A LETTERS JOURNAL EXPLORING THE FRONTIERS OF PHYSICS

OFFPRINT

Link prediction via matrix completion

RATHA PECH, DONG HAO, LIMING PAN, HONG CHENG and TAO ZHOU

EPL, **117** (2017) 38002

A Letters Journal Exploring the Frontiers of Physics

AN INVITATION TO SUBMIT YOUR WORK

epljournal.org

The Editorial Board invites you to submit your letters to EPL

EPL is a leading international journal publishing original, innovative Letters in all areas of physics, ranging from condensed matter topics and interdisciplinary research to astrophysics, geophysics, plasma and fusion sciences, including those with application potential.

The high profile of the journal combined with the excellent scientific quality of the articles ensures that EPL is an essential resource for its worldwide audience. EPL offers authors global visibility and a great opportunity to share their work with others across the whole of the physics community.

Run by active scientists, for scientists

EPL is reviewed by scientists for scientists, to serve and support the international scientific community. The Editorial Board is a team of active research scientists with an expert understanding of the needs of both authors and researchers.

OVER
568,000
full text downloads in 2015

18 DAYS
average accept to online publication in 2015

20,300
citations in 2015

"We greatly appreciate the efficient, professional and rapid processing of our paper by your team."

Cong Lin
Shanghai University

Six good reasons to publish with EPL

We want to work with you to gain recognition for your research through worldwide visibility and high citations. As an EPL author, you will benefit from:

1. **Quality** – The 60+ Co-editors, who are experts in their field, oversee the entire peer-review process, from selection of the referees to making all final acceptance decisions.
2. **Convenience** – Easy to access compilations of recent articles in specific narrow fields available on the website.
3. **Speed of processing** – We aim to provide you with a quick and efficient service; the median time from submission to online publication is under 100 days.
4. **High visibility** – Strong promotion and visibility through material available at over 300 events annually, distributed via e-mail, and targeted mailshot newsletters.
5. **International reach** – Over 3200 institutions have access to EPL, enabling your work to be read by your peers in 100 countries.
6. **Open access** – Articles are offered open access for a one-off author payment; green open access on all others with a 12-month embargo.

Details on preparing, submitting and tracking the progress of your manuscript from submission to acceptance are available on the EPL submission website **epletters.net**.

If you would like further information about our author service or EPL in general, please visit **epljournal.org** or e-mail us at **info@epljournal.org**.

EPL is published in partnership with:

European Physical Society | Società Italiana di Fisica | EDP Sciences | IOP Publishing

A Letters Journal Exploring the Frontiers of Physics

February 2017

EPL, **117** (2017) 38002 www.epljournal.org
doi: 10.1209/0295-5075/117/38002

Link prediction via matrix completion

Ratha Pech[1], Dong Hao[1,2(a)], Liming Pan[1], Hong Cheng[3] and Tao Zhou[1,2(b)]

[1] *CompleX Lab, University of Electronic Science and Technology of China - Chengdu 611731, PRC*
[2] *Big Data Research Center, University of Electronic Science and Technology of China - Chengdu 611731, PRC*
[3] *Center for Robotics, University of Electronic Science and Technology of China - Chengdu 611731, PRC*

received 6 June 2016; accepted in final form 15 February 2017
published online 22 March 2017

PACS 89.65.-s – Social and economic systems
PACS 89.20.Ff – Computer science and technology
PACS 89.75.Hc – Networks and genealogical trees

Abstract – Inspired by the practical importance of social networks, economic networks, biological networks and so on, studies on large and complex networks have attracted a surge of attention in the recent years. Link prediction is a fundamental issue to understand the mechanisms by which new links are added to the networks. We introduce the method of robust principal component analysis (robust PCA) into link prediction, and estimate the missing entries of the adjacency matrix. On the one hand, our algorithm is based on the sparse and low-rank property of the matrix, while, on the other hand, it also performs very well when the network is dense. This is because a relatively dense real network is also sparse in comparison to the complete graph. According to extensive experiments on real networks from disparate fields, when the target network is connected and sufficiently dense, whether it is weighted or unweighted, our method is demonstrated to be very effective and with prediction accuracy being considerably improved compared to many state-of-the-art algorithms.

Introduction. – In the past decade, the rapid expansion of studies on complex networks has brought together different disciplines including physics, mathematics, computer science, sociology, economics, biology and so on [1,2]. The theory of complex networks provides us with novel insights for understanding the real-world linking patterns. The real-world linked datasets are usually dynamically changing and subjected to unobservability. On the one hand, the datasets are growing and changing over time through the increment of new links [3]. On the other hand, the missing or unobservable entries extensively exist in the datasets [4]. Therefore, predicting missing links is of great importance for studying the newly appeared and unobserved relations between data entries.

The link prediction problem essentially concerns the knowledge discovery and topology remodeling for large volumes of dynamic and noisy datasets [5], which also aims at uncovering to what extent the evolution of networks can be modeled and analyzed according to the intrinsic features and structures of the network itself [6]. So far it has been generally accepted as a fundamental paradigm not only in physics but also in bioinformatics, sociology, statistics and computer science.

Great effort has been made to solve the link prediction problem [7–15] and most of the algorithms are based on the similarity between vertex pairs since these algorithms are designed according to the fact that similar vertices are more likely to connect to each other. These algorithms are called similarity-based algorithms. Roughly, the similarity indices can be classified into three categories [8], *i.e.*, local [16,17], global [18,19] and quasi-local [20,21] indices. The most popular methods are the local ones because they are simple and applicable for very large-scale networks. Although they are computational efficient, the local similarity-based link prediction algorithms are sometimes less accurate.

The global topological information can be exploited through the adjacency matrix, where the nonzero entries denote the connections between vertices, while missing links and nonexisting links are both denoted by zero entries. In most cases, a very small fraction of zero entries (called hidden nonzero entries or hidden entries) represent the missing links and the rest (called null entries) represent the nonexisting links. Essentially speaking, a link prediction algorithm aims at recovering the hidden

(a) E-mail: haodong@uestc.edu.cn
(b) E-mail: zhutou@ustc.edu

nonzero entries from the real null entries according to all the nonzero entries in the adjacency matrix. However, for a real-world network, the adjacency matrix is usually very sparse (*i.e.*, most of its entries are zeros), providing highly limited information. How to precisely predict the missing links based on the sparse information is a challenging issue.

In this work, we introduce the robust principal component analysis (robust PCA) [22] method, namely low rank (LR), into link prediction and design a novel global information-based prediction algorithm based upon the low-rank and sparse property of the adjacency matrix. Then, by minimizing the nuclear norm of the matrix which fits the training data, we reconstruct a network that is close to the original network and accordingly identify the missing links. It is shown that when the target network is connected and sufficiently dense, we can find out the missing links with much higher accuracy compared to some of the state-of-the-art algorithms.

Method. – An undirected network consists of a set of vertices V and a set of links E. We do not consider multiple links and self-connections. Suppose we have an observed network represented by the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, which is a snapshot or a subset of an original network $\mathbf{G}^*$. The sets of links in $\mathbf{A}$ and $\mathbf{G}^*$ are denoted by E^T and E, respectively. Denote the rest of the links in E as E^P, namely $E = E^T \cup E^P$ and $E^T \cap E^P = \emptyset$. Then E^T is the training set for learning and prediction and E^P is the probe set for verifying the prediction accuracy. Without loss of generality, in the experiment we dynamically take $80\%, 85\%, 90\%$, and 95% of all the links in $\mathbf{G}^*$ as the training set and the rest as the probe set, respectively.

The objective of the link prediction is to find out the missing links of the original network $\mathbf{G}^*$, or equivalently, to recover a network $\mathbf{G}$ which is sufficiently close to $\mathbf{G}^*$, based on the observed entries of $\mathbf{A}$ (it is worth noting that it is generally intractable to recover exactly $\mathbf{G}^*$). Assume that i) $\mathbf{X}^* \in \mathbb{R}^{n\times n}$ conveys the pattern as to how the network evolves (how predicted links are added and some existing links are eliminated) and we call $\mathbf{X}^*$ the backbone network; ii) $\mathbf{X}$ is the subset of $\mathbf{X}^*$ containing only the predicted links, which can be obtained by resorting the values of elements corresponding to nonzero entries in $\mathbf{A}$ to be zero. $\mathbf{X}^*$ and $\mathbf{X}$ have real number values. Identifying network $\mathbf{X}^*$ is the crucial intermediate step for recovering the original network and predicting the missing links accordingly. The observed network $\mathbf{A}$ is the only information we can utilize. $\mathbf{X}^*$ can be represented by subtracting an error/noise matrix $\mathbf{E} \in \mathbb{R}^{n\times n}$ from $\mathbf{A}$ and this noise matrix should be much sparser than either $\mathbf{A}$ or $\mathbf{X}^*$. Therefore, $\mathbf{X}^*$ can be written as

$$\mathbf{X}^* = \mathbf{A} - \mathbf{E}, \tag{1}$$

where $\mathbf{E}$ is the noise matrix in which positive entries are the spurious links and negative entries represent the missing links that appear in $\mathbf{X}^*$. The relationship among

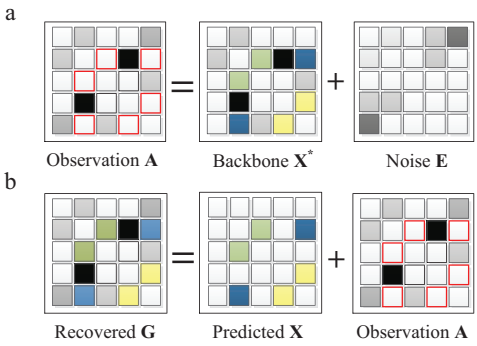

Fig. 1: (Color online) (a) The relationship between the observed network $\mathbf{A}$, the corresponding backbone network $\mathbf{X}^*$ containing newly appearing links and some existing links and the noise $\mathbf{E}$ containing spurious links in the observed network $\mathbf{A}$. (b) The relationship among the recovered network $\mathbf{G}$, the predicted network $\mathbf{X}$ containing only newly appearing links and the observed network $\mathbf{A}$. The white squares represent real null entries with value zero, the white squares with red frames represent the missing or likely existing links (values are also zeros), while the other entries indicated by colored squares are the values greater than zero.

$\mathbf{G}$, $\mathbf{A}$, $\mathbf{X}$, $\mathbf{X}^*$ and $\mathbf{E}$ is illustrated in fig. 1. The recovered network $\mathbf{G}$ is obtained as

$$\mathbf{G} = \mathbf{X} + \mathbf{A}. \tag{2}$$

$\mathbf{X}$ contains only newly appeared links and it is defined as

$$x_{ij} = \begin{cases} x^*_{ij}, & a_{ij} = 0. \\ 0, & a_{ij} = 1. \end{cases} \tag{3}$$

The principal component analysis (PCA) can be utilized to obtain $\mathbf{X}^*$ and $\mathbf{E}$ simultaneously by converting the observed network $\mathbf{A}$ into a set of linearly uncorrelated variables called principal components, which captures the backbone network $\mathbf{X}^*$. When the data is slightly corrupted PCA can perform well, however, it cannot perform well with the grossly corrupted data. Therefore, a more robust matrix completion approach against high-dimensional noise $\mathbf{E}$ is required for link prediction in real complex networks. Hence, we apply the robust principal component analysis (robust PCA) in the matrix completion for link prediction.

Mathematically, according to the theory of robust PCA, recovering matrix $\mathbf{X}^*$ can be transformed into the following optimization problem:

$$\min_{\mathbf{X}^*,\mathbf{E}} \operatorname{rank}(\mathbf{X}^*) + \gamma ||\mathbf{E}||_0 \quad \text{subject to} \quad \mathbf{X}^* = \mathbf{A} - \mathbf{E}, \tag{4}$$

where $\operatorname{rank}(\mathbf{X}^*)$ denotes the rank of matrix $\mathbf{X}^*$, the operator $||.||_0$ is the l_0-norm (*i.e.*, the number of nonzero entries of a matrix), and γ is the parameter balancing these two terms. Normally, a precise solution of $\mathbf{X}^*$ guarantees that $\mathbf{G} = \mathbf{G}^*$, which means that the precise solution of $\mathbf{X}^*$ can be used to perfectly recover the original network. Finding

the precise solution of $\mathbf{X}^*$ in eq. (4) is a highly nonconvex optimization problem and its complexity is nondeterministic polynomial. However, some approximate solutions with exponential time complexity can be obtained based on robust PCA [23]. Firstly, since a matrix with rank r has exactly r nonzero singular values, rank($\mathbf{X}^*$) is just the number of nonzero singular values of the matrix $\mathbf{X}^*$. Secondly, according to the pioneering works [24,25], the solution of the l_1-norm is also a sparse solution of the l_0-norm. Hence, the tightest relaxation of rank($\mathbf{X}^*$) and l_0-norm are the nuclear norm and l_1-norm, respectively [26–28]. In a word, the relaxed approximate solution of eq. (4) can be written as

$$\min_{\mathbf{X}^*,\mathbf{E}} ||\mathbf{X}^*||_* + \lambda||\mathbf{E}||_1 \quad \text{subject to} \quad \mathbf{X}^* = \mathbf{A} - \mathbf{E}, \tag{5}$$

where $||.||_*$ denotes the nuclear norm (*i.e.*, the sum of singular values) of the matrix, $||.||_1$ is the l_1-norm (*i.e.*, the sum of the absolute values of the matrix entries), $\mathbf{E}$ is a sparse matrix (*i.e.*, most of its entries are zeros) and λ is the positive weighting parameter balancing the low-rank property and sparsity.

On the one hand, the backbone network $\mathbf{X}^*$ contains the predicted links not in $\mathbf{A}$; on the other hand, it also eliminates some possible links in $\mathbf{A}$. After obtaining $\mathbf{X}^*$ we check only the newly appearing links and ignore the observed links in $\mathbf{A}$, as shown in eq. (3), then we merge $\mathbf{X}$ with $\mathbf{A}$ to recover a matrix $\mathbf{G}$ as illustrated in eq. (2). This matrix is recovered from the observed data $\mathbf{A}$ through the above procedure, and it is supposed to be close to the original network $\mathbf{G}^*$.

Each pair of vertices (*e.g.*, x and y) in $\mathbf{G}$ is bundled with a score S_{xy} corresponding to nonzero entries of $\mathbf{X}$. The scores in $\mathbf{G}$ or $\mathbf{X}$ denote the likelihoods of missing or newly emerging links such that the higher they are, the more chances they have to be the missing or predicted links. It is worth noting that the above approach can also be applied to solve the link prediction problem in directed networks [29]. Finally, we sort the score of unobserved links in a descending order and select the top L links. In this work, L is the cardinality of the probe set. We check whether each of these L links really appears in the probe set and record the number of appearing links as L_r. As we set the L as the cardinality of the probe set, the precision value is also equal to the recall value at this point [8] as

$$Pr = L_r/L. \tag{6}$$

Analysis. – One crucial question is: to what extent can we predict the missing links by utilizing the above matrix completion method? In [26], the authors proved that when the m observed entries of an $n \times n$ matrix with rank r satisfy the following inequality,

$$m \geq Cn^{1.2}r\log(n), \tag{7}$$

where C is a positive constant, one can perfectly recover all entries of the matrix with a very high probability through solving a simple convex optimization problem. However, for the real-world data, the adjacency matrix is very sparse where the order of the number of nonzero entries is normally much less than $n^{1.2}r\log(n)$. Fortunately, for the link prediction problem, it is not required to recover all the nonzero entries of the adjacency matrix, since only a small portion of these zero entries are the missing links and the rest of zero entries are the null links. Therefore, we are still able to estimate that the missing and likely existing links even with the nonzero entries are much less than what is required for eq. (7).

The time-consuming part of the proposed algorithm (LR) is to compute the singular value decomposition (SVD) of the adjacency matrix. By utilizing PROPACK [30] the complexity of each iteration of the algorithm reduces from $O(n^3)$ to $O(kn^2)$, where k is the estimated rank of the matrix. This is due to the low-rank property as $k \ll n$ which makes the LR scalable for large networks. For instance, LR takes only about 141.986 seconds to work with a router network containing 5022 nodes and 6258 links on normal Intel(R) Core (TM) i7-6700 PC with 8 GB of RAM[1].

The LR method contains a parameter λ which plays the role to balance the low-rank property of the recovered matrix and sparsity of the noise or spurious link matrix. The optimal value of λ can be obtained from the empirical simulation as it varies according to the sparsity and structure of the networks. The optimal λ of LR for each network in this study as well as the optimal parameters for other global methods are illustrated.

As LR and other global similarity-based algorithms prefer the global structures of the networks, we define a global structure of the adjacency matrix such that

$$g = \frac{D}{\tau}\log(|V|), \tag{8}$$

where D is the density of the network, τ is the ratio between the rank and the dimension value of the adjacency matrix, and $|V|$ is the cardinality of the node set, *i.e.*, the number of nodes. If D is large and τ is small, then the network is homogenous, *i.e.*, its rank is much smaller than it dimension. The logarithmic function on $|V|$ is to normalize the scale of the network. Large networks have more global information for the network structures. Hence, the performances of the global method should have correlation with the global structure. The Pearson correlation coefficients between the precision and the global structure g are displayed in table 3.

Simulation. – We implement LR, the local similarity-based algorithms, quasi-local, and global similarity-based algorithms on the 18 real networks including 16 unweighted and 4 weighted networks in which three of the weighted networks are the same as those in unweighted

[1]See supplementary material at `http://labcomplex.org/data/LR_LP`.

Table 1: The topology of the twelve real networks. $|V|$ and $|E|$ are the number of vertices and links, respectively. C, r and $\langle k\rangle$ are cluster coefficient, assortative coefficient and average degree, respectively. $H = \frac{\langle k^2\rangle}{\langle k\rangle^2}$ is the degree heterogeneity of the networks computed. R, τ and D are the rank of the adjacency matrix, the ratio between rank and dimension and the network density, respectively. LCP-corr is the correlation between CN and LCL. q is the ratio between links with at least one common neighbor and the total number of links.

Networks	$\|V\|$	$\|E\|$	C	r	$\langle k\rangle$	H	R	τ	D	LCP-corr	q
Jazz	198	2742	0.618	0.02	27.697	1.395	198	1.000	0.1406	0.948	0.805
Yeast	2375	11693	0.306	0.45	9.850	3.474	1816	0.765	0.0042	0.969	0.815
Political blogs	1222	19021	0.320	−0.22	27.355	2.970	1093	0.894	0.0224	0.929	0.959
Hamster	300	2503	0.201	−0.082	16.687	1.955	281	0.937	0.0558	0.899	0.860
Router	5022	6258	0.012	−0.14	2.493	5.502	3054	0.608	0.0005	0.807	0.134
FWF	128	2106	0.335	−0.10	32.422	1.231	124	0.969	0.2553	0.912	0.993
World trade	80	875	0.753	−0.39	21.875	1.558	79	0.988	0.2769	0.994	1.000
Contact	264	2108	0.658	−0.48	15.970	3.546	82	0.311	0.0607	0.986	0.966
Metabolic	453	2025	0.646	−0.226	8.940	4.485	450	0.993	0.0198	0.956	0.991
FWM	97	1446	0.468	−0.151	29.814	1.266	95	0.979	0.3106	0.950	0.997
Macaca	94	1515	0.774	−0.151	32.234	1.238	85	0.904	0.3466	0.971	0.999
Karate	34	78	0.571	−0.476	4.588	1.693	24	0.706	0.1390	0.756	0.859
Dolphin	62	159	0.259	−0.044	5.129	1.327	60	0.968	0.0841	0.907	0.761
Email	1133	5451	0.220	0.078	9.622	1.942	1091	0.963	0.0085	0.854	0.776
USAir	332	2126	0.013	−0.21	12.807	4.915	274	0.825	0.0387	0.980	0.969
C. elegans	306	2148	0.647	−0.16	14.039	4.642	282	0.922	0.0460	0.906	0.945
FWE	69	880	0.067	−0.30	25.507	7.972	66	0.957	0.3751	0.960	0.994
Football	35	118	0.353	−0.18	6.743	1.608	35	1.000	0.1983	0.948	0.805

networks. These networks are i) jazz [31] —jazz musician network, the link denotes the relationship between two persons if they used to play together in the same band at least once; ii) yeast [32] —the network of protein-protein interaction; iii) political blogs (PB) [33] —the network of hyperlinks between weblogs on US politics; iv) hamster [34] —the friendship network of users of the web site hamsterer.com; v) router [35] —the router-level topology of the Internet; vi) FWF [36] —the network of predator-prey interactions in Florida Bay in the dry season; vii) world trade (WT) [37] —the network of miscellaneous manufactures of metal among 80 countries in 1994; viii) contact [38] —a contact network between people measured by carried wireless devices; ix) metabolic [39] —the metabolic network of the nematode worm *C. elegans*; x) *C. elegans* [39] —the neural network of worms; xi) FWM [40] —the food web in Mangrove Estuary during the wet season; xii) macaca [41] —cortical networks of the macaque monkey; xiii) karate [42] —the network of relationship among the members in the karate club; xiv) football [43] —the network of American football games consisting of Division IA colleges during the regular season, Fall in 2000. xv) dolphin [44] —network of bottlenose dolphins living in Doubtful Sound (New Zealand); xvi) email [45] —the network of email interchanges between the members of the University of Rovira I Virgili; xvii) USAir [46] —the air transportation network of airports; xviii) FWE [47] —the network of the predator-prey interactions of Everglades Graminoids in the wet season. The topology statistics of the eighteen networks are shown in table 1.

We conducted 100 times the simulations of LR for each network and only report the average values and standard error in this paper. We compare the precision values with six popular unweigted local similarity-based algorithms, *e.g.*, common neighbor (CN) [16], Adamic-Adar (AA) [48], resource allocation (RA) [17], local community paradigm, *e.g.*, Cannistraci-Alanis-Ravasi (CAR), Cannistraci-Adamic-Adar (CAA), and Cannistraci-resource-allocation (CRA) [49]. In addition, we compare LR with the local path (LP) index [20] which is the quasi-local method and three global methods including Katz [18], SPM [7], and LOOP [50]. The detailed results from the ten unweighted-based algorithms and LR are shown in table 2. We can see that LOOP outperforms the rest. The second best global method is SPM, while LR stands as the third. Although LOOP performs better than the others, it is not applicable to large-scale networks. On the other hand, the structure perturbation method (SPM) outperforms LR on 11 of 16 unweighted networks, however, SPM cannot deal with weighted networks. LR, in addition, is better than the hierarchical structure model (HSM) [51] and the stochastic block model (SBM) [4] in terms of computational efficiency. Among the local parameter-free methods, CRA performs best followed by CAA and RA. In terms of mean ranking, CRA is better than LP and Katz, meanwhile LP and Katz are better than RA and other local methods. This is not a surprise since these results have also been reported in others works [52–54] such that they are some of the best mechanistic parameter-free local models

Table 2: The average predicting precision obtained by 100 independent runs on the eight real unweighted networks. The probe set contains 10% of total connections. The methods with asterisks (*) are the global ones, except for LP which is quasi-global. The values in the brackets are the values of optimal parameters of the methods. The highest precisions in global and local methods are respectively shown in boldface.

Networks	LR*	SPM*	LOOP*	Katz*	LP*	RA	CRA	AA	CAA	CN	CAR
Jazz	0.610(.13)	0.674	**0.692**	0.492(.001)	0.491(.01)	0.541	**0.556**	0.528	0.531	0.507	0.518
Yeast	0.526(.14)	0.558	N/A	0.246(.001)	0.166(.10)	**0.259**	0.162	0.163	0.145	0.146	0.141
PB	0.195(.07)	0.235	N/A	0.175(.001)	0.181(.10)	0.146	**0.174**	0.169	0.173	0.171	0.172
Hamster	0.440(.10)	0.462	**0.472**	0.064(.010)	0.070(.20)	0.058	0.060	**0.063**	0.060	0.062	0.057
Router	0.115(.10)	0.159	N/A	0.022(.010)	0.101(.02)	0.006	**0.020**	0.016	0.018	0.019	0.019
FWF	0.565(.14)	0.561	**0.576**	0.103(.001)	0.307(.50)	0.077	**0.079**	0.077	0.076	0.074	0.078
WT	0.442(.12)	**0.489**	0.452	0.419(.010)	0.415(.20)	**0.438**	0.417	0.423	0.397	0.396	0.384
Contact	**0.619**(.10)	0.595	0.580	0.569(.001)	0.591(.50)	**0.562**	**0.562**	**0.562**	0.560	0.561	0.559
Metabolic	0.215(.10)	0.355	**0.365**	0.143(.010)	0.153(.30)	**0.264**	0.207	0.193	0.151	0.140	0.137
C. elegans	0.128(.10)	0.167	**0.200**	0.101(.010)	0.123(.30)	0.106	**0.119**	0.103	0.096	0.095	0.088
FWM	0.552(.14)	0.542	**0.566**	0.151(.010)	0.308(.30)	0.128	**0.134**	0.125	0.130	0.124	0.126
Macaca	0.751(.18)	0.732	**0.755**	0.566(.010)	0.630(.30)	0.516	**0.562**	0.533	0.560	0.544	0.552
Karate	0.114(.23)	0.150	**0.194**	0.149(.001)	0.153(.01)	0.134	**0.208**	0.134	0.208	0.150	0.168
Football	0.213(.17)	**0.238**	0.150	0.206(.100)	0.185(.10)	0.134	**0.171**	0.142	0.159	0.141	0.154
Dolphin	0.069(.25)	0.120	0.125	0.099(.100)	**0.133**(.10)	0.111	0.143	0.120	0.145	0.143	**0.154**
Email	0.063(.16)	**0.151**	N/A	0.133(.001)	0.133(.01)	0.154	**0.159**	0.154	0.145	0.139	0.143

Table 3: The average predicting precision obtained by 100 independent runs on the four real weighted networks. The probe set contains 10% of total connections. The values in the brackets are the values of parameters of the methods.

Networks	LR*	rWRA	rWAA	rWCN	WRA	WAA	WCN	RA	AA	CN
USAir	0.408(.10)	0.423	0.390	0.325	0.395	0.377	0.307	**0.458**	0.391	0.372
C. elegans	**0.129**(.10)	0.109	0.112	0.108	0.109	0.119	0.116	0.104	0.105	0.098
FWE	**0.532**(.10)	0.158	0.156	0.149	0.158	0.213	0.206	0.171	0.162	0.151
Football	**0.220**(.19)	0.058	0.050	0.042	0.058	0.058	0.025	0.083	0.100	0.117

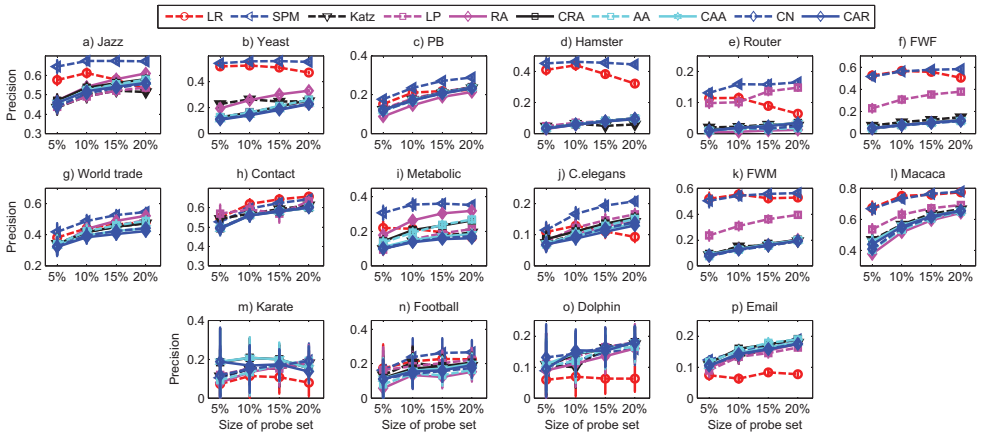

Fig. 2: (Color online) The precision values on the 16 real unweighted networks for different sizes of the probe sets. The results are obtained by 100 independent runs and the short vertical lines represent standard deviations.

for link prediction in both mono-partite and bipartite networks.

The precisions on the router network computed from all the algorithms are very low as the network is very sparse. The traditional algorithms do not perform well on yeast, hamster, FWF, and FWM networks, while LR performs much better. LR performs well on dense networks. For sparse and small networks such as karate, and dolphin, LR fails to predict the missing links. Email is quite large, but it is very sparse and its rank is high, therefore, LR cannot

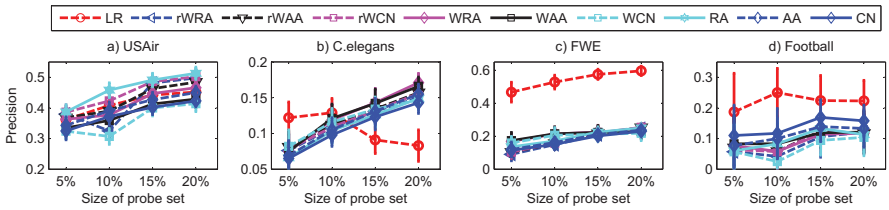

Fig. 3: (Color online) The precision values on the four real weighted networks for different sizes of probe sets. The results are obtained by 100 independent runs and the short vertical lines represent standard deviations.

Table 4: The average ranking of the different methods across all the 16 networks and Pearson correlation coefficients (CC) between precision and global structure (g), defined as in eq. (8), of the 16 unweighted networks.

	LOOP*	SPM*	LR*	CRA	LP*	Katz*	CAA	RA	AA	CAR	CN
Average ranking	N/A	2.063	3.313	4.625	4.688	5.688	6.688	6.938	6.938	7.250	7.875
CC	N/A	0.658	0.713	0.534	0.790	0.587	0.576	0.486	0.543	0.578	0.570

perform well either. The precisions of local methods on the router are very low since the ratio between links with at least one common neighbor and the total links, defined as q in table 1, of this network is also low. This condition is necessary for LCP and other local methods to predict the missing links.

We also compare LR with other six weighted-based algorithms, namely WCN (weighted CN), WAA, WRA [55], rWCN (reliable weighted CN), rWAA and rWRA [56] on four weighted networks. Whenever the weights become 1, WCN, WAA and WRA are equivalent to CN, AA and RA, respectively. Moreover, WCN, WAA and WRA take the sum of the neighbors' weights into consideration, while rWCN, rWAA and rWRA take the multiplication instead (see details in refs. [55–57]). As shown in table 3, the proposed method outperforms the others on *C. elegans*, FWE and football network. However, RA performs as the best on USAir followed by rWRA and LR. The predictions on *C. elegans* fall down when the probe sets are over 15% resulting from sparse and high-rank properties. The precision results for different sizes of the probe set on unweighted and weighted networks are shown in fig. 2 and fig. 3, respectively. The average ranking of the algorithms on the unweighted networks and the Pearson correlation coefficients between the precision and the global structure g, defined in eq. (8), are displayed in table 4.

Conclusion and discussion. – In this work, we adopt the robust principal component analysis to solve the link prediction problem. The adjacency matrix of the target network is decomposed into a low-rank matrix which can be regarded as the backbone of the network containing the true links and a sparse matrix consisting of the corrupted or spurious links in the network. Link prediction, actually, can be regarded as the matrix completion problem from the corrupted or incomplete adjacency matrix. By solving the optimization problem, we obtain the low-rank matrix which later on plays a role as score matrix illustrating the possible connectivity between each pair of vertices.

When the target network is sufficiently dense and connected, the proposed method performs better than the traditional algorithms. In other words, the local similarity-based methods do not perform well on the dense network. This indicates that the low-rank matrix recovery can well utilize the dense information in the adjacency matrix while the local similarity indices cannot. All of the networks we employ in this paper are undirected, however, the proposed method can be extended to deal with directed networks.

One disadvantage of the proposed method is the parameter λ. This parameter plays a very important role in the low rank and sparsity decomposition of the networks. In this work, we tune the parameter based on the empirical simulations to obtain the optimal values for each network. However, in the real-world problem we do not have the probe set to testify the parameter λ, but we can still divide the existing links into training set and probe set, and train the data to obtain a good value for this parameter. Although, the trained optimal value of λ from the training set may not be identical to the optimal value for the whole set of existing links, these two values should be close to each other if the data set is large.

Although LR can be used to predict the missing links in both weighted and unweighted networks, some networks, such as neural networks and signed opinion networks, should be treated as special cases. Those weights should be regularized to some positive values in advance. After prediction, one can adjust the weights in an inverse way. This kind of extension deserves further investigation, and currently we do not know how LR performs. We leave these problems to further work.

* * *

The authors thank Jian Gao, Qian-Ming Zhang, and Dr Junming Huang for useful discussions. This work was partially supported by the National Natural Science Foundation of China (NNSFC) under Grant Nos. 61433014

and 71601029 and the Fundamental Research Funds for the Central Universities No. ZYGX2014J056.

REFERENCES

[1] Albert R. and Barabási A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.
[2] Newman M. E. J., *Networks: An Introduction* (Oxford University Press, Oxford) 2010.
[3] Holme P. and Saramäki J., *Phys. Rep.*, **519** (2012) 97.
[4] Guimerà R. and Sales-Pardo M., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 22073.
[5] Getoor L. and Diehl C. P., *ACM SIGKDD Explor. Newslett.*, **7** (2005) 3.
[6] Wang W. Q., Zhang Q.-M. and Zhou T., *EPL*, **98** (2012) 28004.
[7] Lü L., Pan L., Zhou T., Zhang Y. and Stanley H. E., *Proc. Natl. Acad. Sci. U.S.A.*, **112** (2015) 2325.
[8] Lü L. and Zhou T., *Physica A*, **390** (2011) 1150.
[9] Liu Z., Dong W. and Fu Y., *Chaos*, **25** (2015) 013115.
[10] Scholz C., Atzmueller M. and Stumme G., arXiv:1407.2161 (2014).
[11] Berlusconi G., Calderoni F., Parolini N., Verani M. and Piccardi C., *PLoS ONE*, **11** (2016) e0154244.
[12] Zhu B. and Xia Y., *PLoS ONE*, **11** (2016) e0148265.
[13] Zhang P., Wang X., Wang F., Zeng A. and Xiao J., *Sci. Rep.*, **6** (2016) 18881.
[14] Barzel B. and Barabási A. L., *Nat. Biotechnol.*, **31** (2013) 720.
[15] Liu J.-H., Zhu Y.-X. and Zhou T., *Physica A*, **447** (2016) 199.
[16] Newman M. E. J., *Phys. Rev. E*, **64** (2001) 025102.
[17] Zhou T., Lü L. and Zhang Y. C., *Eur. Phys. J. B*, **71** (2009) 623.
[18] Katz L., *Psychometrika*, **18** (1953) 39.
[19] Brin S. and Page L., *Comput. Netw.*, **56** (2012) 3825.
[20] Lü L., Jin C. H. and Zhou T., *Phys. Rev. E*, **80** (2009) 046122.
[21] Liu W. and Lü L., *EPL*, **89** (2010) 58007.
[22] Candès E. J., Li X., Ma Y. and Wright J., *J. ACM*, **58** (2011) 11.
[23] Wright J., Ganesh A., Rao S., Peng Y. and Ma Y., *Advances in Neural Information Processing Systems (NIPS 2009)*, Vol. **22** (Neural Information Processing Systems Foundation, Inc.) 2009, p. 2080.
[24] Candès E. J. and Tao T., *IEEE Trans. Inf. Theory*, **51** (2005) 4203.
[25] Donoho D. L., *Commun. Appl. Math.*, **59** (2006) 797.
[26] Candès E. J. and Recht B., *Found. Comput. Math.*, **9** (2009) 717.
[27] Wright J., Yang A. Y., Ganesh A., Sastry S. S. and Ma Y., *IEEE Trans. Pattern Anal. Mach. Intell.*, **31** (2009) 210.
[28] Lin Z., Chen M. and Ma Y., Technical Report UILU-ENG-09-2215, UIUC, arXiv: 1009.5055 (2009).
[29] Zhang Q.-M., Lü L., Wang W.-Q. and Zhou T., *PLoS ONE*, **8** (2013) e55437.
[30] Larsen R. M., at `http://sun.stanford.edu/~rmunk/PROPACK/` (2004).
[31] Gleiser P. M. and Danon L., *Adv. Complex Syst.*, **6** (2003) 565.
[32] von Mering C., Krause R., Snel B., Cornell M., Oliver S. G., Fields S. and Bork P., *Nature*, **417** (2002) 399.
[33] Adamic L. A. and Glance N., *Proceedings of the 3rd International Workshop on Link Discovery*, Vol. **417** (ACM) 2005, pp. 36–43.
[34] Kunegis J., *Hamsterster friendships network dataset - KONECT*, `http://konect.uni-koblenz.de/networks/petster-friendships-hamster` (2013).
[35] Spring N., Mahajan R., Wetherall D. and Anderson T., *IEEE/ACM Trans. Netw.*, **12** (2004) 2.
[36] Ulanowicz R. E., Bondavalli C. and Egnotovich M. S., *Network Analysis of Trophic Dynamics in South Florida Ecosystem, FY 97: The Florida Bay Ecosystem*, Ref. No. [UMCES]CBL 98-123 (Chesapeake Biological Laboratory, Solomons, MD) 1998.
[37] De N. W., Mrvar A. and Batagelj V., *Exploratory Social Network Analysis with Pajek* (Cambridge University Press) 2011.
[38] Kunegis J., *Proceedings of the 22nd International Conference on World Wide Web Companion, 2013*, pp. 1343–1350.
[39] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.
[40] Baird D., Luczkovich J. and Christian R. R., *Florida Estuar. Coast. Shelf Sci.*, **47** (1998) 329.
[41] da F. Costa L., Kaiser M. and Hilgetag C. C., *BMC Syst. Biol.*, **1** (2007) 16.
[42] Zachary W. W., *J. Anthropol. Res.*, **33** (1977) 452.
[43] Girvan M. and Newman M. E. J., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 7821.
[44] Cohen J. E., Schittler D. N., Raffaelli D. G. and Reuman D. C., *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009) 22335.
[45] Guimerà R., Danon L., Díaz-Guilera A., Giralt F. and Arenas A., *Phys. Rev. E*, **68** (2003) 065103.
[46] Batageli V. and Mrvar A., at `http://vlado.fmf.uni-lj.si/pub/networks/data/mix/USAir97.net`.
[47] Heymans J., Ulanowicz R. and Bondavalli C., *Ecol. Model.*, **149** (2002) 5.
[48] Adamic L., Buyukkokten O. and Adar E., *Soc. Netw.*, **25** (2003) 211.
[49] Cannistraci C. V., Alanis L. G. and Ravasi T., *Sci. Rep.*, **3** (2013) 1613.
[50] Pan L., Zhou T., Lü L. and Hu C.-K., *Sci. Rep.*, **6** (2016) 22955.
[51] Clauset A., Moore C. and Newman M. E. J., *Nature*, **453** (2008) 98.
[52] Tan F., Xia Y. and Zhu B., *PLoS ONE*, **9** (2014) e107056.
[53] Daminelli S., Thomas J. M., Durán C. and Cannistraci C. V., *New J. Phys.*, **17** (2015) 113037.
[54] Wang W., Cai F., Jiao P. and Pan L., *Sci. Rep.*, **6** (2016) 38938.
[55] Murata T. and Moriyasu S., *Preceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2007*, pp. 85–88.
[56] Zhao J., Miao L., Yang J., Fang H., Zhang Q.-M., Nie M., Holme P. and Zhou T., *Sci. Rep.*, **5** (2015) 12261.
[57] Lü L. and Zhou T., *EPL*, **89** (2010) 18001.