into a training G_t and a predicting set G_p. In such process, we randomly choose r = 80% and leave the other 20% as missing links for test. The problem of link prediction is to build a probabilistic model f based on G_t and to decide whether potential links exist or not:

$$f(G_t, G_p) \rightarrow \{0,1\}$$

In the predicting stage, the model function f will output a score to demonstrate the link formulation probability by leveraging the information from the training network G_t. We then infer the binary existence {0,1} of relationships in the target network G_p with the ground truth.

4. The Proposed Model

4.1 The Basic Model

Before going to detail about how to model the weighted structure information, we first give a brief review of LNB link prediction model [12], which has been proved to be effective in handling link prediction task. LNB assumes the link formulation probability is conditionally depending on its local structural property—common neighbors. Then, the posterior probability between nodes v_i and v_j can be respectively given by:

$$P\left(e_{ij}\middle|CN(v_i,v_j)\right)=\frac{P(e_{ij})}{P(CN(v_i,v_j))}\cdot P\left(CN(v_i,v_j)\middle|e_{ij}\right) \tag{1}$$

$$P\left(\bar{e}_{ij}\middle|CN(v_i,v_j)\right)=\frac{P(\bar{e}_{ij})}{P(CN(v_i,v_j))}\cdot P\left(CN(v_i,v_j)\middle|\bar{e}_{ij}\right) \tag{2}$$

where $P(e_{ij})$ and $P(\bar{e}_{ij})$ denote the constant network property, $P\left(CN(v_i,v_j)\middle|e_{ij}\right)$ and $P\left(CN(v_i,v_j)\middle|\bar{e}_{ij}\right)$ control the $CN(v_i,v_j)$ contribution, respectively. A popular choice is to substantiate Eq. (1) and Eq. (2). Then, the similarity score $s^{LNB}_{v_i,v_j}$ between v_i and v_j becomes:

$$s^{LNB}_{v_i,v_j}=\frac{P(e_{ij})}{P(\bar{e}_{ij})}\times\frac{P\left(CN(v_i,v_j)\middle|\bar{e}_{ij}\right)}{P\left(CN(v_i,v_j)\middle|e_{ij}\right)}=\underbrace{\frac{P(e_{ij})}{P(\bar{e}_{ij})}\prod_{\omega\in CN(v_i,v_j)}\frac{P(e_{ij})}{P(\bar{e}_{ij})}}_{\text{constant value}}\times\underbrace{\prod_{\omega\in CN(v_i,v_j)}\frac{P\left(e_{ij}\middle|\omega\right)}{P\left(\bar{e}_{ij}\middle|\omega\right)}}_{\text{contribution value}} \tag{3}$$

where ω denotes the common neighbors of v_i and v_j. The first part of Eq. (3) is a constant value and the factor in the bracket is the common neighbors' contribution (weight) to the potential link $< v_i, v_j >$.

4.2 Weighted Local Link Prediction Model

In Eq. (3), the conditional probability $P\left(e_{ij}\middle|\omega\right)$ means that given a common neighbor ω, the link formulation possibility among its neighbors, which is equal to the definition of local cluster coefficient c_ω:

$$P(e_{ij}|\omega)=c_{\omega}=\frac{2T(\omega)}{|N(\omega)|\times(|N(\omega)|-1)} \quad (4)$$

In this equation, the clustering coefficient is defined without taking into consideration the link weights as $T(\omega)$ and $N(\omega)$ only care about the correlation among nodes. The loss of weight information could lead to obscureness of the latent information in local structure. To illustrate such disadvantage, let us consider the right figure in Fig. 1 again, where each link is assigned a weight value. It is clearly that the potential links formulation probability is determined by its local structure, including not only node influence but also weighted contributions of links. If we assume links are unweighted, some potential and useful information (weighted value) will be missed.

Thus, the performance of LNB in unweighted network may be different from the weighted one as the link weighted in the local structure should be considered. Therefore, a WLNB link prediction model is proposed to overcome the limitation of LNB. As such model is based on CNs, we also call it WLNBCN. The key idea of WLNBCN is to adapt a weighted cluster coefficient (WCC) [19] into the LNB model. The task now turns to extend c_{ω} into a weighted scene.

First, we define node strength ns_{v_i} which measures the strength of nodes v_i in terms of the total weight of their connections:

$$ns_{v_i} = \sum_{v_j \in N(v_i)} w_{v_i,v_j} \quad (5)$$

where v_i is the neighbors of v_j. If the weighted network is considered as unweighted network, ns_{v_i} turns to the degree of v_i, then the WCC can be defined as:

$$c_{\omega}^{w}=\frac{1}{ns_{\omega}(k_{\omega}-1)}\sum_{(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{2}a_{v_i,v_j}\,a_{v_i,\omega}a_{\omega,v_j} \quad (6)$$

In such equation, ω is the common neighbor of node v_i and v_j and k_{ω} is the degree of ω, a_{v_i,v_j} is binary symbol, where 1 means there exists a link between v_i, v_j and 0 represents the opposite. $\sum_{(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{2}a_{v_i,v_j}\,a_{v_i,\omega}a_{\omega,v_j}$ is the triangle number in such local structure formulated by the potential links and its common neighbors. The normalization factor $ns_{\omega}(k_{\omega}-1)$ accounts for the weight of each edge times the maximum possible number of triangles it may participate, and it ensures that $0\leq c_{\omega}^{w}\leq 1$ [19].

This is a measure of the local cohesiveness that takes into account the importance of the clustered structure on the basis of weight information actually found on the local triples [19]. Indeed, c_{ω}^{w} is counting for each triple formed in the neighborhood of the node v_i. In this way, we are not just considering the number of closed triangles in the neighborhood of a node but also their total relative weight with respect to the node strength.

As $\frac{P(e_{ij})}{P(\bar{e_{ij}})}$ is the ratio r between the number of connection links $|E|$ and disconnection links scale $|V|(|V|-1)/2-E$, the left part $\frac{P(e_{ij})}{P(\bar{e_{ij}})}\prod_{\omega\in CN(v_i,v_j)}\frac{P(e_{ij})}{P(\bar{e_{ij}})}$ of Eq. (3) becomes a constant value $|CN(v_i,v_j)|r$,

Combing Eq. (3) and Eq. (4), the right part $\prod_{\omega\in CN(v_i,v_j)}\frac{P(e_{ij}|\omega)}{P(\bar{e}_{ij}|\omega)}$ can be transformed to $\sum_{\omega\in CN(v_i,v_j)}\log\frac{c_\omega}{1-c_\omega}$. Then we substitute c_ω^w with c_ω and find that directly realizing Eq. (3) is difficult, we normalize the right part through a logarithmic function that maps it to a value between 0 and 1. Then the similarity in WLNBCN model can be calculated as:

$$s_{v_i,v_j}^{WLNBCN}=|CN(v_i,v_j)|\times logr+\sum_{\omega\in CN(v_i,v_j)}\log\frac{c_\omega^w}{1-c_\omega^w} \tag{7}$$

where the right part is the weighted contribution of common neighbors and the outcome of the whole process is the similarity score between node v_i and v_j.

4.3 Model Extension

To further prove the effectiveness of the proposed WLNB model, we try to extend it to other classic similarity metrics, such as AA (Adamic and Adar) [7] and RA (Resource Allocation) [15]. Thus, the corresponding WLNB forms of AA and RA metrics respectively are:

$$s_{v_i,v_j}^{WLNBAA}=\sum_{\omega\in CN(v_i,v_j)}\frac{logr}{\log|N(\omega)|}+\sum_{\omega\in CN(v_i,v_j)}\frac{\log\frac{c_\omega^w}{1-c_\omega^w}}{\log|N(\omega)|} \tag{8}$$

$$s_{v_i,v_j}^{WLNBRA}=\sum_{\omega\in CN(v_i,v_j)}\frac{logr}{N(\omega)}+\sum_{\omega\in CN(v_i,v_j)}\frac{\log\frac{c_\omega^w}{1-c_\omega^w}}{|N(\omega)|} \tag{9}$$

5. Experiments

In this section, we conduct experiments to evaluate the robustness and effectiveness of the proposed WLNB. We first build up the experiment setting and introduce several weighted complex networks. Then we compare the performance of WLNB with other baseline methods.

5.1 Evaluation Settings

In a weighted network G, we randomly sample r% as the training set and the remaining $1-r\%$ as test set. The default value of r is set to be 80. Then the random sampling is carried out 10 times independently and the average results are reported. Two widely used evaluation metrics, i.e., precision (Prec@N) and area under curve (AUC) are adopted to evaluate the link prediction performance. Specifically, Prec@N is defined as the ratio between the ground truths with candidate set scale N, the default value N is set to be 1000. AUC is equally to $\widehat{K}+0.5\overline{K}/K$. In K times of independent comparisons between potential links and non-existing links, the former has $\widehat{K}$ times higher similarity and $\overline{K}$ times equal. We compare our model with the following alternative baselines.

- The weighted format of some classic baseline methods [22].

$WCN(v_i,v_j)=\sum_{\omega\in CN(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{2}$, $WAA(v_i,v_j)=\sum_{\omega\in CN(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{2\log(1+ns_\omega)}$,

$WRA(v_i,v_j)=\sum_{\omega\in CN(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{ns_\omega}$

- Other quantitative baselines [4].

  $WJaccard(v_i,v_j)=\sum_{\omega\in CN(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{ns_{v_i}+ns_{v_j}-w_{v_i,v_j}}$ $WSalton(v_i,v_j)=\sum_{\omega\in CN(v_i,v_j)}\frac{w_{v_i,\omega}+w_{\omega,v_j}}{\sqrt{ns_{v_i}\times ns_{v_j}}}$

- The unweighted format of LNB models (which are sometimes called LNBs): LNBCN, LNBAA, LNBRA.

  WLNBCN, WLNBAA and WLNBRA are our proposed metrics and WLNBs denotes the summarization of these three metrics.

5.2 Datasets

We conduct experiments on five classic real-world weighted network datasets in this paper.

- astro-ph [23]. It contains the collaboration network of scientists posting preprints on the astrophysics archive at www.arxiv.org, 1995–1999, as compiled by M. Newman. The network is weighted, with weights assigned as described in the original papers.
- cond-mat [23]. It contains a collaboration network of scientists posting preprints on the condensed matter archive at www.arxiv.org. This version is based on preprints posted to the archive between January 1, 1995 and June 30, 2003.
- hep-th [23]. It contains the collaboration network of scientists posting preprints on the high-energy theory archive at www.arxiv.org, 1995–1999. The network is weighted, with weights assigned also as described in the original papers.
- citation [24]. It is a geographical view of collaboration network at the city level, where the nodes are cities and weighted undirected links indicate the presence and frequency of collaborations between scholars of different cities.
- facebook [25]. The Facebook-like Forum Network was attained from the same online community as the online social network. In this network, a weight can be assigned to the ties based on the number of messages or characters that a user posted to a topic.

The statistics of the datasets are shown in Table 1.

Table 1. Statistics of datasets

Dataset	Nodes	Links	Average degree	Average weighted
astro-ph	7171	56547	7.6529	2 .8005
cond-mat	5638	20912	1.8841	3.2632
hep-th	8255	15751	35.878	1.8334
citation	226	28869	286.6814	27443
facebook	897	142760	317.5362	37.4516

Table 2. The AUC and Prec@N performance of different methods

	WCN	WAA	WRA	WSalton	WJaccard	LNBCN	LNBAA	LNBRA	WLNBCN	WLNBAA	WLNBRA
AUC											
astro-ph	0.971	0.973	0.973	0.969	0.969	0.972	0.971	0.972	0.972	0.977	**0.978**
cond-mat	0.928	**0.929**	0.926	0.924	0.926	0.923	0.927	0.924	0.927	0.928	0.928
hep-th	0.906	0.913	0.912	0.911	0.911	0.907	0.904	0.912	0.908	0.909	**0.913**
citation	0.893	0.897	0.926	0.366	0.541	0.142	0.142	0.142	0.996	**0.997**	0.959
facebook	0.811	0.835	0.832	0.458	0.908	0.981	0.982	0.992	0.974	0.989	**0.998**
Prec											
astro-ph	0.793	0.888	0.952	0.447	0.619	0.952	0.963	0.991	0.947	0.997	**0.998**
cond-mat	0.289	0.448	0.664	0.365	0.453	0.666	0.803	0.757	0.672	**0.812**	0.777
hep-th	0.242	0.319	0.426	0.168	0.213	0.517	0.656	0.626	0.521	**0.798**	0.698
citation	0.863	0.866	0.788	0.569	0.699	0.101	0.101	0.092	0.896	0.904	**0.912**
facebook	0.899	0.912	0.896	0.924	0.894	0.943	0.935	0.941	0.937	0.948	**0.948**

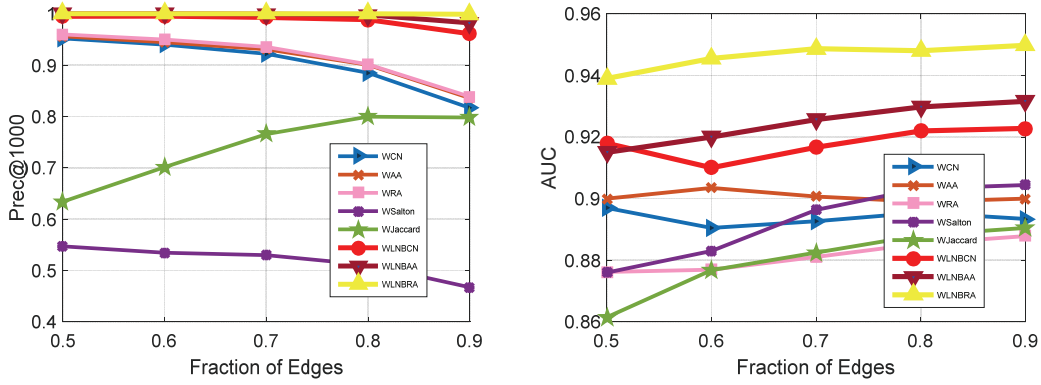
5.3 Performance Comparison

Table 2 summarizes the AUC and Prec@1000 (we also refer to Prec for convenience) of all the methods in the prediction task. The higher the AUC and Prec are, the better performance the model achieves. In each row, the best results are highlighted in boldface. We can clearly observe that the best performance is constantly achieved by WLNBs on both AUC and Prec. If we overlook the model difference, WLNBCN, WLNBAA, WLNBRA on average achieve 0.051, 0.043, 0.039 higher AUC and 0.249, 0.182, 0.122 higher precision comparing to the corresponding WCN, WAA, WRA, respectively. Such better performance can be ascribed to the fact that leveraging the role of common neighbors from the weighted network. On the other hand, we observe that the LNBs perform relative poor in two metrics by 0.231, 0.1752, 0.181 higher AUC and 0.1694, 0.168, 0.162, respectively. It shows that LNBs might not work perfectly in cases where link weight exists. WJaccard and WSalton perform almost the worst among all the approaches. The introduction of union of common neighbor is a possible reason for this performance. Another phenomenon is that the performance of WLNBs is correlated with the model specification as WLNBAA and WLNBRA perform slightly better than WLNBCN in almost all cases. This is mainly because that the degree information is considered in WLNBAA and WLNBRA. These results support the fact that our proposed model can take advantage of the common neighbors influence and link weight for the prediction task.

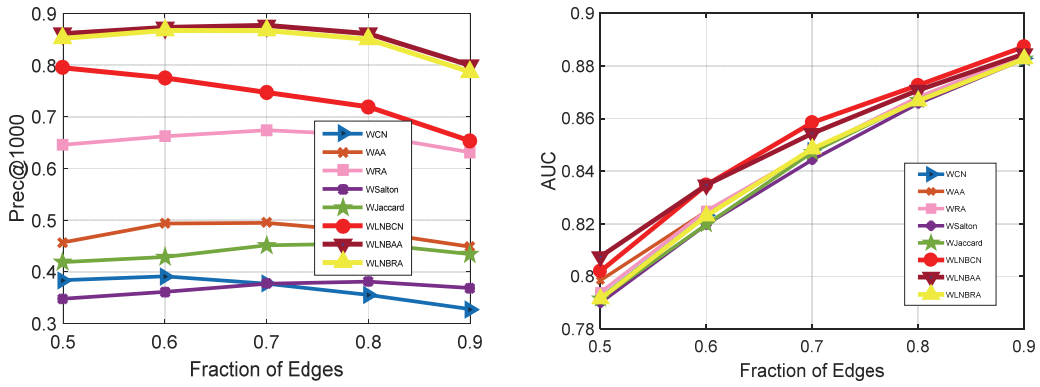
To evaluate the statistical significance, we perform a t-test between WLNBs and the best competitive method of citation network in the predictive task. The result is 2.5365e320, which suggests the significance of our proposed model. We also have the similar observations on other datasets.

We now evaluate the variation of performance of the framework with different proportions of training data. Such process will help assess the robustness of WLNBs. We vary the proportion r from 50% to 80% in step 5% for learning and show the results of different methods on facebook and cond-mat in Fig. 2. In each subgraph, the x-axis indicates the training rate and the y-axis represents the AUC

and Prec. We make the following observation. (1) The AUC performance of all the methods decreases when introduces more training instance and contrast phenomenon appears in Prec. For example, on facebook, the performance decreases down 1.28% to 3.17% in terms of Prec and increases up 2.13% to 17.26% in terms of AUC. We can see that more training data is beneficial for increasing the AUC performance and is helpless to Prec metric. (2) WLNBs consistently outperforms all baselines with significant gain in most cases. The reason is that the alternative metrics lack discriminative features stemming from common neighbors. WLNBs takes its advantage and achieves better performance. As aforementioned, WJaccard and WSalton undoubtedly stay in the bottom level. (3) It can also be observed that all the three methods of WLNBs perform stable and have small amplitudes. Even on cond-mat dataset, the framework outperforms the nearest baseline for tiny proportions for training data (50%), demonstrating that it performs well for low training data sizes. Such performance shows consistent trends in all five datasets. In summary, the results demonstrate that the framework can stably learn from a small training set, and it effectively utilizes training data to predict links in weighted scenario. The framework WLNBs consistently performs well across all proportions of training data and hence is robust to its variations.

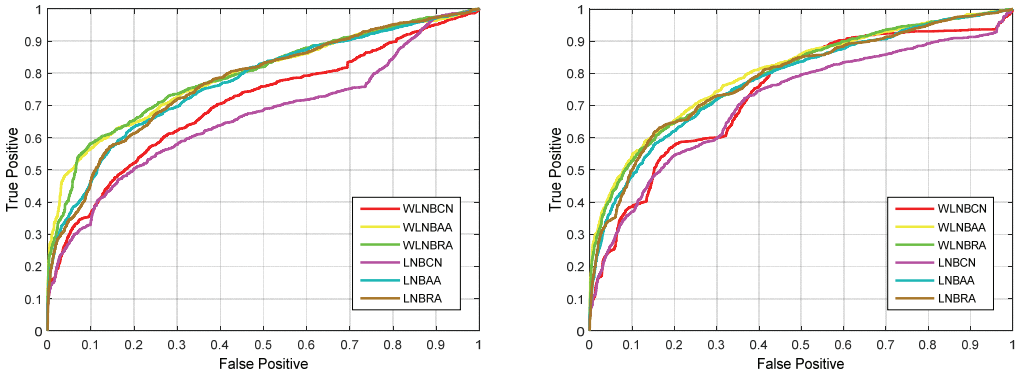


(a)



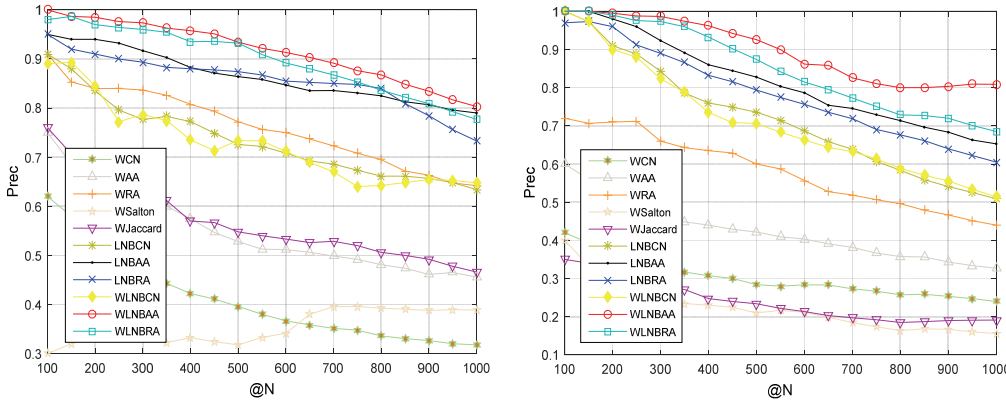
(b)

Fig. 2. Performance of AUC and Prec by varying the percentage of training network: (a) facebook, (b) cond-mat. The x-axis indicates the training size, and the y-axis represents the corresponding score.



(a) (b)

Fig. 3. The ROC curve: (a) cond-mat, (b) hep-hp. The x-axis indicates TPR (true positive rate), and the y-axis represents the FPR (false positive rate).



(a) (b)

Fig. 4. @N performance: (a) cond-mat, (b) hep-hp. The x-axis indicates precision, and the y-axis represents the N value.

5.4 Performance Analysis

First we extend the experiments to show the advantage of WLNBs against LNBs by ROC (receiver operating characteristic curve). A ROC curve is generated by plotting the TPR (true positive rate) on the vertical axis and the FPR (False Positive Rate) on the horizontal orient. In ROC, the closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. We show the results on cond-mat and hep-hp dataset in Fig. 3. All the methods did well in the predictive task and we observe only small difference. The major differences between WLNBs and LNBs appear in their different weighted and unweighted formats since WLNBs consistently lie below the corresponding LNBs. For instance, on hep-hp dataset, we can observe that the yellow lines (WLNBAA) is always on top of blue lines (LNBAA), which indicates that our proposed model has the largest area under its curve. Similar phenomenon proves the effectiveness of WLNBs again.

Secondly, as the ranking task is to find which potential links have the highest similarity score and

whether such set of high score is the ground truth. Thus, in this paragraph we evaluate the experiment by Top-N evaluation and set N=100 to 1000 of Prec@N in steps of 50, respectively. We only report the experimental results on cond-mat and hepth datasets since we have similar observations on other datasets. The results are reported in Fig. 4. It can be seen that three kinds of WLNBs still lie below the corresponding WCN, WAA, WRA and LNBs by a larger margin when N goes up. This matches our intuition as they neither ignore the common neighbors' different influence nor neglect the diverse link weight while WLNBs captures both weights, and thus obtains better performance. Besides, with the increment of N, the performance of all the methods becomes worse. For example, on cond-mat dataset, the accuracy of WLNBAA decreases gradually from 1 to 0.805 as N increase from 100 to 1000. This is because more noisy information is added as the candidate set becomes larger. This matches the intuition of link prediction—the more similar the pair of users is, the more likely they are linked. Besides, comparing WLNBAA with WLNBRA, the performance of both methods stands almost the same place (0.998 and 0.972) on top as N is 100. When N decreases, both performance drops. However, we can observe that the decreasing rate of WLNBAA is smaller than that of WLNBRA when N gets larger, which means WLNBAA is less sensitive to N and holds a stable performance.

6. Conclusions and Future Work

In this work, we have introduced WLNB, an unsupervised link prediction model that overcomes the limitation of similarity based weighted metrics. We first introduce the prediction problem using a local Bayes model and then propose a weighted format of LNB by considering the weighted network property: weighted cluster coefficient. Then we extend WLNB to other two representative local similarity metrics to validate the robustness of our proposed model. The effectiveness and stability of WLNB are demonstrated through extensive experiments conducted on real-world weighted networks.

Building an effective link prediction model for weighted network is important for complex network analysis. We are interested in incorporating other Bayes machine learning models to enhance the prediction task in future work. It is also interesting to extend this work to multi-layer weighted network and investigate the layer correlation into the proposed model.

Acknowledgement

This paper is supported by Guangdong Provincial Higher outstanding young teachers Training Program (No. YQ2015177). This work is in part supported by Natural Science Foundation of China (Nos. 61003174 and 61070090), Natural Science Foundation of Guangdong Province (No. S2012010009961), Specialized Research Fund for Guangdong Special Education Research Project (No. GDJY-2014-B-B200), Guangdong Higher Vocational Education Research Project (No. GDGZ14Y037), Major engineering technical and commercial services applied research project (No. GDGM2015-ZZ-C03), the Project funded by China Postdoctoral Science Foundation (No. 2016M602674), and the Fundamental Research Funds for the Central Universities of China (No. A03012023601042).

References

[1] A. L. Barabasi, *Linked: How Everything is Connected to Everything Else and What It Means for Business, Science, and Every Life*. Cambridge: Penguin Group, 2002.

[2] T. Zhou, L. Lu, and Y. C. Zhang, "Predicting missing links via local information," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 71, no. 4, pp. 623-630, 2009.

[3] A. Clauset, C. Moore, and M. E. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, 2008, vol. 453, no. 7191, pp. 98-101. 2008.

[4] D. K. Wind and M. Morup, "Link prediction in weighted networks," in *Proceedings of 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, 2012, pp. 1-6.

[5] H. Gui, Y. Sun, J. Han, and G. Brova, "Modeling topic diffusion in multi-relational bibliographic information networks," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, Shanghai, China, 2014, pp. 649-658.

[6] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the Association for Information Science and Technology*, vol. 58, no. 7, pp. 1019-1031, 2007.

[7] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211-230, 2003.

[8] S. Guha, R. Rastogi, and K. Shim, "ROCK: a robust clustering algorithm for categorical attributes," in *Proceedings of 15th International Conference on Data Engineering*, Los Alamitos, CA, 1999, pp. 512-521.

[9] M. E. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, article no. 056131, 2004.

[10] Z. Yu, C. Wang, J. Bu, X. Wang, Y. Wu, and C. Chen, "Friend recommendation with content spread enhancement in social networks," *Information Sciences*, vol. 309, pp. 102-118, 2015.

[11] T. Murata and S. Moriyasu, "Link prediction of social networks based on weighted proximity measures," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, Silicon Valley, CA, 2007, pp. 85-88.

[12] Z. Liu, Q. M. Zhang, L. Lu, and T. Zhou, "Link prediction in complex networks: a local naïve Bayes model," *EPL (Europhysics Letters)*, vol. 96, no. 4, article no. 48007, 2011.

[13] A. Zarezade, A. Khodadadi, M. Farajtabar, H. R. Rabiee, and H. Zha, "Correlated cascades: compete or cooperate," 2015 [Online]. Available: https://arxiv.org/abs/1510.00936.

[14] Y. Sun, J. Tang, J. Han, M. Gupta, and B. Zhao, "Community evolution detection in dynamic heterogeneous information networks," in *Proceedings of the 8th Workshop on Mining and Learning with Graphs*, Washington, DC, 2010, pp. 137-146.

[15] L. Lu and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150-1170, 2011.

[16] M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proceedings of the 4th Workshop on Link Analysis, Counterterrorism and Security (SDM06)*, Bethesda, MD, 2006.

[17] W. Liu and L. Lu, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, article no. 58007, 2010.

[18] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla, "New perspectives and methods in link prediction," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, 2010, pp. 243-252.

[19] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 11, pp. 3747-3752, 2004.

[20] Z. Yang, D. Fu, Y. Tang, Y. Zhang, Y. Hao, C. Gui, X. Ji, and X. Yue, "Link prediction based on weighted networks," in *AsiaSim 2012*. Heidelberg: Springer, 2012, pp. 119-126.

[21] H. R. De Sa and R. B. Prudencio, "Supervised link prediction in weighted networks," in *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)*, San Jose, CA, 2011, pp. 2281-2288.

[22] L. Lu and T. Zhou, "Link prediction in weighted networks: The role of weak ties," *EPL (Europhysics Letters)*, vol. 89, no. 1, article no. 18001, 2010.

[23] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404-409, 2001.

[24] S. Scellato, A. Noulas, and C. Mascolo, "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*, San Diego, CA, 2011, pp. 1046-1054.

[25] R. K. Pan, K. Kaski, and S. Fortunato, "World citation and collaboration networks: uncovering the role of geography in science," *Scientific Reports*, vol. 2, article no. 902, 2012.

JieHua Wu

He received B.S. in School of Computer Science and Engineering from Central South University 2005 and M.S. degrees in School of Software Engineering from Sun Yat-sen University in 2007, respectively. Since June 2007, he is with the School of Computer Science and Engineering from Guangdong Polytechnic of Industry and Commerce as an associate professor. Since March 2012, he is also with the School of Computer Science and Engineering from South China University of Technology University as a PhD candidate. His current research interests include data mining, network analysis.

GuoJi Zhang

He is a Professor in the School of Sciences, South China University of Technology, Guangzhou, China. He received his B.Sc. degree in Computer Application and Ph.D. degree in Circuit and System from South China University of Technology, in 1977 and 1999, respectively. His research interests include computational intelligence, computational electromagnetic and cryptology. He has published over 50 research papers.

YaZhou Ren

He is a Lecturer in the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. He received his B.Sc. degree in Information and Computation Science and Ph.D. degree in Computer Science from South China University of Technology, Guangzhou, China in 2009 and 2014, respectively. He visited the Data Mining Laboratory at George Mason University, USA, for two years from 2012 to 2014. His current research interests include clustering, semi-supervised learning and self-paced learning.

XiaYan Zhang

She received B.S degrees in School of Computer Science and Engineering from South China University of Technology in 2012. Since March 2012, she is with the School of Computer Science and Engineering from South China University of Technology University as a PhD candidate. Her current research interests include cryptography, cryptanalysis and cipher theory.

Qiao Yang

She received M.S. degrees in School of Computer Science and Engineering from South China University of Technology in 2015. Her current research interests include social network analysis.