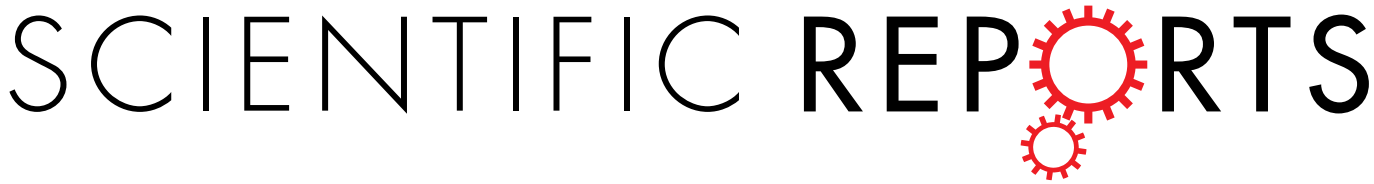

OPEN

The reconstruction of complex networks with community structure

Received: 02 April 2015
Accepted: 28 October 2015
Published: 01 December 2015

Peng Zhang[1], Futian Wang[1], Xiang Wang[1], An Zeng[3] & Jinghua Xiao[2,1]

Link prediction is a fundamental problem with applications in many fields ranging from biology to computer science. In the literature, most effort has been devoted to estimate the likelihood of the existence of a link between two nodes, based on observed links and nodes' attributes in a network. In this paper, we apply several representative link prediction methods to reconstruct the network, namely to add the missing links with high likelihood of existence back to the network. We find that all these existing methods fail to identify the links connecting different communities, resulting in a poor reproduction of the topological and dynamical properties of the true network. To solve this problem, we propose a community-based link prediction method. We find that our method has high prediction accuracy and is very effective in reconstructing the inter-community links.

Many complex systems can be naturally described by complex networks, which has largely deepened our understanding of the structure of real systems. For example, many topological properties, such as small-world[1], scale-free[2], assortativity[3], community[4] and rich club[5], have been uncovered in not only the social and technology systems we are using everyday[6–11], but also the biology systems within our bodies[12–14]. In addition, network representation is useful from practical point of view. It allows us to optimize the systems for higher functionality[15–17] and predict the future evolution of real systems[18,19]. Link prediction is one of these significant research problems[20]. It aims to estimate the likelihood of the existence of a link between two nodes, based on observed links and nodes' attributes in a network. With this problem solved, a large amount of cost in lab experiment for identifying the missing data could be reduced[20].

Link prediction methods assume that similar nodes are those that have similar connectivity patterns. Therefore, the essential problem in link prediction is to objectively estimate the similarity between nodes[21]. Up to now, many similarity metrics on link prediction have been proposed. The most straightforward method is the so-called Common Neighbor index which directly computes the number of overlapped neighbors between two nodes to determine their similarity[22]. This index, though simple, has many shortcomings. It is strongly biased to the large degree nodes and it works poorly in sparse networks. To solve these problems, many other methods, such as Jaccard[23], Resource Allocation[24], Local Path methods[25] etc, are designed. Recently, some attention has also been paid to study link prediction in weighted[26,27], directed[28,29], bipartite[30,31] networks. Moreover, some link prediction methods have been introduced to detect the spurious connections in complex networks[32].

In order to quantify the quality of link prediction, the index called *area under the receiver operating characteristic curve* (AUC) is usually used[33]. In practice, it calculates the probability that a true link has a higher link prediction score than a nonexisting link. In the case of predicting missing links, the predicted links need to be added to the observed networks to obtain the reconstructed networks[20]. The AUC index can only reflect the fraction of corrected links added to the network, but cannot capture whether the reconstructed network has the same or similar structural and dynamical properties as the true

[1]School of Science, Beijing University of Posts and Telecommunications, Beijing 100876, P.R. China. [2]State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China. [3]School of Systems Science, Beijing Normal University, Beijing 100875, P.R. China. Correspondence and requests for materials should be addressed to A.Z. (email: anzeng@bnu.edu.cn)

network. This is especially important in the networks with community structure[34]. It can happen in such networks that a link prediction method correctly identifies many missing links, but completely neglects those links connecting different communities. These inter-community links actually play an important role in the networks. They characterize the interactions between different clusters[35]. They are also strongly related to many global network properties such as average shortest path and the betweenness centrality[36]. Without these links, some dynamical properties such as bond percolation will be largely distorted[37].

In this paper, we apply several representative link prediction methods to reconstruct complex networks, namely to add the missing links with high likelihood of existence back to the networks. Even though large AUC is achieved, the reconstructed networks from these existing methods are found to be very different from the true networks, especially in terms of the average *betweenness* of the predicted links. This result indicates that the missing inter-community links are seldom captured by the existing link prediction methods. To solve this problem, we propose a community-based link prediction method. Our method can effectively identify the inter-community links by slightly sacrificing the prediction accuracy. The final obtained network can thus well reproduce the structural and dynamical properties of the true network.

Results

We consider an undirected network $G(V, E)$ where V is the set of nodes and E is the set of links. In link prediction, the original links E are first randomly divided into two parts: the training set (E^T) and the probe set (E^P). The training set contains 90% of the original links and the link prediction methods run on it. The probe set consists of the remaining 10% of the original links (The results of other division ratios are shown in SI). The probe set is used to test the accuracy of the link prediction methods. The accuracy is usually measured by the AUC value (see the Methods section for details), the higher the better. Besides accuracy, we consider also whether the link prediction methods can effectively recover the structural properties of the original network. Normally, the link prediction methods predict missing links by assigning each unconnected node pair a score which estimates the likelihood for each node pair to have a missing link between them. An accurate link prediction method will assign high score to the true missing links and low score to the nonexistent links. Unfortunately, for most of the existing link prediction methods, there is no obvious score gap between the true missing links and nonexistent links. Therefore, in order to reconstruct the network, one has to assume that the number of true missing links L is roughly known. In this fashion, one can add L top-ranking links in the link prediction methods to the observed network to reconstruct the predicted network. The approach is widely used in the literature[38,39]. Consistent with the previous works, we also assume that we know roughly the total number of true missing links. The L node pairs ($L = |E^P|$) with the highest score (denoted as the "predicted links") will be added to the training set E^T to obtain the reconstructed network $G'(V, E')$. A well-performed link prediction method should not only aim at achieving a high AUC value, but also make the structural properties of $G'(V, E')$ close to $G(V, E)$.

In this paper, we focus on the networks with community structure. According to the definition, the nodes within a community are densely connected while the nodes across communities are much more sparsely connected. In this kind of networks, the inter-community links are in general more difficult to be predicted. Without these inter-community links, the average shortest path length of the reconstructed networks would be much higher than the original networks, and the transportation dynamics[40] in this network would be much slower and congested in the reconstructed networks. In order to solve this problem, we propose a community-based link prediction method. We first detect the communities by using the EO algorithm[41] in the training set. Then the similarity scores between unconnected node pairs are computed by some classic local similarity measures (i.e. the CN or RA methods, see the Methods section for definitions). We also consider three global link prediction methods[32,39,42], the results are similar to those of CN and RA (see Supplementary Information (SI)). A tunable parameter $\beta \in [0, 1]$ is proposed to combine the information of communities and node similarity for link prediction. In practice, the node pairs are classified as intra-community pairs and inter-community pairs. Within each classification, the node pairs are ranked in descending order according to the similarity measures. β controls the probability that the intra-community node pairs ranked higher than the inter-community node pairs (see the Methods section for details). This method is inspired by ref. 43 but used here for a different goal. For convenience, when the method is combined with common neighbor similarity, it is called community-based CN method (CBCN). Similarly, it is called community-based RA method (CBRA) when it is combined with the resource allocation similarity. The illustration of the method is shown in Fig. 1. Like previous works[43], we adopt AUC to evaluate the accuracy of the link prediction. In addition, we propose to monitor the average edge-betweenness $\langle B \rangle$ of the predicted links (calculated by adding those predicted links to the network). If the average edge-betweenness is high, more inter-community links are predicted (For the solid evidences, see SI). In fact, measuring the average betweenness of the reconstructed network is also a good evaluation metric for this issue. Despite some quantitative difference, the results are qualitatively consistent with the results when $\langle B \rangle$ is used (see results in SI).

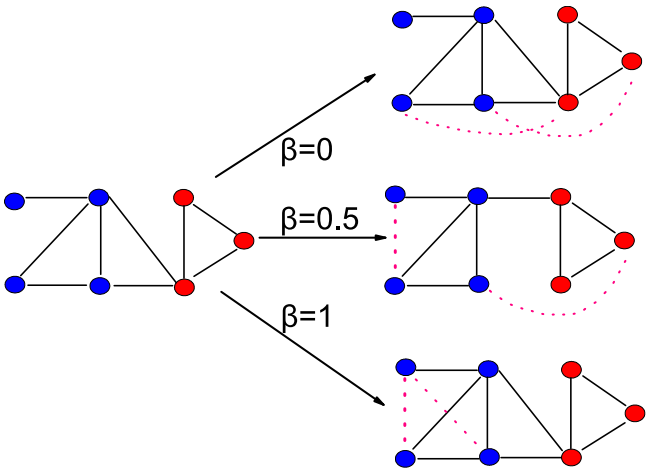

Figure 1. The illustration of the community-based link prediction method. The network on the left is the network consisting of the links in the training set. The nodes within one community are marked by the same color. The solid links represent the observed links and the dashed links stand for the predicted links. When $\beta = 0$, the inter-community missing links are ranked higher than the intra-community missing links in the prediction list. Therefore, mainly inter-community links are added to the network by the link prediction method. When $\beta = 1$, the intra-community missing links are ranked higher than the inter-community missing links in the prediction list, and mainly intra-community links are added to the network. When $\beta = 0.05$, the results are mixed, both inter- and intra-community missing links are added to the network. The similarity measure used in this toy network is CN.

We first test our method in a classical artificial network: GN-benchmark network[35] which is widely used in the research of community structure. In the GN-benchmark network, $n = 128$ nodes equally distribute in 4 communities, and each node has on average $k_{in} + k_{out} = 16$ links where k_{in} is the average number of neighbors within the same community ($8 \leq k_{in} \leq 15$) and k_{out} is the average number of neighbors between different communities ($1 \leq k_{out} \leq 8$). As k_{in} increases, the community structure of network becomes clear. Given an observed network, the obtained similarity score between nodes is deterministic if CN and RA similarity measurements are applied. However, the community detection algorithm has randomness. Therefore, there is some stochasticity in the link prediction process coming from the community detection algorithm. In this paper, we use the extremal optimization (EO) algorithm to detect communities. As stated in ref. 41, the performance of this algorithm is rather stable. Therefore, the stochasticity of the link prediction process is expected to be relatively small. We perform several times of realizations and find that the variance is much smaller than the mean value. Therefore, we mainly report the results of the mean value of different realizations.

In Fig. 2, we show the dependence of AUC and $\langle B \rangle$ on β under different k_{in}. The CBCN and CBRA are used in Fig. 2(a–d), respectively. One can see that AUC increases with β, indicating that the links within the communities are easier to be predicted. The results of CBCN and CBRA are similar and the increment of AUC is more significant when the community structure is more obvious (i.e. larger k_{in}). This result is consistent with a recent finding in ref. 43. In Fig. 2(a,b), the dashed lines mark the AUC of the original CN and RA methods (without β to adjust the ranking of the intra- and inter-community missing links). One can see that the AUC of CBCN and CBRA can be respectively higher than the AUC of CN and RA when β is large.

In Fig. 2(c,d), it shows that $\langle B \rangle$ actually decreases with β. This is natural as a larger β means more intra-community missing links are ranked higher, thus the predicted links are mainly within communities. In Fig. 2(c,d) the dashed lines mark the $\langle B \rangle$ of the links in the probe set. Clearly, if one only considers AUC, $\beta = 1$ is the optimal solution. However, this setting of β would make $\langle B \rangle$ of the predicted links smaller than that of the true missing links. A good link prediction method should not only have high AUC but also make $\langle B \rangle$ of the predicted links close to that of the true missing links. Interestingly, we observe that when β is large, a small change in β can result in a significant decrease in $\langle B \rangle$ but little influence on AUC. This observation indicates the possibility to adjust β for a satisfactory results in both AUC and $\langle B \rangle$.

We also examine our method on four real networks: *ZK* is a social network in the zahcary karate club[44], *NS* is the largest connected component of a co-authorship network of scientists who are publishing on the topic of network science[45], *Email* is an email network of an university built by regarding each email address as a node and linking two nodes if there is an email communication between them[46], *C.elegans* is a neural network of the worm Caenorhadities elegans with each neuron as a node and each

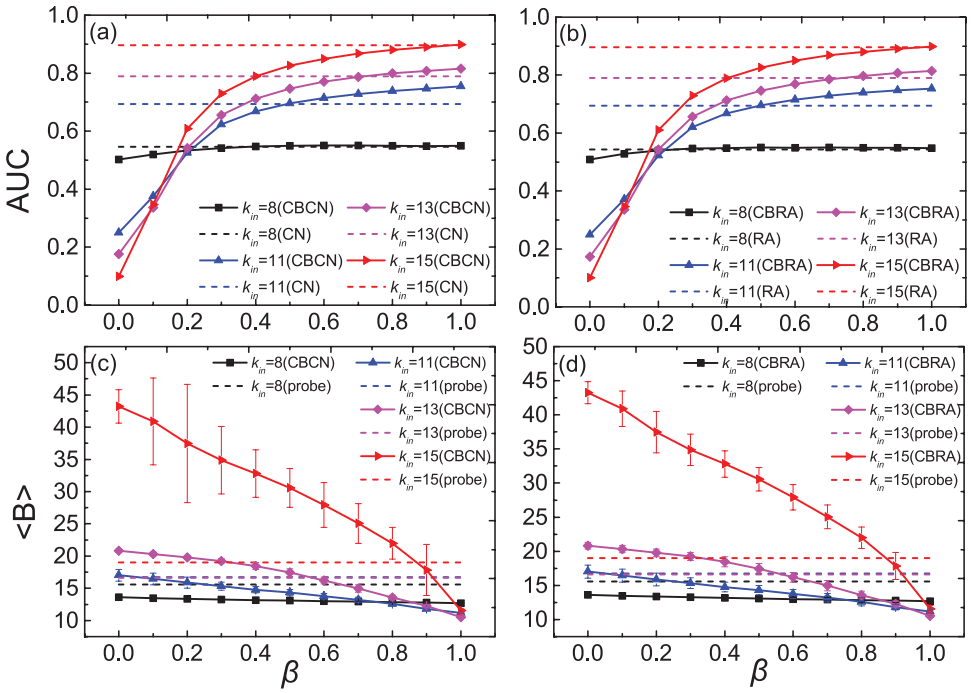

Figure 2. The influence of β on *AUC* and $\langle B \rangle$ in the GN-benchmark networks. (**a**–**d**) are the results of CBCN and CBRA, respectively. The solid lines are the results of the community-based link prediction methods (CBCN and CBRA) and the dashed lines are the results of the classic link prediction methods (CN and RA). The results are averaged over 100 independent realizations.

Network	*N*	*E*	$\langle k \rangle$	β^*		*AUC*			
				CBCA	*CBRA*	*CBCN*	*CN*	*CBRA*	*RA*
ZK	34	78	4.59	0.72	0.82	0.724	0.695	0.752	0.734
NS	379	914	4.82	0.35	0.74	0.982	0.978	0.984	0.981
email	1133	5451	9.62	0.26	0.63	0.875	0.855	0.873	0.856
C.elegans	297	2148	14.5	0.80	0.93	0.852	0.850	0.847	0.870

Table 1. Basic structural properties (network size *N*, edge number *E*, average degree $\langle k \rangle$) of the real networks, and β^* of CBCN and CBRA and *AUC* of the four methods when applied to these networks (*AUC* of CBCN and CBRA is obtained when $\beta = \beta^*$). The results are averaged over 100 independent realizations.

synapse or gap junction as a link[47]. All of these real networks are widely used in the literature and the basic structural properties of them are listed in Table 1. Here we use them to examine our methods. Figure 3 shows the performance of the community-based link prediction methods on these real networks. One can see that the results are qualitatively the same as those in the GN-benchmark networks. In these real networks, as the community structure is not as obvious as the GN-benchmark, the effect of β on AUC is even smaller, especially after $\beta > 0.1$. However, the influence of β on $\langle B \rangle$ is still strong.

We denote β^* as the β that can make $\langle B \rangle$ of the predicted links the same as that of the true missing links (i.e. the links in the probe set). Accordingly, the AUC under β^* is denoted as AUC^*. The quantitative results of β^* and AUC^* in four real networks are reported in Table 1. Clearly, the AUC^* of CBCN and CBRA can still be higher than the AUC of CN and RA, respectively.

To further understand the performance of each method, we compute the number of correctly predicted inter- and intra-links and the number of inter- and intra-links in the predicted links (results are shown in SI). We find that when the existing link prediction methods are used in GN-benchmark, the number of inter-links in the predicted links is almost zero, indicating that these existing methods tend to neglect inter-links. On the contrary, CBCN and CBRA have many inter-links in the predicted links. However, if we look at the number of correctly predicted inter-links in our methods, the number is also small. This is because the inter-links are sparsely and randomly connected in GN-benchmark (i.e. almost

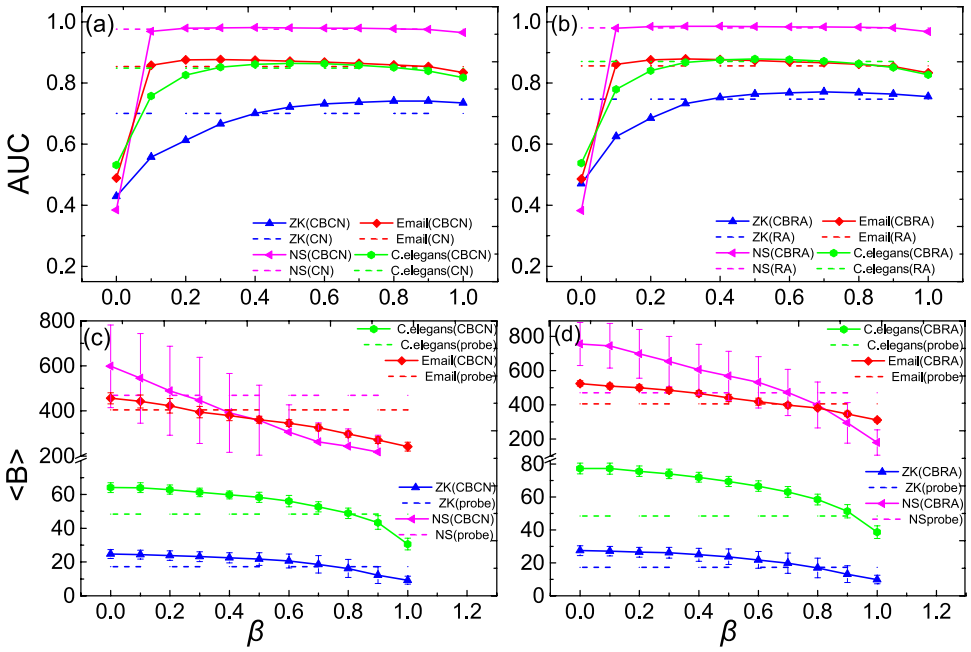

Figure 3. The influence of β on *AUC* and $\langle B \rangle$ in four real networks. (**a**–**d**) are the results of CBCN and CBRA, respectively. The solid lines are the results of the community-based link prediction methods (CBCN and CBRA) and the dashed lines are the results of the classic link prediction methods (CN and RA). The results are averaged over 100 independent realizations.

form no triangle) and it is difficult for CBCN and CBRA to capture their similarity to other links. In real networks, however, the inter-links form more triangles than thus are easier to be predicted. We test the NS real network with clear community structure (collaboration network between network scientists). We find that CN and RA can correctly predict 17.6 and 30.0 inter-links while CBCN and CNRA can correctly predict 23.7 and 31.5 inter-links (For more detailed results in NS network, see SI). These results indicates that CBCN and CNRA can respectively outperforms CN and RA in real networks as well.

In Fig. 4, we further investigate the influence of k_{in} on β^* and AUC^* in the GN-benchmark networks. In Fig. 4(a,b), one can see that β^* has an abrupt change after $k_{in} > 10$. After this value, β^* significantly increases with k_{in}. This is because when the community structure is obvious ($k_{in} > 10$), we don't have to sacrifice too much AUC and a large β can already make $\langle B \rangle$ close to the true value. In Fig. 4(c,d), we show the dependence of AUC^* on k_{in}. One can see that when k_{in} is large, AUC^* is very close to the AUC of the original CN or RA. However, when k_{in} is relatively small, AUC^* can be much smaller than AUC of CN or RA. This is because when k_{in} is small, β needs to be adjusted to a very small value in order to keep $\langle B \rangle$ of the predicted links the same as the real links (as shown in Fig. 2). In this case, a large amount of AUC needs to be sacrificed for a higher $\langle B \rangle$.

So far, we have already shown that adjusting β in the community-based link prediction methods can indeed help the methods predict more high-betweenness links in the networks. A natural question to ask at this point is how to choose β in real use. Even though β^* can be chosen at the value where $\langle B \rangle$ of the predicted links becomes the same as the real links. However, as $\langle B \rangle$ of the real links is unknown information, the above strategy seems to be an inapplicable way. To solve this problem, one has to learn the optimal β^* from the observed data. To mimic this process, we use a so-called threefold validation where a small part (usually 10% of all links) is moved from the previously introduced training set E^T to a learning set E^L[48]. The threefold validation is usually used to avoid model over-fitting in machine learning. In our case, by checking at which β the predicted links from E^T can have the same $\langle B \rangle$ as the links in E^L, one can determine the estimated optimal parameter β_e^*.

One concern for the learning process is that the missing links may largely change the structural properties. To check this, we first conduct the community detection algorithm (EO algorithm) on the original true network and denote the obtained communities as the "true detected communities". Then we randomly remove a fraction of links from the true network to obtain the observed network. We do again the community detection algorithm on the observed network and compute the fraction of nodes classified correctly by comparing the obtained communities with the so-called "true detected communities". We find that the fraction of nodes classified correctly is rather high, especially when the community

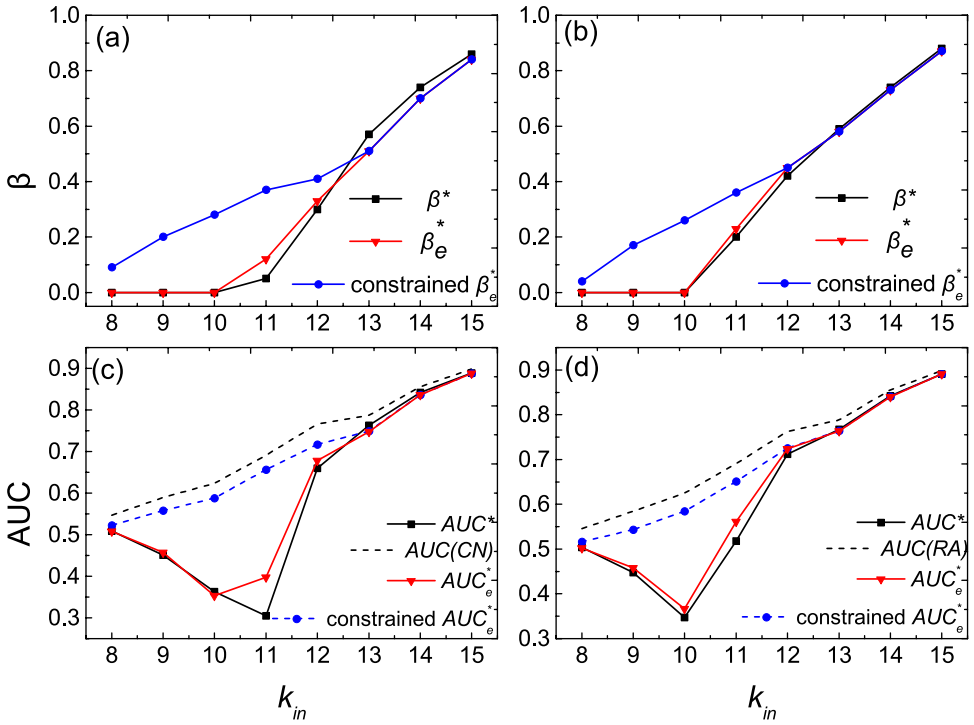

Figure 4. The influence of κ_{in} on β^* and AUC^* in the GN-benchmark networks. (**a**–**d**) are the results of CBCN and CBRA, respectively. The solid lines are the results of the community-based link prediction methods (CBCN and CBRA) and the dashed lines are the results of the classic link prediction methods (CN and RA). The results are averaged over 100 independent realizations.

structure is obvious (correct rate is over 80% when $k_{in} \geq 10$). Moreover, we compare β_e^* with β^* determined with E^P in Fig. 4(a,b). One can see that $\beta_e^* \simeq \beta^*$ at different k_{in}.

The learned optimal parameter β_e^* is then used to predict missing links based on $E^T \cup E^L$ which are then compared with entries in E^P to finally measure the link prediction accuracy AUC_e^*. The results are shown in Fig. 4(c,d). One can see that AUC_e^* is indeed close to AUC^*. As discussed above, the β^* is usually too small when $k_{in} < 13$, which directly results in a low AUC in link prediction. Therefore, we propose an additional constraint in the learning process: when determining the optimal β_e^* with the learning set E^L, we also monitor the prediction AUC of these links in E^L (denoted as AUC_{E^L}). In order to make sure the optimal β_e^* will not be too small, we assume that at most we can sacrifice 5% of the accuracy. Here, we define the AUC of the original method CN or RA as AUC_o. If before AUC_{E^L} drops to 95% of AUC_o, the predicted links can have the same $\langle B \rangle$ as the links in E^L, β_e^* is chosen as this cross-over point. If not, β_e^* is chosen as the value where AUC_{E^L} equals to 95% of AUC_o. The β_e^* obtained in this way is denoted as "constrained β_e^*". The results of the constrained β_e^* and its prediction accuracy "constrained AUC_e^*" are shown in Fig. 4 as well. So far, we have discussed three parameters: β^*, β_e^* and constrained β_e^*. A summary of these three parameters is given in Table 2. Note that even though the amount of missing links is not known, the estimation of β_e^* and constrained β_e^* will not be influenced. This is because β_e^* and constrained β_e^* are obtained from the learning process in which the amount of links in the learning set E^L is known.

Moreover, we study whether the structural and dynamical properties of the reconstructed networks from CBCN and CBRA are truly closer to the true networks. We take into account six indices, including the average shortest path of the networks ($\langle d \rangle$), clustering coefficient (C)[47], assortativity coefficient (r)[3], congestibility (D)[49], synchronizability (Q)[50] and spreading ability (μ_c)[51]. The results of different link prediction methods are listed in Table 3. The original real networks are denoted as A_0. We first randomly divide the links in A_0 to three parts: training set E^T (with 80% of the links), learning set E^L (with 10% of the links) and probe set E^P (with 10% of the links). We apply the community-based link prediction methods to compute the constrained β_e^* with E^T and E^L. Then we do $E^T \cup E^L$ to obtain a complete E^T. We apply the community-based link prediction methods with the constrained β_e^* on the complete E^T. The $|E^P|$ number of links with the highest link prediction score are then added to E^T to create the reconstructed network A^*. We also create the reconstructed networks with β arbitrarily set as 0 and 1, and denote these networks as A_1 and A_2, respectively. For comparison, the reconstructed networks with the

Parameter	Data division	Description
β^*	10% E^P, 90% E^T	determined when $\langle B\rangle$ of the top-10% ranking links equals to $\langle B\rangle$ in E^P
β_e^*	10% E^P, 10% E^L, 80% E^T	determined when $\langle B\rangle$ of the top-10% ranking links equals to $\langle B\rangle$ in E^L
Constrained β_e^*	10% E^P, 10% E^L, 80% E^T	(1) If $AUC(\beta_e^*) \geq 0.95 * AUC_o$ in E^L, constrained $\beta_e^* = \beta_e^*$. (2) If $AUC(\beta_e^*) < 0.95 * AUC_o$ in E^L, constrained β_e^* is set as the β which makes AUC equal to $0.95 * AUC_o$ in E^L

Table 2. The description of the parameters β^*, β_e^* and Constrained β_e^*. Here, AUC_o means the AUC value of the original link prediction methods such as CN or RA.

Net	properties	A_0	*CBCN*			*CN*	*CBRA*			*RA*
			A_1	A^*	A_2	A_3	A_1	A^*	A_2	A_3
ZK	$\langle d\rangle$	2.41	2.28	**2.37**	2.45	2.58	2.25	**2.40**	2.46	2.49
	C	0.571	**0.583**	0.595	0.550	0.612	0.611	0.623	**0.584**	0.668
	r	−0.476	−0.369	−0.388	**−0.438**	−0.193	−0.389	−0.430	**−0.469**	−0.204
	D	462	502	**466**	466	469	492	466	**460**	469
	Q	38.7	57.6	42.5	**40.7**	48.6	52.9	42.4	**40.9**	49.2
	μ_c	7.77	8.64	**7.77**	7.54	8.66	8.41	**7.79**	7.46	8.80
NS	$\langle d\rangle$	6.04	**6.12**	6.24	6.42	6.80	5.83	**6.16**	6.42	7.12
	C	0.741	0.654	0.668	0.667	**0.685**	0.694	**0.728**	0.713	0.724
	r	−0.0817	**0.0037**	0.0183	0.0335	0.0670	−0.1004	**−0.0834**	−0.0712	0.0485
	D	$5.7{\cdot}10^4$	$\mathbf{5.7{\cdot}10^4}$	$6.1{\cdot}10^4$	$6.2{\cdot}10^4$	$5.9{\cdot}10^4$	$\mathbf{5.7{\cdot}10^4}$	$6.3{\cdot}10^4$	$6.0{\cdot}10^4$	$5.7{\cdot}10^4$
	Q	2305	**2747**	3421	2861	3447	**2458**	3128	2787	3150
	μ_c	8.02	8.73	8.70	**8.15**	9.21	8.38	**8.07**	7.76	8.74
Email	$\langle d\rangle$	3.61	**3.63**	3.65	3.68	3.75	3.57	**3.61**	3.63	3.71
	C	0.220	**0.223**	0.232	0.238	0.233	0.327	0.358	**0.314**	0.339
	r	0.0782	0.163	0.165	**0.150**	0.238	0.0753	0.0756	**0.0782**	0.212
	D	$5.1 \cdot 10^4$	$7.7 \cdot 10^4$	$6.8 \cdot 10^4$	$5.7 \cdot 10^4$	$\mathbf{5.6 \cdot 10^4}$	$6.3 \cdot 10^4$	$5.6 \cdot 10^4$	$5.6 \cdot 10^4$	$\mathbf{5.3 \cdot 10^4}$
	Q	217	331	307	**304**	372	235	**205**	262	273
	μ_c	18.7	21.7	21.5	**20.5**	21.6	19.7	19.4	**19.1**	19.9
C.elegans	$\langle d\rangle$	2.45	**2.44**	2.47	2.49	2.53	2.40	**2.47**	2.48	2.56
	C	0.292	0.333	0.351	**0.333**	0.349	**0.369**	0.385	0.369	0.384
	r	−0.163	−0.113	−0.0980	**−0.135**	−0.0405	−0.130	−0.129	**−0.162**	−0.0428
	D	$2.6 \cdot 10^4$	$3.0 \cdot 10^4$	$2.9 \cdot 10^4$	$\mathbf{2.5 \cdot 10^4}$	$2.8 \cdot 10^4$	$3.1 \cdot 10^4$	$\mathbf{2.7 \cdot 10^4}$	$2.5 \cdot 10^4$	$2.4 \cdot 10^4$
	Q	159	176	**168**	148	195	185	**163**	151	217
	μ_c	26.1	29.7	28.2	**26.9**	31.5	29.7	27.3	**26.4**	32.1

Table 3. The properties of the reconstructed networks when different link prediction methods are applied. A_0 represents the original networks, and A_1, A^*, A_2 stand for the reconstructed networks, when $\beta = 0$, $\beta =$ constrained β_e^*, $\beta = 1$ respectively. A_3 is the reconstructed networks of the traditional methods CN and RA. ($\langle d\rangle$), C, r, D, Q, μ_c in turn, represent the average shortest path, the clustering coefficient, the assortativity coefficient, congestability, synchronizability and spreading ability of the networks. We highlight the values that are closest to the original networks in bold font. The results are averaged over 100 independent realizations.

traditional link prediction methods (e.g. CN and RA) are denoted as A_3. From Table 3, we can see that the reconstructed networks from the community-based link prediction methods (i.e. A_1, A_2 and A^*) have more similar network properties to the real network A_0 than those obtained by the traditional link prediction methods (A_3). The best results sometimes appear in A_1 and A_2. However, when A_1 is closest to A_0, A_2 is very different from A_0, and vice versa. A^* keeps a reasonable trade-off between these two methods: A^* best reproduces the network properties of A_0 in many cases; when A^* is not the best, A^* is the closest one to the best. These results confirm the importance of the parameter learning process.

Finally, we discuss the computational complexity of our method. The method is actually a combination of local link prediction algorithm and the community detection algorithm. For the local link prediction algorithm such as CN and RA, the computational complexity is $O(N * k^2)$ where N is the number of nodes and k is the mean degree of the network. In this paper, we use the extremal optimization (EO) algorithm for community detection, with computational complex $O(N^2 * lnN)$. Apparently, the computational complexity in our method is mainly determined by the community detection algorithm. If the method is applied to large networks, one can choose a faster community detection algorithm, such as the method in ref. 52 with complexity $O(N + L)$ in which L is the number of edges in the network.

Discussion

Predicting the missing or future links is a very important research topic itself and has applications in many different domains. Although many link prediction methods have been proposed in the literature, they consider all the missing links homogeneous (i.e. all the missing links are considered equally important). In this paper, we argue that in the networks with community structure, the links connecting different communities are actually of more significance and more difficult to be predicted. We propose a community-based link prediction method which allows us to predict more missing inter-community links (with high edge-betweenness) in both artificial and real networks. The results show that our method can predict more high betweenness links without losing much link prediction accuracy. As the community-based link prediction method has a parameter to tune, we propose a learning process to determine the optimal parameter. We finally apply the community-based link prediction method to reconstruct networks. The results show that the reconstructed networks by our method have very similar network properties with the real networks.

Even though our paper tries to solve a specific problem, it points out several long-neglected important issues in link prediction research: (i) Links in the network are not with equal importance. The algorithms should give priority to those important links. (ii) Prediction results should be evaluated not only by accuracy but also by how much the predicted links can recover the properties of the true network. (iii) The parameters in the link prediction algorithms should be estimated via a learning process before applied to real prediction. These issues will encourage researchers to reconsider the existing works in link prediction and may inspire a series of more effective algorithms in the future.

In this paper, we proposes an effective method to predict the inter-community links. Compared to the existing methods which all fail to predict the inter-community links (especially when the community structure is obvious), our method has a large proportion of inter-community links in the top ranking. We admit that the improved precision of these inter-community links is not high, this is because those links have a very low probability of existing. However, by including more inter-community links in the prediction list, we manage to obtain reconstructed networks with closer topological properties to the true networks. Predicting important links in networks is a scientific problem which cannot be completely solved in one paper, it surely asks for more studies in the future. Therefore, our paper raises up some important questions for future research. The method in this paper use the classic EO community algorithm to detect communities. An interesting question would be comparing the performance of different community algorithms in helping link prediction algorithms identify inter-community links. In the networks without clear community structure, the links with high edge-betweenness are still more important than the low edge-betweenness links. In these networks, the method proposed in this paper cannot be directly applied as it relies on the community detection method. Therefore, how to predict high edge-betweenness links in networks without community structure is an important extension. Finally, our study highlights the fact that the missing links are not with equal importance. Besides betweenness, the importance of links can be measured by other properties such as degree-product, clustering coefficient, link salience[53] etc. We hope the method in this paper will shed some light on designing methods to predict these kinds of important links in complex networks.

Methods

Classic link prediction algorithms. We use two representative classic link prediction algorithms in this paper: common neighbors (CN) and resource allocation (RA). After the network data is divided into the training set E^T and probe set E^P, these two methods generate the predicted links by estimating the similarity values between different node pairs in E^T. We denote the set of neighbors of node x by $\Gamma(x)$.

CN simply measures the similarity between node x and node y with the number of overlapped neighbors,

$$s_{xy} = |\Gamma(x) \bigcap \Gamma(y)|. \tag{1}$$

RA is a variant of CN. In RA, the weight of each common neighbor is negatively proportional to its degree. The similarity is thus computed as

$$s_{xy} = \sum_{z \in O_{xy}} \frac{1}{k_z}, \tag{2}$$

where k_z is the degree of node z and O_{xy} is the set of the common neighbors between x and y. After obtaining s_{xy} for each node pairs, the missing links is ranked by sorting s_{xy} in descending order.

Community detection. The community detection method in the paper is the EO method[41]. It detects communities by optimizing the modularity Q with a heuristic search. The modularity Q is defined as

$$Q = \frac{1}{2M}\sum_j q_j = \frac{1}{2M}\sum_j \left(\gamma_{c(j)} - k_j a_{c(j)}\right), \tag{3}$$

where q_j is the contribution of individual node j given a certain partition into communities. $\gamma_{c(j)}$ is the number of links node j has with nodes in the same community $c(j)$, $c(j)$ is the community which node j belongs to. k_j is the degree of node j and $a_{c(j)}$ is the fraction of links that have one or two nodes inside of the community $c(j)$. M is the number of the links in the network.

Community-based link prediction method. After computing s_{xy}, the node pairs are classified into two sets according to the community detection results: intra-community node pairs and inter-community node pairs. The node pairs in each set are ranked according to s_{xy} in descending order. The ranking list in intra-community node pairs is denoted as R_{inter} and the ranking list in inter-community node pairs is denoted as R_{inter}. The parameter β is used when R_{inter} and R_{inter} are combined. Initially, R is empty. The node pairs are then moved from R_{inter} and R_{inter} to R one by one from top to bottom. In each step, R_{inter} is picked with probability β and R_{inter} is picked with probability $1 - \beta$. For instance, if there is already n node pairs in R and in next step R_{inter} is picked, highest ranked node pair in R_{inter} is removed and placed in the $n + 1$ position in R. Note that the ranking list R_{inter} and R_{inter} become shorter and shorter while the ranking list R becomes longer and longer. The procedure is terminated if both R_{inter} and R_{inter} are empty.

Result evaluation. The results of the link prediction are evaluated by AUC and $\langle B \rangle$. AUC (area under the ROC curve) is a way to quantify the accuracy of prediction algorithms[54]. At each time, we randomly select a nonexisting link in the original network and a link in the probe set to compare their positions in R. After n times of comparison, there are n' times the probe set links have a higher rank and n'' times the probe set links have the same rank as the nonexisting links, then the AUC value is

$$AUC = \frac{n' + 0.5n''}{n}. \tag{4}$$

Besides AUC, we considered another important metric called *Precision*. It is defined as the fraction of correctly predicted links in the top-L ranking list. Here, L is set as the total number of missing links. The results are shown in SI. Despite some quantitative difference, the results of precision are qualitatively consistent with that of AUC (i.e. prediction accuracy increases with β).

$\langle B \rangle$ is defined as the average betweenness of the predicted links when they are added to the networks. The predicted links are just $|E^P|$ number of top ranking links in R. The betweenness of a link B_{ij} is defined as the ratio of the shortest paths which pass through the edge e_{ij} among all the shortest paths in the network,

$$B_{ij} = \sum_{l \neq m \neq i \neq j} \left[N_{lm}(e_{ij})/N_{lm}\right] \tag{5}$$

N_{lm} is the number of shortest routes between node l and m, $N_{lm}(e_{ij})$ is the number of the shortest paths between node l and m which pass through the edge e_{ij}.

References

1. Milgram, S. The small world problem. *Psychol. Today* **2,** 60–67 (1967).
2. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
3. Newman, M. E. Assortative mixing in networks. *Phys. Rev. Lett.* **89,** 208701 (2002).
4. Radicchi, F. *et al.* Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **101,** 2658–2663 (2004).

5. Zhou, S. & Mondragón, R. J. The rich-club phenomenon in the Internet topology. *IEEE Commun. Lett.* **8,** 180–182 (2004).
6. Amaral, L. A. N., Scala, A., Barthélémy, M. *et al.* Classes of small-world networks. *Proc. Natl. Acad. Sci. USA* **97,** 11149–11152 (2000).
7. Borgatti, S. P. *et al.* Network analysis in the social sciences. *Science* **323,** 892–895 (2009).
8. Zhao, K. *et al.* Social network dynamics of face-to-face interactions. *Phys. Rev. E* **83,** 056105 (2011).
9. Barabási, A. L., Albert, R. & Jeong, H. Scale-free characteristics of random networks: the topological of the world wide web. *Physica A* **281,** 68–77 (2000).
10. Pastor-Satorras, R., Vázquez, A. & Vespignani, A. Dynamical and correlation properties of the Internet. *Phys. Rev. E* **87,** 258701 (2001).
11. Barthélemy, M. Spatial networks. *Phys. Rep*. **499,** 1–101 (2011).
12. Barabási, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5,** 101–113 (2004).
13. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).
14. Vidal, M., Cusick, M. E. & Barabási, A. L. Interactome networks and human disease. *Cell* **144,** 986–995 (2011).
15. Newman, M. E. The structure and function of complex networks. *Siam. Rev.* **45,** 167 (2003).
16. Buldyrev, S. V. *et al.* Networks formed from failures in interdependent networks. *Nature* **464,** 1025–1028 (2010).
17. Gao, J. *et al.* Networks formed from interdependent networks. *Nat. Phys*. **8,** 40–48 (2012).
18. Albert, R. & Barabási, A. L. Statistics mechanics of complex networks. *Rev. Mod. Phys* **74,** 47 (2002).
19. Dorogovtsev, S. N. & Mendes, J. F. Evolution of networks. *Adv. Phys*. **51,** 1079 (2002).
20. Lü, L. & Zhou, T. Link prediction in complex networks: A survey. *Physica A* **390,** 1150–1170 (2011).
21. Lin, D. An information-theoretic definition of similarity. in *Proceedings of the 15th International Conference on Machine Learning*, 296–304 (Madison, Wisconsin, USA, 1998).
22. Lorrain, F. & White, H. C. Structural equivalence of individuals in social networks. *J. Math. Sociol.* **27,** 49–80 (1971).
23. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. vaudoise sci. nat.* **37,** 547–579 (1901).
24. Zhou, T., Lü, L. & Zhang, Y. C. Predicting missing links via local information. *Eur. Phys. J. B.* **71,** 623 (2009).
25. Liu, W. & Lü, L. Link prediction based on local random walk. *Europhys. Lett.* **89,** 58007 (2010).
26. Murata, T. & Moriyasu, S. Link prediction of social networks based on weighted proximity measures. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, 85–88 (Washington, DC, USA, 2007).
27. Wind, D. K. & Morup, M. Link prediction in weighted networks. In *Proceedings of the IEEE international workshop on machine learning for signal processing* 1–6 (Stander, Spain, 2012).
28. Brzozowski, M. J. & Romero, D. M. Who should I follow? Recommending people in directed social networks. In *Proceedings of the 5th international conference on weblogs and social media* 458–461 (Barcelona, Catalonia, Spain, 2011).
29. Núria, R. *et al.* Predicting future conflict between team-members with parameter-free models of social networks. *Sci. Rep.* **3,** 1999 *(*2013*)*.
30. Kunegis, J., De, Luca, E. W. & Albayrak, S. The link prediction problem in bipartite networks. *Computational intelligence for knowledge-based systems design* **6178,** 380–389 (2010).
31. Guimerà, R. *et al.* Predicting human preferences using the block structure of complex social networks. *PLoS One* **7,** e44620 (2012).
32. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* **106,** 22073 (2009).
33. Hanely, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).
34. Santo, F. Community detection in graphs. *Phys. Rep*. **486,** 75–174 (2010).
35. Girvan, M. & Newman, M. E. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **99,** 7821–7826 (2002).
36. Freeman, L. C. A set of measures of centrlity based on betweenness. *Sociometry* **40,** 35–41 (1977).
37. Wu, C. *et al.* Multiple hybrid phase transition: Bootstrap percolation on complex networks with communities. *Europhys. Lett.* **107,** 48001 (2014).
38. Stetter, O., Battaglia, D., Soriano, J. & Geisel, T. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLoS computational biology* **8(8),** e1002653 (2012).
39. Lü, L. *et al.* Toward link predictability of complex networks. *Proc. Natl. Acad. Sci. USA* **112,** 2325–2330 (2015).
40. Zheng, J. F., Gao, Z. Y. & Zhao, X. M. Properties of transportation dynamics on scale-free networks. *Physica A* **373,** 837–844 (2007).
41. Duch, J. & Arenas, A. Community detection in complex networks using extremal optimization. *Phys. Rev. E* **72,** 027104 (2005).
42. Katz, L. A new status index derived from sociometric analysis. *Psychometrika* **18,** 39 (1953).
43. Yan, B. & Gregory, S. Finding missing edges in networks based on their community structure. *Phy. Rev. E* **85,** 056112 (2012).
44. Zachary, W. W. An Information flow model for conict and fission in small groups. *J. Anthropol. Res* **33,** 452C473 (1977).
45. Newman, M. E. Finding community structure in networks using the eigenvectors of matrices. *Phy. Rev. E* **74,** 036104 (2006).
46. Guimerà, R., Danon, L., Díaz-Guilera, A., Giralt, F. & Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E* **68,** 065103 (2003).
47. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-word' networks. *Nature* **393,** 440–442 (1998).
48. Feng, C. X. J., Yu, Z. G. S., Kingi, U. & Baig, M. P. Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data. *J. Manuf. Syst.* **24,** 93–107 (2005).
49. Guimerà, R. *et al.* Optimal network topologies for local search with congestion. *Phys. Rev. Lett.* **89,** 248701 (2002).
50. Arenas, A., Díaz-Guilera, A., Kurths, J., Moreno, Y. & Zhou, C. Synchronization in complex networks. *Phys. Rep.* **469,** 93C153 (2008).
51. Boguá, M. & Pastor-Satorras, R. Epidemic spreading in correlated complex networks. *Phy. Rev. E* **66,** 047104 (2002).
52. Wu, F. & Huberman, A. finding communities in linear time: a physics approach, *Eur. Phys. J. B* **38,** 331 (2004).
53. Grady, D., Thiemann, C. & Brockmann, D. Robust classification of salient links in complex networks. *Nat. Commun.* **3,** 864 (2012).
54. Fawcett, T. An introduction to ROC analysis. *Pattern. Recogn. Lett*. **27,** 861 (2006).

Acknowledgements

This work was supported by The National Natural Science Foundation of China (Grant No. 61403037). AZ acknowledges the support from the Youth Scholars Program of Beijing Normal University (grant no. 2014NT38).

Author Contributions

P.Z., A.Z. and J.X. designed the research and wrote the manuscript. F.W. and X.W. performed the simulation. All authors analyzed the results and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Zhang, P. *et al.* The reconstruction of complex networks with community structure. *Sci. Rep.* **5**, 17287; doi: 10.1038/srep17287 (2015).