Published in "Europhysics Letters 96(4): 48007, 2011"
which should be cited to refer to this work.

Link prediction in complex networks: A local naïve Bayes model

Zhen Liu[1], Qian-Ming Zhang[1], Linyuan Lü[2(a)] and Tao Zhou[1]

[1] *Web Sciences Center, University of Electronic Science and Technology of China - Chengdu 610054, PRC*
[2] *Department of Physics, University of Fribourg - Chemin du Musée 3, Fribourg CH-1700, Switzerland*

Abstract – The common-neighbor–based method is simple yet effective to predict missing links, which assume that two nodes are more likely to be connected if they have more common neighbors. In the traditional method, each common neighbor of two nodes contributes equally to the connection likelihood. In this letter, we argue that different common neighbors may play different roles and thus contributes differently, and propose a local naïve Bayes model. Extensive experiments were carried out on nine real networks. Compared with the traditional method, the present method can provide more accurate predictions.

http://doc.rero.ch

Introduction. – The problem of link prediction aims at estimating the likelihood of the existence of a link between two nodes in a given network, based on the observed links [1]. Recently, the study of link prediction has attracted much attention from disparate scientific communities. In the theoretical aspect, accurate prediction indeed gives evidence to some underlying mechanisms that drive the network evolution [2]. Moveover, it is very possible to build a fair evaluation platform for network modeling under the framework of link prediction, which might be interested by network scientists [1,3]. In the practical aspect, for biological networks such as protein-protein interaction networks and metabolic networks [4–6], the experiments of uncovering the new links or interactions are costly, and thus to predict in advance and focus on the links most likely to exist can sharply reduce the experimental costs [7]. In addition, some of the representative methods of link prediction have been successfully applied to address the classification problem in partially labeled networks [8,9], as well as to identify the spurious links resulting from inaccurate information in the data [10].

Many methods for link predication have been proposed: some algorithms are based on Markov chains[11–13] and machine learning [14,15], and another group of algorithms are based on node similarity [1,16]. Common Neighbors (CN) is one of the simplest similarity indices. The basic assumption of CN is that two nodes are more likely to be connected if they have more common neighbors. Empirically, Kossinets and Watts [17] analyzed a large-scale social network, suggesting that two students having many mutual friends are very probable to be friend in the future. Also in online society like Facebook, the users tend to be friend if they have common friends and therefore form a community. Extensive analysis on disparate networks suggests that CN index performs well in predicting the missing links [16,18]. Lü *et al.* [19] suggested that in a network with large clustering coefficient, CN can provide competitively accurate predictions compared with the indices making use of global information. Very recently, Cui *et al.* [20] revealed that the nodes with more common neighbors are more likely to form new links in a growing network. Simply counting the number of common neighbors indicates that each common neighbor gives equal contribution to the connection likelihood. However, different common neighbors may play different roles. For instance, the common close friends of two people who do not know each other may contribute more to their possibly future friendship than their common nodding acquaintances. In this letter, we propose a probabilistic model based on the Bayesian theory, called Local Naïve Bayes (LNB) model, to predict the missing links in complex networks. Based on the LNB model, two node pairs with the same number of common neighbors could have different connection likelihoods. Experiments on nine real networks demonstrate that our method can effectively identify the different roles of common neighbors to the connection likelihood and thus give more accurate prediction than CN.

(a) E-mail: linyuan.lue@unifr.ch

http://doc.rero.ch

Problem description. – Consider an undirected network $G(V,E)$, where V and E are the sets of nodes and links, respectively. The multiple links and self-connections are not allowed. Each non-existent link, namely a link $(x,y)\in U-E$, where $x,y\in V$ and U denotes the universal set, will be assigned a score to quantify its existence likelihood. Higher score means higher probability that nodes x and y are connected, and vice versa. All the non-existent links are sorted in descending order according to their scores, and the links at the top are most likely to exist. To test the algorithm's accuracy, the observed links, E, are randomly divided into two parts: the training set, E^T, is treated as known information, while the probe set, E^P, is used for testing and no information therein is allowed to be used for prediction. Clearly, $E=E^T\cup E^P$ and $E^T\cap E^P=\phi$. In this Letter, the training set always contains 90% of links, and the remaining 10% of links constitute the probe set[1]. Hereinafter, the links in E^P are called missing links and the links in $U-E^T$ are called non-observed links.

We apply two standard metrics to quantify the prediction accuracy: *AUC* (area under the receiver operating characteristic curve) [21] and *precision* [22]. The AUC evaluates the algorithms performance according to the whole list. Provided the scores of all non-observed links, AUC can be interpreted as the probability that a randomly chosen missing link is given a higher score than a randomly chosen non-existent link. In the implementation, among n times of independent comparisons, if there are n' times the missing link having higher score and n'' times that they have the same score, the AUC value is

$$\texttt{AUC}=(n'+0.5n'')/n. \tag{1}$$

If all the scores are generated from an independent and identical distribution, AUC should be about 0.5. Therefore, the degree to which the value exceeds 0.5 indicates how much better the algorithm performs than pure chance. Different from AUC, precision only focuses on the L links with highest scores. It is defined as the ratio of relevant items selected to the items selected. Among the top-L links, if L_r links are accurately predicted (there are L_r links in the probe set), then the precision equals L_r/L.

Method. – Given a pair of disconnected nodes (x, y) (*i.e.*, a non-observed link), our task is to calculate their connecting probability on the basis of the condition: x and y might have a group of common neighbors, each of which has a couple of conditional probabilities corresponding to encouraging and hampering the connections of its two neighbors (see eqs. (6) and (7)).

The naïve Bayes classifier. A naïve Bayes classifier is a simple probabilistic classifier based on Bayesian theory with strong (naïve) independence assumptions that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature [23]. Abstractly, the probability model for a classifier is a conditional model $P(C|F_1,\cdots,F_n)$, where C is a dependent class variable and $F_1,\ F_2,\cdots,F_n$ are feature variables. According to the Bayesian theory[2], the posterior probability $P(C|F_1,F_2,\cdots,F_n)$ is

$$P(C|F_1,F_2,\cdots,F_n)=\frac{P(C)\cdot P(F_1,F_2,\cdots,F_n|C)}{P(F_1,F_2,\cdots,F_n)}. \tag{2}$$

Consider the naïve assumption that each feature F_i is conditionally independent to every other feature F_j ($j\neq i$), then we have

$$P(C|F_1,F_2,\cdots,F_n)=\frac{P(C)\cdot\prod_{i=1}^{n}P(F_i|C)}{P(F_1,F_2,\cdots,F_n)}. \tag{3}$$

Local naïve Bayes model. Given a training set $G(V,E^T)$, link prediction questions which links are more likely to exist among all the non-observed links. Denote by A_1 and A_0 the class variables of *connection* and *disconnection* respectively. The prior probabilities of A_1 and A_0 can be calculated through

$$P(A_1)=\frac{M^T}{M}, \tag{4}$$

$$P(A_0)=\frac{M-M^T}{M}, \tag{5}$$

where $M=|V|(|V|-1)/2$ and $M^T=|E^T|$. Each node w owns two conditional probabilities $\{P(w|A_1),P(w|A_0)\}$, where $P(w|A_1)$ is the probability that node w is a common neighbor of two connected nodes, and $P(w|A_0)$ is the probability that node w is a common neighbor of two disconnected nodes. According to Bayesian theory, these two probabilities are

$$P(w|A_1)=\frac{P(w)\cdot P(A_1|w)}{P(A_1)}, \tag{6}$$

$$P(w|A_0)=\frac{P(w)\cdot P(A_0|w)}{P(A_0)}, \tag{7}$$

where $P(w)$ means the probability that a given node w is a common neighbor of a pair of nodes. For a pair of disconnected nodes (x, y), denote by O_{xy} the set of their common neighbors that are considered as the feature variables and assumed to be independent of each other. Then according to the naïve Bayes classifier theory, the posterior probability of connection and disconnection of

[1] Besides the 90% *vs.* 10% division, we have also tested the cases 75% *vs.* 25% and 50% *vs.* 50%, and we found that the main conclusions are not changed.

[2] Bayesian theory [24,25] is a probabilistic approach which relates the conditional and marginal probabilities of events A and B, provided that the probability of B does not equal zero: $P(A|B)=\frac{P(A)\cdot P(B|A)}{P(B)}$, where $P(A)$ is the prior probability (or marginal probability) of A which does not take into account any information about B, and $P(A|B)$ is the conditional probability of A given B. $P(A|B)$ is also called the posterior probability.

node x and y are

$$P(A_1|O_{xy}) = \frac{P(A_1)}{P(O_{xy})} \prod_{w \in O_{xy}} P(w|A_1), \tag{8}$$

$$P(A_0|O_{xy}) = \frac{P(A_0)}{P(O_{xy})} \prod_{w \in O_{xy}} P(w|A_0). \tag{9}$$

For a given node pair, comparing these two probabilities we can obtain whether they are likely to connect. However, it can not tell us which non-observed links are more likely to exist than the others. In order to compare the existence likelihood between node pairs, we define the likelihood score of node pair (x, y) as the ratio of $P(A_1|O_{xy})$ to $P(A_0|O_{xy})$. Substituting eqs. (6) and (7), we have

$$r_{xy} = \frac{P(A_1)}{P(A_0)} \prod_{w \in O_{xy}} \frac{P(A_0) \cdot P(A_1|w)}{P(A_1) \cdot P(A_0|w)}. \tag{10}$$

$P(A_1|w)$ is equal to the clustering coefficient, C_w, as

$$P(A_1|w) = C_w = \frac{N_{\triangle w}}{N_{\triangle w} + N_{\wedge w}}, \tag{11}$$

where $N_{\triangle w}$ and $N_{\wedge w}$ are respectively the number of connected and disconnected node pairs whose common neighbors include w. Obviously $N_{\triangle w} + N_{\wedge w} = \frac{k_w \times (k_w - 1)}{2}$, where k_w is the degree of node w. Since $P(A_1|w) + P(A_0|w) = 1$, we have

$$P(A_0|w) = 1 - C_w = \frac{N_{\wedge w}}{N_{\triangle w} + N_{\wedge w}}. \tag{12}$$

Substituting eqs. (4), (5) and eqs. (11), (12), the likelihood score of node pair (x, y) is

$$r_{xy} = s^{-1} \prod_{w \in O_{xy}} s \frac{N_{\triangle w} + 1}{N_{\wedge w} + 1}, \tag{13}$$

where $s = \frac{P(A_0)}{P(A_1)} = \frac{M}{M^T} - 1$. Note that we here apply the *add-one smoothing* to prevent the score from being equal to 0. For a given node w, we directly define its role function as

$$R_w = \frac{N_{\triangle w} + 1}{N_{\wedge w} + 1}. \tag{14}$$

Therefore eq. (13) can be written as

$$r_{xy} = s^{-1} \prod_{w \in O_{xy}} s R_w. \tag{15}$$

Clearly, if $R_w = 1$ for all nodes in the network, then the score of nodes x and y, r^{xy}, will become a monotone increasing function of the number of their common neighbors. In this case, eq. (15) is equivalent to CN $(= |O_{xy}|)$. Different common neighbors are generally of different contributions to the connecting probability, according to the corresponding R_w.

It has been pointed out that the common neighbor's degree plays an important role in link prediction. Suppressing the contributions of common neighbors with high degrees can improve the prediction accuracy [18]. Two indices are designed in this way, the Adamic-Adar index (AA) [26] and the Resource Allocation (RA) [18,27]. The scores are defined respectively as

$$r_{xy}^{\texttt{AA}} = \sum_{w \in O_{xy}} \frac{1}{\log k_w}, \tag{16}$$

and

$$r_{xy}^{\texttt{RA}} = \sum_{w \in O_{xy}} \frac{1}{k_w}. \tag{17}$$

With the same motivation, we add an exponent $f(k_w)$ to the item sR_w in eq. (15), where f is a function of the node's degree. Using a logarithmic function on both sides and neglecting the constant factor s^{-1}, we obtain a linear formula of connection likelihood:

$$r'_{xy} = \sum_{w \in O_{xy}} f(k_w) \log(sR_w). \tag{18}$$

Here we consider three forms of function f, namely $f(k_w) = 1$, $f(k_w) = \frac{1}{\log k_w}$ and $f(k_w) = \frac{1}{k_w}$, which are corresponding to the Local Naïve Bayes (LNB) form of CN, AA and RA indices, respectively:

$$r_{xy}^{\texttt{LNB-CN}} = |O_{xy}| \log s + \sum_{w \in O_{xy}} \log R_w, \tag{19}$$

$$r_{xy}^{\texttt{LNB-AA}} = \sum_{w \in O_{xy}} \frac{1}{\log k_w} (\log s + \log R_w), \tag{20}$$

$$r_{xy}^{\texttt{LNB-RA}} = \sum_{w \in O_{xy}} \frac{1}{k_w} (\log s + \log R_w). \tag{21}$$

Obviously, when $R_w = 1$, the LNB-CN, LNB-AA and LNB-RA will degenerate to CN, AA and RA, respectively.

Results. – Nine networks are considered in our experiments: i) USAir [28]: The network of US air transportation system, which contains 332 airports and 2126 airlines. ii) EEAir [29]: The air transportation network of Eastern Europe. iii) *C. elegans* (CE) [30]: The neural network of the nematode worm *C. elegans*, in which an edge joins two neurons if they are connected by either a synapse or a gap junction. iv) Political Blogs (PB) [31]: A network of the US political blogs. The original links are directed, here we treat them as undirected links. v) Yeast [32]: A protein-protein interaction network containing 2617 proteins and 11855 interactions. Although this network is not well connected (it contains 92 components), most of the nodes belong to the giant component, whose size is 2375. vi) NetScience (NS) [33]: A network of coauthorships between scientists who are themselves publishing on the topic of networks. The network contains 1589 scientists, and 128 of which are isolated. NS consists of 268 connected components, and the size of the largest connected component is only 379. vii) Foodweb1 (FW1) [34]: A network

http://doc.rero.ch

Table 1: The basic topological features of the nine networks. $|V|$ and $|E|$ are the number of nodes and links. C and r are clustering coefficient [37] and assortative coefficient [38], respectively. $\langle k \rangle$ and $\langle d \rangle$ are the average degree and the average shortest distance. H denotes the degree heterogeneity defined as $H = \langle k^2 \rangle / \langle k \rangle^2$.

Networks	USAir	EEAir	Yeast	CE	PB	NS	FW1	FW2	FW3
$\|V\|$	332	115	2375	297	1222	379	128	69	97
$\|E\|$	2126	299	11693	2148	16717	914	2106	880	1446
C	0.749	0.668	0.388	0.308	0.361	0.798	0.335	0.552	0.468
r	−0.208	−0.287	0.454	−0.163	−0.221	−0.082	−0.104	−0.298	−0.151
$\langle k \rangle$	12.81	5.20	9.85	14.46	27.36	4.82	32.90	25.51	29.81
$\langle d \rangle$	2.74	2.53	5.10	2.46	2.74	6.04	1.77	1.64	1.69
H	3.46	3.43	3.48	1.80	2.97	1.663	1.231	1.27	1.27

Table 2: The prediction accuracy measured by AUC and precision (top-100) on nine networks. Each value is obtained by averaging over 100 implementations with independently random divisions of training set and probe set.

	Index	USAir	EEAir	Yeast	CE	PB	NS	FW1	FW2	FW3
AUC	CN	0.953	0.920	0.916	0.848	0.924	0.980	0.606	0.689	0.710
	LNB-CN	0.959	0.941	0.916	0.862	0.926	0.982	0.694	0.733	0.748
	AA	0.965	0.942	0.916	0.865	0.927	0.984	0.608	0.697	0.713
	LNB-AA	0.967	0.944	0.916	0.866	0.928	0.984	0.697	0.733	0.750
	RA	0.972	0.944	0.916	0.870	0.928	0.984	0.613	0.704	0.716
	LNB-RA	0.972	0.944	0.917	0.867	0.929	0.984	0.697	0.730	0.750
Precision	CN	0.597	0.120	0.685	0.131	0.419	0.356	0.087	0.146	0.133
	LNB-CN	0.612	0.124	0.689	0.138	0.409	0.391	0.106	0.192	0.161
	AA	0.615	0.140	0.699	0.135	0.378	0.527	0.090	0.156	0.139
	LNB-AA	0.629	0.133	0.703	0.136	0.380	0.528	0.104	0.193	0.161
	RA	0.630	0.138	0.506	0.126	0.247	0.547	0.086	0.169	0.145
	LNB-RA	0.633	0.137	0.625	0.129	0.259	0.548	0.104	0.196	0.170

of foodweb in Florida Bay during wet season. viii) Foodweb2 (FW2) [35]: A network of foodweb in Everglades Graminoids during wet season. ix) Foodweb3 (FW3) [36]: A network of foodweb in Mangrove Estuary during wet season. Here we only consider the giant component and every network is treated as undirected network. The basic topological features of these nine networks are summarized in table 1.

The prediction accuracy, measured by AUC and precision, on the nine real networks are shown in table 2. In general, the LNB forms outperform their corresponding basic forms. It shows that for AUC, except the result of RA in CE network, LNB model gives higher accurate prediction for all nine networks. Especially the improvements of the foodwebs are significant. When measured by precision, except the result of CN in PB network and the AA and RA in EEAir, LNB model also improve the accuracy. In accordance with our analysis, CN assigns equal weight to the common neighbors (*i.e.*, $R_w = 1$ for all nodes), while LNB-CN can effectively capture the different roles of common neighbors. Notice that, some elaborately designed algorithms [7,10] could provide more accurate prediction, although they are more time consuming. In particular, the CN-based methods are poor for low-clustering networks, yet for these networks, the hierarchical model [7] and stochastic block model [10] still work well.

Case study. – In this section, we give detailed analysis on USAir network which contains 332 airports and 2126 airlines. This network has a very specific structure: the hierarchical organization consisting of hubs, local centers and small local airports. We rank all the airports according to their degrees in descending order. The top 17 airports who own about 1/3 of the total degree are defined as hubs (Hub), the last 273 airports who also own 1/3 of the total degree are local airports (LA), and the rest 42 airports are local centers (LC). Therefore there are six kinds of links: Hub-Hub (97.06%), Hub-LC (72.83%), LC-LC (36.24%), LA-Hub (12.5%), LA-LC (2.75%) and LA-LA (0.72%). The number in the bracket indicates the connecting probability: the ratio of the number of real links to its possibly maximal value. For example, there are 132 links connecting two hubs among $\frac{17\times16}{2} = 136$ possible links, and thus the connection probability for two hubs is 97.06%, indicating that hubs are densely connected with each other, which is a specific feature of air transportation network [39]. Due to this reason, both CN and LNB-CN methods can provide accurate predictions on the missing Hub-Hub links. Moveover, we find that although the number of common neighbors of (LC, LC) are higher than that of (Hub, LC), LCs are more likely to connect with Hubs (see $72.83\% > 26.24\%$). Therefore by simply counting the number of common neighbors, CN tends to assign higher score to (LC, LC) than (Hub,

http://doc.rero.ch

http://doc.rero.ch

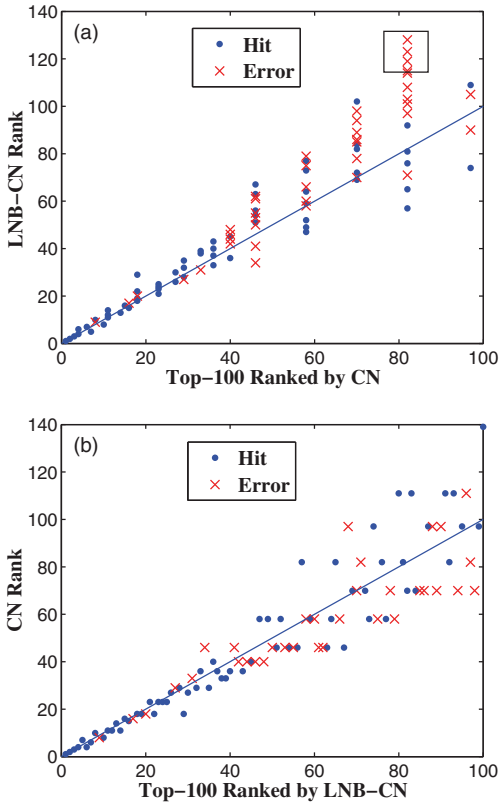

Fig. 1: (Colour on-line) Comparison of the top-100 node pairs, respectively, ranked by CN and LNB-CN on the USAir network. (a) The top-100 node pairs ranked by CN and their corresponding ranks assigned by LNB-CN. (b) The top-100 node pairs ranked by LNB-CN and their corresponding ranks assigned by CN. "Hit" denotes the correctly predicted link (*i.e.*, links in probe set), while "Error" indicates the non-existent link. The diagonal is presented by a blue solid line.

LC) and thus leads to poor predictions. Since LBN-CN is sophisticate to identify the negative roles of (LC, LC)'s common neighbors, it will depress the score of (LC, LC) and provide more accurate predictions. Compared with the first three kinds of links, the remaining three kinds of links involving local airports are difficult to predict (none of such links are included in top-100 with both CN and LNB-CN). Because LAs usually connected with Hubs and rarely have common neighbors with other airports, the nodes pairs involving LAs are given very small scores and thus ranked lower. This is a common drawback of CN-based methods.

We further focus on the top-100 node pairs, respectively, ranked by CN and LNB-CN. Figure 1(a) shows the top-100 node pairs ranked by CN and their corresponding ranks assigned by LNB-CN, and fig. 1(b) shows the top-100 node pairs ranked by LNB-CN and their corresponding ranks assigned by CN. The links who are correctly predicted are labeled by blue dots while the non-existent links are labeled by red crosses. From fig. 1(a), we can see that among top-100 node pairs ranked by CN, 11 node pairs are ranked above 100 by LNB-CN, within which only 2 of them

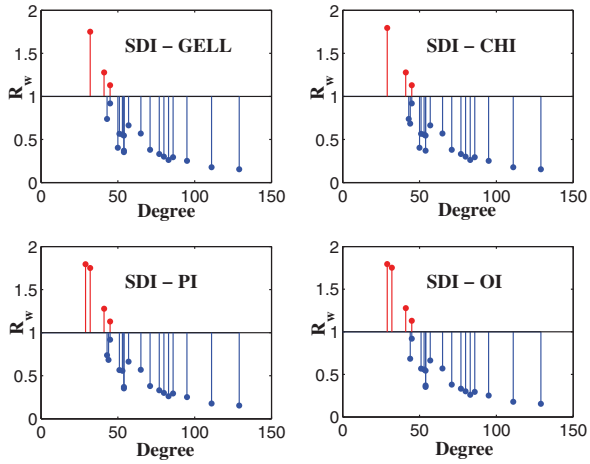

Fig. 2: (Colour on-line) The dependence of common neighbors' R_w on its degree for the four pairs.

are predicted right, which implies that LNB-CN tends to rank the non-existent links lower than CN. Among the remaining 9 node pairs, 6 of them are LC-LC links and 3 are Hub-LC pairs. This result further demonstrates that LNB-CN can give more accurate judgements on the LC-LC and Hub-LC links. Similarly, in fig. 1(b), 6 pairs are ranked above 100 by CN, within which 5 links are predicted right, which indicates that LNB-CN tends to rank missing links higher than CN.

We take four pairs within the rectangle in fig. 1(a) as examples. Five airports are involved, including *San Diego Intl* (SDI), *General Edward Lawrence Logan* (GELL), *Cleveland-Hopkins Intl* (CHI), *Philadelphia Intl* (PI) and *Orlando Intl* (OI). Therein, PI, with over 62 airlines is a hub, others are local airports. These four pairs, ranked 128, 123, 119 and 114, respectively, by LNB-CN, are corresponding to four non-existent airlines, namely (GELL, SDI), (CHI, SDI), (OI, SDI) and (PI, SDI). Each of them has 21 common neighbors and is ranked 82th by CN. Since SDI is located at the west coast while other four airports are all at the east of USA, there are no direct airlines connecting SDI and other three airports. Instead the passengers need to transfer at their common neighbors (most of which are hubs). This feature can be well captured by LNB-CN. Figure 2 shows the dependence of common neighbors' R_w on its degree for these four pairs. For all four pairs, most of the common neighbors play negative roles and thus hamper the connections. Therefore, LNB-CN will rank this kind of pairs lower than CN by assigning them small scores. In a word, compared with the CN method, LNB-CN is more powerful to uncover the different roles of nodes in forming a network, which is also an attractive issue in recent studies [10,40].

Conclusion. – In this letter, we proposed a local naive Bayes model (LNB) to predict missing links in complex networks. The advantage of this method is that it can well capture the different roles of common neighbors

and assign them different weights. To test the method, we compared three representative local indices, namely Common Neighbors (CN), Adamic-Adar (AA) index and Resource Allocation (RA) index, with their corresponding LNB forms. Extensive analysis on nine real networks drown from disparate fields shows that for all these three indices the LNB forms outperform their corresponding original indices. In particular, the improvement is remarkable on the foodwebs where the hierarchical structure is obvious and there are rare links within the same level. Finally, we gave a detailed analysis on US air transportation network. Although some pairs of airports have many common neighbors, there are no direct airlines connecting them because of the long geographical distance [41]. LNB methods can well capture this feature and thus give more accurate predictions. In addition, some researchers found that the local clustering property can be utilized to improve the accuracy of link prediction [42], yet they did not give any solid reason about their method, while the present model provides a theoretical base on the usage of local clustering.

* * *

We acknowledge Chong-Jing Sun for helpful assistance in manuscript preparation and Duanbing Chen for valuable suggestions. This work is partially supported by the National Natural Science Foundation of China under Grant Nos. 11075031, 60903073 and 60973069. LL acknowledges the Swiss National Science Foundation under Grant No. 200020-132253.

REFERENCES

[1] Lü L. and Zhou T., *Physica A*, **390** (2011) 1150.
[2] Liu H.-K., Lü L. and Zhou T., *Sci. China Ser. G*, **41** (2011) 816.
[3] Wang W.-Q., Zhang Q.-M. and Zhou T., *Evaluating Network Evolving Models: A Likelihood Analysis*, in preparation.
[4] Yu H., Braun P., Yildirim M. A., Lemmens I., Venkatesan K., Sahalie J., Hirozane-Kishikawa T., Gebreab F., Li N., Simonis N., Hao T., Rual J. F. and Dricot A., *Science*, **322** (2008) 104.
[5] Stumpf M. P. H., Thorne T., de Silva E., Stewart R., An H. J., Lappe M. and Wiuf C., *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008) 6959.
[6] Amaral L. A. N., *Proc. Natl. Acad. Sci. U.S.A.*, **105** (2008) 6795.
[7] Clauset A., Moore C. and Newman M. E. J., *Nature*, **453** (2008) 98.
[8] Gallagher B., Tong H., Eliassi-Rad T. and Faloutsos C., *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press) 2008, pp. 256–264.
[9] Zhang Q.-M., Shang M.-S. and Lü L., *Int. J. Mod. Phys. C*, **21** (2010) 813.
[10] Guimerà R. and Sales-Pardo M., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 22073.
[11] Bilgic M., Namata G. M. and Getoor L., *7th IEEE International Conference on Data Mining* (IEEE Press) 2007, pp. 381–386.
[12] Zhu J., Hong J. and Hughes J. G., *Lect. Notes Comput. Sci.*, **2311** (2002) 55.
[13] Sarukkai R. R., *Comput. Netw.*, **33** (2000) 377.
[14] Popescul A. and Ungar L., *Proceedings of the workshop on Learning Statistical Models from Relational Data at JMCAI-2003* (ACM Press) 2003, pp. 81–87.
[15] Wang C., Satuluri V. and Parthasarathy S., *7th IEEE International Conference on Data Mining* (IEEE Press) 2007, pp. 1–10.
[16] Liben-Nowell D. and Kleinberg J., *J. Am. Soc. Inf. Sci. Technol.*, **58** (2007) 1019.
[17] Kossinets G., *Soc. Netw.*, **28** (2006) 247.
[18] Zhou T., Lü L. and Zhang Y.-C., *Eur. Phys. J. B*, **71** (2009) 623.
[19] Lü L., Jin C.-H. and Zhou T., *Phys. Rev. E*, **80** (2009) 046122.
[20] Cui A.-X., Fu Y., Shang M.-S., Chen D.-B. and Zhou T., *Acta Phys. Sin.*, **60** (2011) 038901.
[21] Hanley J. A. and McNeil B. J., *Radiology*, **143** (1982) 29.
[22] Herlocker J. L., Konstann J. A., Terveen K. and Riedl J. T., *ACM Trans. Inf. Syst.*, **22** (2004) 5.
[23] `http://en.wikipedia.org/wiki/Naive_bayes`.
[24] Bernardo J. M. and Smith A. F M., *Meas. Sci. Technol.*, **12** (2001) 221.
[25] `http://en.wikipedia.org/wiki/Bayes%27_theorem`.
[26] Adamic L. A. and Adar E., *Soc. Netw.*, **25** (2003) 211.
[27] Liu W. and Lü L., *EPL*, **89** (2010) 58007.
[28] Batageli V. and Mrvar A., `http://vlado.fmf.uni-lj.si/pub/networks/data/default.htm`.
[29] Guimerà R., Mossa S., Turtschi A. and Amaral L. A. N., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 7794.
[30] White J. G., Southgate E., Thomson I. N. and Brenner S., *Philos. Trans. R. Soc. B*, **314** (1986) 1.
[31] Ackland R., `http://incsub.org/blogtalk/images/robertackland.pdf`.
[32] von Merging C., Krause R., Snel B., Cornell M., Oliver S. G., Fields S. and Bork P., *Nature*, **417** (2002) 399.
[33] Newman M. E. J., *Phys. Rev. E*, **74** (2006) 036104.
[34] Ulanowicz R. E., Bondavalli C. and Egnotovich M. S., *Technical report, CBL 98-123* (1998), `http://www.cbl.umces.edu/~atlss/FBay701.html`.
[35] Ulanowicz R. E., Heymans J. J. and Egnotovich M. S., *Technical report, TS-191-99* (2000), `http://www.cbl.umces.edu/~atlss/swgras701.html`.
[36] Baird D., Luczkovich J. and Christian R. R., *Estuarine, Coastal, Shelf Sci.*, **47** (1998) 329.
[37] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.
[38] Newman M. E. J., *Phys. Rev. Lett.*, **89** (2002) 208701.
[39] Colizza V., Flammini A., Serrano M. A. and Vespignani A., *Nat. Phys.*, **2** (2006) 110.
[40] Guimerà R., Sales-Pardo M. and Amaral L. A. N., *Nat. Phys.*, **3** (2007) 63.
[41] Xie Y.-B., Zhou T., Bai W.-J., Chen G., Xiao W.-K. and Wang B.-H., *Phys. Rev. E*, **75** (2007) 036106.
[42] Feng X., Zhao J. and Xu K., arXiv:1103.4919 (2011).

http://doc.rero.ch