IOP PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND THEORETICAL

J. Phys. A: Math. Theor. **43** (2010) 395002 (18pp) doi:10.1088/1751-8113/43/39/395002

Iterative node matching between complex networks

Qi Xuan[1,3], **Fang Du**[2] **and Tie-Jun Wu**[2]

[1] Department of Automation, Zhejiang University of Technology, Hangzhou 310023, People's Republic of China
[2] Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, People's Republic of China

E-mail: crestxq@hotmail.com

Received 8 June 2010, in final form 26 July 2010
Published 25 August 2010
Online at stacks.iop.org/JPhysA/43/395002

Abstract

How to reveal corresponding identities of an individual in different complex systems is an ongoing problem in various areas, which can be transferred to a node-matching problem between different complex networks. In this paper, a novel iterative node-matching algorithm and two appropriate revealed pairwise matched nodes' selection strategies are proposed. It is proven that the algorithm has the approximately linear time complexity, which is much lower than that of the traditional matching algorithms in graph theory. Besides it seems that a tiny error imported in the early stage of the iterative process will not be magnified as the algorithm further proceeds. Therefore, the iterative algorithm can produce surprisingly higher matching precision on interactional scale-free networks generated by the BA model, especially when there are only a small fraction of pairwise matched nodes revealed in advance. The algorithm is finally tested by a pair of real-world networks and a feasible result is obtained.

PACS numbers: 89.75.Hc, 89.75.Da, 89.70.Eg

(Some figures in this article are in colour only in the electronic version)

1. Introduction

We live in a network world [1, 2]. For instance, our bodies contain protein networks where proteins or genes are nodes and the interactions among them are links [3–5]. Our languages can also be represented by networks where words are nodes and co-occurrence of words in the same sentences are links [6–8]. Even our friendships can be characterized by a network where nodes represent ourselves and links denote pairwise friends [9]. Such friendship network can be partially revealed with the help of advanced communication technologies, i.e. telephone

[3] Author to whom any correspondence should be addressed.

[10, 11], electronic mail [12], blog [13, 14] and so on. Interestingly, many of these complex networks in various areas present several similar topological properties [15], such as small-world [16], scale-free [17], self-similarity [18], symmetry [19], etc. In order to explain these properties, a large number of models have been proposed [3, 16, 17, 19–22].

In most cases, a network must not be isolated [23–26], that is, the individuals belonging to different complex networks may also interact with each other all along. There may be various types of interactions between different networks, and the most direct one may be caused by the various identities of the same individual in different systems [26]. For instance, an ancient protein may evolve into various homologous proteins in different species, a concept can be expressed by different words in different languages and a person may use different tools to communicate with others, and thus may have different identities in communication networks, such as telephone numbers and email addresses, etc. Although an individual may be active in different systems with different identities, in most cases the corresponding relation between them is still unknown. In fact, after a long-time evolution, there must be distinct shape differences between homologous proteins playing similar roles in different organisms or between words representing same concepts in different languages, which make it quite difficult to reveal homologous proteins or translate ancient words before the structure of protein networks or language networks is clearly presented. Similarly, due to the anonymous essence of the Internet, it is also inappropriate to say that all the identities in different communication networks correspond to the same individual just because they have similar formations.

As we can see, revealing the different identities of an individual in different systems has practical significance in many areas, e.g. revealing homogeneous proteins, auto-translating languages, inter-network filtrating information and so on. Because all these systems can be described by networks, these problems of different areas can be transferred to a common node-matching problem between different complex networks, and thus can be solved in the same framework. However, as is referred in [26], with the reason that real-world complex networks are always highly symmetric and only partially overlapped, it seems unpractical to match nodes between different networks just by considering their local topological similarities. Fortunately, the researchers of different areas can use their own dedicated methods, such as chemical, semantic and others, to reveal a part of matched nodes, although their high economical or computational cost makes it almost impossible to examine and compare each pair of nodes between different large-scale networks. It is more intuitive on the Internet where individuals may leave their email addresses when they register blog accounts, and thus several pairs of matched nodes in these networks may have been revealed beforehand. Here, it should be noted that the selection of the pairwise matched nodes which are revealed *a priori* by the dedicated methods is somewhat controllable. For example, Internet administrators can encourage (but not force) a number of selected people to provide their real-name identities, such as telephone numbers, when they register anonymous Internet accounts. Such information can help researchers better finish node matching between real-name networks and anonymous networks, which may be especially useful for the police to track the suspects in real-name networks when their illegal behaviors were first noticed in anonymous networks. Therefore, we can propose revealed pairwise matched nodes' strategies to improve the performance of node-matching selection algorithms.

Through calculating the similarities between nodes of different networks by their connections to several pairs of preliminarily revealed matched nodes, the node-matching problem between different networks can be transferred to a maximum weighted bipartite matching problem [26] and then can be solved by some well-known optimal matching algorithms in graph theory [27]. This method seems reasonable because, in many cases, an individual may behave similarly in different systems, which can be partially reflected

by the similar local structural properties in corresponding networks. However, the method fails to achieve acceptable results when there are only a relatively small number of pairwise matched nodes revealed beforehand, which, as well as its long running time, hinders its efficient application in real-world networks of quite large size. In this paper, we propose a novel iterative node-matching algorithm to solve the node-matching problem between two different networks. The whole iterative process is like a network domino. Compared with the optimal node-matching algorithm proposed in [26], the iterative algorithm proposed here has approximatively linear time complexity and can produce much better matching results on scale-free networks generated by the BA model especially when there are only a small number of pairwise matched nodes revealed beforehand.

The rest of the paper is organized as follows. In the next section, the node-matching problem is briefly reviewed and an iterative node-matching algorithm is proposed. Then in section 3, the iterative algorithm is adopted to solve the node-matching problem between scale-free networks generated by the BA model and many interesting phenomena are revealed and properly explained. In section 4, the iterative algorithm is applied to solve the node-matching problem between a pair of real-world networks, i.e. the chat network and the friendship network, obtained from the database of the *Alibaba trademanager* [28]. The paper is finally concluded in section 5.

2. The iterative node-matching algorithm

At first, we briefly introduce the node-matching problem between two different networks [26]. The two networks are denoted by $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $V_i = \left\{v_1^i, \ldots, v_{N_i}^i\right\}$ and E_i represent the node set and the link set of network i ($i = 1, 2$), respectively. Assuming there are M ($M \leqslant \min\{N_1, N_2\}$) pairs of matched nodes $v_j^1 \leftrightarrow v_j^2$ defined by $\left\{v_1^i, \ldots, v_M^i\right\} \subseteq V_i$ ($i = 1, 2$) and P_r ($P_r < M$) pairs of them have been already revealed, named as revealed matched nodes and denoted by $\left\{v_1^i, \ldots, v_{P_r}^i\right\} \subset V_i$ ($i = 1, 2$), the purpose is to design a method to find the other $M - P_r$ pairs of matched nodes in the two different networks. If P_c ($P_c \leqslant M - P_r$) pairs of them are revealed correctly by some algorithm, the matching precision ϕ can be calculated by

$$\phi = \frac{P_c}{M - P_r}. \tag{1}$$

We introduced an optimal node-matching algorithm in [26] to solve this problem, where the similarity between each pair of unmatched nodes belonging to two different networks was first calculated by numerating their connections to several pairs of preliminarily revealed matched nodes, and thus a weighted bipartite graph (a weighted link represents the similarity between a pair of nodes) was created. Then, the problem was transferred to a traditional maximum weighted bipartite matching problem and was solved by the KM algorithm in graph theory. This method has academic value but cannot be practically applied due to its low matching precision and long running time.

In this paper, we mainly focus on proposing a novel and far more efficient iterative node-matching algorithm with only approximately linear time complexity to solve this problem. Here, the similarity between a pair of nodes belonging to different networks is defined and calculated as before, and the difference is that, at each time, only a pair of unmatched nodes with the largest similarity is selected as a pair of matched nodes which is further considered as a pair of newly revealed matched nodes to recalculate the similarities between the remaining pairwise unmatched nodes, and so forth, until some terminal conditions are satisfied. Correspondingly, the degree-based revealed pairwise matched nodes' selection strategies proposed in [26] also

need to be improved. Specially, the iterative algorithm is composed of the following four steps.

(1) *Revealed pairwise matched nodes' selection.* Due to the fact that the interactional networks under study are usually not completely identical, it seems unpractical to match nodes between different networks just by their local structural properties. Therefore, a few pairs of matched nodes would be better revealed as references before the node-matching algorithm is implemented. How to select proper revealed matched nodes to improve the performance of the algorithm is always a big challenge. Because the matched nodes are supposed unknown beforehand in reality, the selection process has to be divided into two steps: first select a subset of nodes in one network by some of their local structural properties, and then reveal their matched nodes in the other network by some more dedicated methods. Generally, in a network, a node with more neighbors can provide more structural information; therefore, as was proven in [26], higher matching precision can always be obtained by selecting those nodes with larger degree in one network and their matched nodes in the other network as the revealed matched nodes. Because the structural properties of the two networks may be different when the interaction between them is not symmetric, a node of large degree in one network may be matched to a node of small degree in the other network. As a result, different matching results can be expected by adopting the large degree priority selection strategy in the different networks. Therefore, we proposed two large degree priority selection strategies: one is first selecting large degree nodes in G_1 and the other is in G_2. The difference between them is theoretically and experimentally studied in [26] when the interaction between the two networks is not symmetric. While for an iterative node-matching algorithm, the revealed pairwise matched nodes would better be centralized to a local world in the networks so as to improve the matching precision in the first round, then the second round and so on. And correspondingly, here we propose two improved revealed pairwise matched nodes' selection strategies specially for the iterative node-matching algorithm. The first one is *centralized large degree priority in G_1 (CLDP1).* That is, a set R_1 ($|R_1| = P_r$) of nodes in the network G_1 are picked up according to their degrees by the following process. The node of the largest degree in G_1 is firstly selected as the only member of R_1. Denoting the neighbor set of R_1 as U_1 ($U_1 \cap R_1 = \emptyset$), i.e. each node in U_1 (but none of the nodes in $V_1 \backslash (U_1 \cup R_1)$) is at least connected to one node in R_1, at each time the nodes in $V_1 \backslash R_1$ are sorted by the number of neighbors belonging to U_1 in descending order and the top one is selected to join in R_1. Update R_1 and U_1 and repeat the selecting process until the set R_1 contains exactly P_r nodes. Then the set R_1 of nodes in G_1 as well as their matched nodes in G_2 are selected as the revealed pairwise matched nodes. A simple example is shown in figure 1. The second one is *centralized large degree priority in G_2 (CLDP2)* which is parallel with the first one.

(2) *Similarity calculation.* The similarity between two nodes belonging to different networks can be measured by the number of revealed pairwise matched nodes around them. Specially, the similarity [26] between nodes v_i^1 and v_j^2 can be calculated by

$$S\left(v_i^1, v_j^2\right) = \frac{n_M\left(v_i^1, v_j^2\right)}{n_L\left(v_i^1\right) + n_L\left(v_j^2\right) - n_M\left(v_i^1, v_j^2\right)}, \tag{2}$$

where $n_M\left(v_i^1, v_j^2\right)$ denotes the number of revealed pairwise matched nodes $\left(v_k^1, v_k^2\right)$ that the nodes v_i^1 and v_j^2 are mutually connected to, i.e. v_i^1 is connected to v_k^1 and v_j^2 is connected to v_k^2, in the corresponding networks, and $n_L\left(v_i^1\right)$ (or $n_L(v_j^2)$) represents the total number of nodes connected to the node v_i^1 (or v_j^2) in the network G_1 (or G_2). Equation (2) guarantees

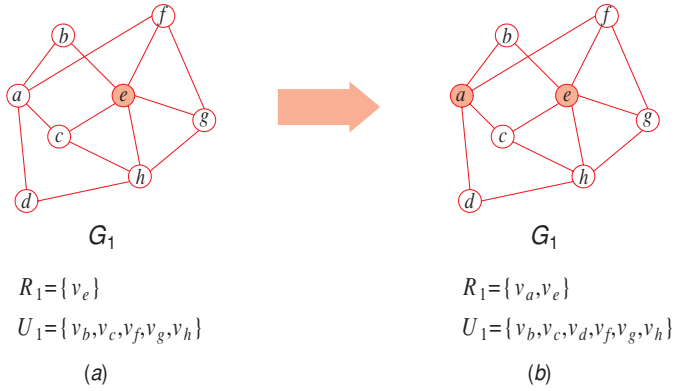

Figure 1. The sketch map for CLDP1 which aims to select a set R_1 ($|R_1| = P_r = 2$) of nodes in the network G_1 with the node set $V_1 = \{v_a, v_b, v_c, v_d, v_e, v_f, v_g, v_h\}$. (*a*) Initially, the node of the largest degree in G_1, i.e. the filled node v_e, is selected as the only member of R_1. Then the neighbors of v_e are grouped as the neighbor set of R_1, denoted by $U_1 = \{v_b, v_c, v_f, v_g, v_h\}$. (*b*) The nodes in $V_1 \backslash R_1$ are sorted by the number of neighbors belonging to U_1 in the descending order, i.e. $v_a(3, \{v_b, v_c, v_f\} \subset U_1)$, $v_g(2, \{v_f, v_h\} \subset U_1)$, $v_h(2, \{v_c, v_g\} \subset U_1)$, $v_c(1, \{v_h\} \subset U_1)$, $v_d(1, \{v_h\} \subset U_1)$, $v_f(1, \{v_g\} \subset U_1)$, $v_b(0)$ and the top one v_a is selected to join in R_1, i.e. $R_1 = \{v_a, v_e\}$. Then the neighbors of v_a and v_e are grouped as the neighbor set of R_1, denoted by $U_1 = \{v_b, v_c, v_d, v_f, v_g, v_h\}$. Thus v_a and v_e as well as their matched nodes in G_2 (not shown) are selected as the revealed pairwise matched nodes in this example.

that the similarity between two nodes of different networks has the normalized value in [0, 1].

(3) *Node matching.* At each time, a pair of unmatched nodes belonging to different networks with the largest similarity are selected as a pair of matched nodes. Then this pair of matched nodes are considered as a pair of newly revealed matched nodes and turn to step 2 to recalculate the similarities, and so forth.

(4) *Termination.* In reality, the value of M is always unknown, that is, one cannot know the exact number of pairwise matched nodes between target real-world networks *a priori*. In such a situation, a threshold $\theta \in (0, 1]$ must be provided and once the similarity between each pair of unmatched nodes is smaller than or equal to θ, the iterative node-matching algorithm is terminated. However, it is a real challenge to provide a proper threshold θ because the similarity between a pair of unrevealed matched nodes is not only determined by the topological structure of the target networks but also significantly influenced by the number of revealed pairwise matched nodes. For simplification, in this paper, the algorithm will not be terminated until all of the nodes in one target network have been matched.

Take ring networks, for example, as shown in figure 2, G_1 and G_2 are set to be identical for convenience, that is, G_2 is just a copy of G_1 and a node in G_1 and its copy in G_2 form a pair of matched nodes. In such a situation, CLDP1 and CLDP2 must be equal. Interestingly, when two pairs of matched nodes are revealed by CLDP1 or CLDP2 beforehand, all of the other pairwise matched nodes between the two ring networks can be revealed correctly no matter how many nodes these two networks contain. The matching process is very similar to a domino. A more complicated example is shown in figure 3, where G_1 and G_2 are both scale-free networks generated by the BA model and are also set to be identical. Each network has 30 nodes and 57 edges. In this example, there are three pairs of revealed

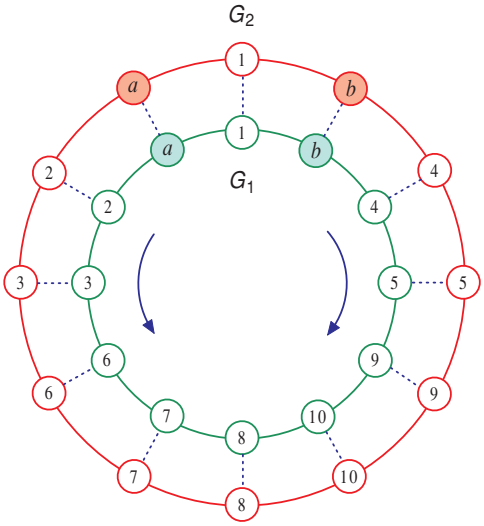

Figure 2. Node matching between ring networks G_1 and G_2 which are set to be identical for simplification. Nodes connected by a dotted line represent a pair of matched nodes. Following CLDP1 or CLDP2, two pairs of nodes denoted by filled nodes $v_1^a \leftrightarrow v_2^a$ and $v_1^b \leftrightarrow v_2^b$ have been revealed before the matching algorithm is implemented. Except these two pairs of nodes, each node in the figure is marked by a number representing the round at which it is matched. Naturally, the pair of nodes between the two revealed pairwise matched nodes are firstly matched due to their largest similarity, and then the remaining nodes are matched clockwise and anticlockwise like a domino. Interestingly, by the algorithm, two pairs of revealed matched nodes are enough to reveal all of the other pairwise matched nodes in two ring networks no matter how many nodes they contain.

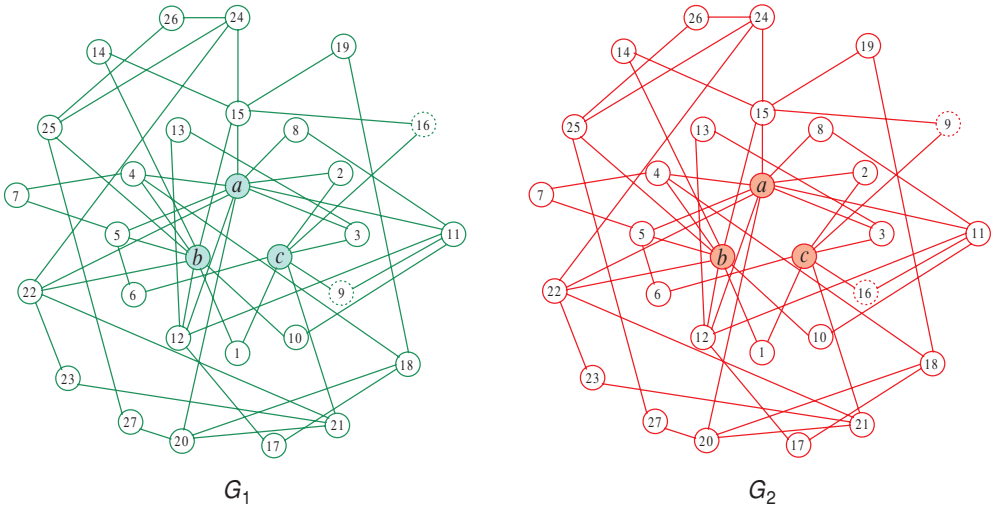

Figure 3. Node matching between a pair of identical scale-free networks. Each network has totally 30 nodes and 57 edges. Following CLDP1 or CLDP2, three pairs of nodes denoted by filled nodes $v_1^a \leftrightarrow v_2^a$, $v_1^b \leftrightarrow v_2^b$ and $v_1^c \leftrightarrow v_2^c$ have been revealed before the matching algorithm is implemented. Except these three pairs of nodes, each node in the figure is marked by a number representing the round at which it is matched. In this example, there are two pairs of wrongly matched nodes (marked by '9' and '16'); hence, the matching precision is $(27-2)/27 = 92.6\%$.

matched nodes and the matching precision is 92.6%, i.e. only two pairs of nodes are wrongly matched.

The time complexity of the algorithm mainly depends on the recalculation of the similarities in step 2. Generally, once a pair of nodes from different networks are matched at the $(\tau - 1)$st round, one need to recalculate the similarities of about $k_\tau^1 k_\tau^2$ pairs of nodes mutually connected to that pair of matched nodes at the τth round, where k_τ^i $(i = 1, 2)$ represents the degree of the matched node in G_i at the $(\tau - 1)$st round. Provided $N_1 = N_2 = M = N$, the running time of the algorithm, denoted by Γ, can be statistically calculated by

$$\Gamma \sim E\left(\sum_{\tau=1}^{N} k_\tau^1 k_\tau^2\right). \tag{3}$$

If the two networks under study are strongly dependent on each other, i.e. extremely G_1 and G_2 are identical and a node in one network can only be matched to a node of equal degree in the other network, equation (3) can be replaced by

$$\Gamma \sim \sum_{\tau=1}^{N} E\left(\left(k_\tau^1\right)^2\right). \tag{4}$$

For scale-free networks generated by the BA model, the degree distribution follows $p(k) \sim k^{-3}$; thus, equation (4) can be simplified by

$$\Gamma \sim N \int_1^N k^2 k^{-3}\,\mathrm{d}k \sim N \ln N. \tag{5}$$

However, in reality, the two target networks tend to be relatively independent although there are interactions between them, i.e. a node with large degree in one network can be matched to a node with small degree in the other network. In such a situation, equation (3) can be approximatively transferred to

$$\Gamma \sim \sum_{\tau=1}^{N} E\left(k_\tau^1\right) E\left(k_\tau^2\right) \sim N\langle k^1\rangle\langle k^2\rangle, \tag{6}$$

where $\langle k^i\rangle$ represents the average degree of the network G_i. In most cases, $\langle k^i\rangle$ can be considered as a constant; therefore, equation (6) suggests a linear time complexity $O(N)$ of the iterative algorithm.

3. Simulation and analysis

In order to compare with the optimal node-matching algorithm [26], here, we will adopt the same model to create tested interactional artificial networks. The model is introduced as follows and the parameters are set to be $N_1 = N_2 = M = N$ for convenience.

(1) *Networks initialization.* Two networks G_1 and G_2 with N nodes respectively are generated by the same rule, where all the nodes are randomly matched, i.e. N pairs of randomly matched nodes $v_i^1 \leftrightarrow v_i^2$ are provided.
(2) *Interaction.* If v_i^1 (or v_i^2) and v_j^1 (or v_j^2) is connected in G_1 (or G_2) while v_i^2 (or v_i^1) and v_j^2 (or v_j^1) is not connected in G_2 (or G_1), then connect v_i^2 (or v_i^1) and v_j^2 (or v_j^1) with probability η_1 (or η_2), as shown in figure 4.

The initial networks could be random networks, small-world networks or scale-free networks, as studied in [26]. In this paper, we only focus on the scale-free networks due to their popularity in reality. Particularly, G_1 and G_2 are generated by the BA model [17]:

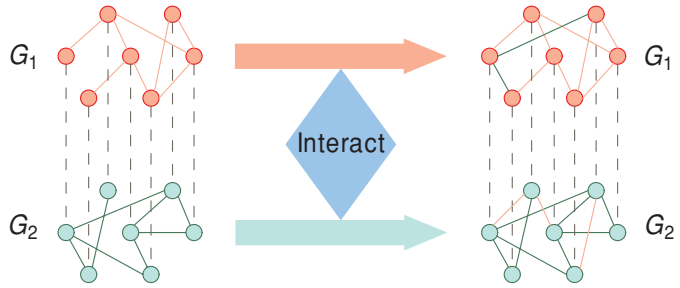

Figure 4. In the beginning, two networks G_1 and G_2 are obtained by the same model, where all the nodes are randomly matched (connected by dashed lines). Then the pair of interactional networks G_1 and G_2 are obtained by interacting with each other, i.e. two non-linked nodes in the network G_1 are connected by a line with probability η_2 if their corresponding matched nodes in G_2 are linked while two non-linked nodes in G_2 are connected by a line with probability η_1 if their corresponding matched nodes in G_1 are linked. η_1 and η_2 are named as the interactional degree.

starting with a small number m_0 of nodes connected with each other, adding a new node at every time step, and connecting it to m ($m \leqslant m_0$) different nodes which are selected with a probability linearly proportional to their degrees, then after T time steps, a scale-free network with $N = m_0 + T$ nodes is generated. Here, the parameters are set to be $m = m_0 = 4$ and $N = 500$. Then, G_1 and G_2 are randomly matched and interact with each other with different interactional degrees $\eta_1 = 0.9$ and $\eta_2 = 0.1$. The resultant networks are adopted to test the iterative node-matching algorithm.

The primary worry about an iterative algorithm is that whether a tiny error imported in the early stage will be magnified as the algorithm further proceeds so as to produce a meaningless result. Denoting $P_c(\tau)$ as the number of pairwise matched nodes revealed correctly by the algorithm in the first τ rounds, the matching precision in the first τ rounds then can be calculated by

$$\phi(\tau) = \frac{P_c(\tau)}{\tau}. \tag{7}$$

The matching precision $\phi(\tau)$ in the first τ rounds as functions of τ for different revealed pairwise matched node selection strategies CLDP1 and CLDP2 and various $P_r = 1, 2, 3, 4, 5, 8, 10, 20$ are plotted in figure 5. It is found that when $P_r \geqslant 3$ for CLDP1 and $P_r \geqslant 5$ for CLDP2, after a whistle drop in the early stage, $\phi(\tau)$ climbs up all along until the matching process is close to termination, which suggests that, in most cases, the iterative node-matching algorithm could be considered convergent, i.e. the error imported in the early stage will not be magnified boundlessly as the algorithm further proceeds.

This phenomenon can be well explained by introducing the local symmetry [19] between nodes of the same network. Denoting the number of common neighbors of two non-linked nodes v_i and v_j as χ_{ij}^c and the number of total neighbors as χ_{ij}^t, the symmetry between the two non-linked nodes in a network is defined by

$$\omega_{ij} = \frac{\chi_{ij}^c}{\chi_{ij}^t}. \tag{8}$$

If nodes v_i and v_j are connected, release the link and then calculate the symmetry between them following equation (8), as shown in figure 6(*a*) and (*b*). Naturally, it is impossible to distinguish two nodes v_i and v_j in a network with the symmetry $\omega_{ij} = 1$ (i.e. they share

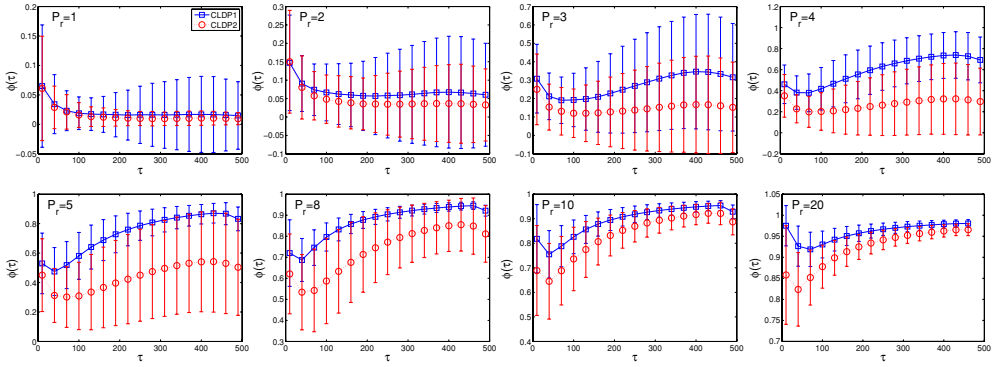

Figure 5. The matching precision $\phi(\tau)$ in the first τ rounds as functions of τ for different revealed pairwise matched nodes' selection strategies CLDP1 and CLDP2 and various $P_r = 1, 2, 3, 4, 5, 8, 10, 20$. It could be found that even for very small P_r ($P_r \geqslant 3$ for CLDP1 and $P_r \geqslant 5$ for CLDP2), after experiencing a whistle drop in the early stage, $\phi(\tau)$ climbs up all along until the matching process is close to termination. The iterative node-matching algorithm is implemented on 100 different pairs of scale-free networks generated by the BA model with the same parameters $m = m_0 = 4$ and $N = 500$.

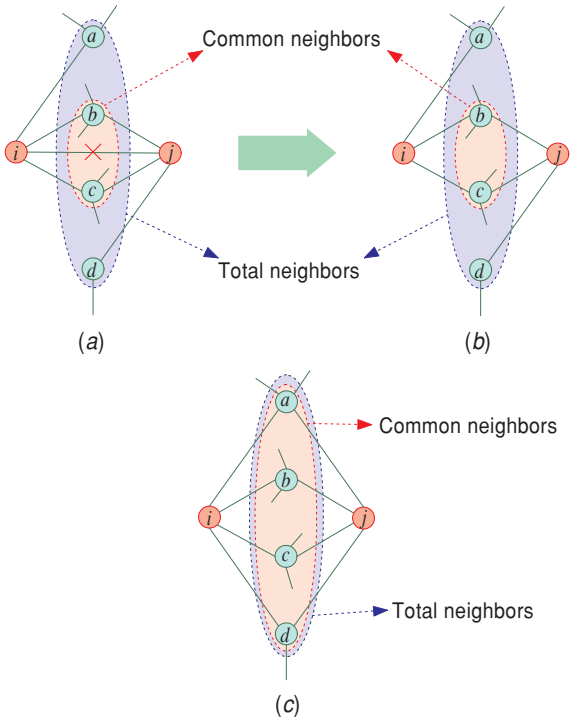

Figure 6. The symmetry of two nodes is defined by the ratio of the number of their common neighbors versus the number of their total neighbors. (*a*) If two nodes v_i and v_j are connected, (*b*) release the link and then calculate the symmetry ω_{ij} between them following equation (8). (*c*) Two nodes v_i and v_j are fully symmetric ($\omega_{ij} = 1$) when and only when they share exactly same neighbors (excluding themselves).

the same neighbors (excluding themselves), as shown in figure 6(*c*)) just by adopting their topological information. Therefore, those highly symmetric nodes in one network may be wrongly matched to the nodes in the other network with quite a high probability. However, also with the reason that the symmetric nodes are topological equal for any of the other nodes in the same network, there is little difference for the posterior matching process whether the symmetric nodes in one network were wrongly matched with each other to the nodes in the other network or not. This is the reason why the matching precision $\phi(\tau)$ can rebound after experiencing a whistle drop in the early stage.

In fact, the iterative node-matching algorithm behaves surprisingly good in revealing pairwise matched nodes between scale-free networks. Figure 7(*a*) shows that by adopting the iterative node-matching algorithm, most of the pairwise matched nodes ($\phi \geqslant 80\%$) can be revealed correctly even when there are only a very small number ($P_r \geqslant 5$ for CLDP1 and $P_r \geqslant 8$ for CLDP2) of revealed matched nodes. While at the same condition (the sample ratio $\gamma = P_r/N \in [0.01, 0.02]$) the matching precision ϕ is close to 0 by adopting the optimal node-matching algorithm [26]. Moreover, it is also shown that the matching precision obtained by CLDP1 is higher than that obtained by CLDP2 almost everywhere provided $\eta_1 > \eta_2$ because in such a situation the revealed pairwise matched nodes obtained by CLDP1 can provide more topological information for unmatched nodes [26]. Such priority is especially remarkable when the matching precision ϕ experiences a rapid ascending for the both revealed pairwise matched nodes' selection strategies, as shown in figure 7(*b*).

More interestingly, figure 7(*a*) also shows that large variance always appears in the rapid ascending process of ϕ as P_r increases, i.e. $P_r \in [2, 5]$ for CLDP1 and $P_r \in [2, 8]$ for CLDP2. Considering there are totally 100 different pairs of tested scale-free networks generated by the BA model with the same parameters $m = m_0 = 4$ and $N = 500$ in the experiment, in order to provide a more clear image, we define N_ϕ as the number of pairwise tested networks with the final matching precision in $[\psi - \delta, \psi + \delta)$ and present the relationships between N_ϕ and ϕ for $\delta = 0.05$ and different revealed pairwise matched nodes' selection strategies by two bar figures in figure 7(*c*) and (*d*), respectively. As expected, it is shown that the distribution of the matching precision ϕ tends to be polarized rather than centralized around the average matching precision, i.e. in most cases, a much higher ($\phi \geqslant 0.8$) or a much lower ($\phi < 0.1$) matching precision could be obtained by the iterative node-matching algorithm. This phenomenon may suggest that, given a pair of tested scale-free networks, there is a critical point of P_r below which the matching precision tends to 0 while above which a much higher matching precision could be obtained, and such bifurcation always comes into being in the early stage of the iterative matching process.

Denoting $s(\tau)$ as the maximal similarity between the pairwise matched nodes revealed by the algorithm at the τth iterative round, like equation (7), the average similarity of the pairwise matched nodes revealed in the first τ rounds can be calculated by

$$S(\tau) = \frac{\sum_{t=1}^{\tau} s(t)}{\tau}. \tag{9}$$

It is interesting to study the trend of $S(\tau)$ as the algorithm proceeds step by step since, in reality, the maximal similarity $s(\tau)$ is the only reference to judge the correctness of the pairwise matched nodes revealed at that round. Particularly, the relationships between $S(\tau)$ and τ for different revealed pairwise matched node selection strategies CLDP1 and CLDP2 and various $P_r = 1, 2, 3, 4, 5, 8, 10, 20$ are shown in figure 8. Similarly, after experiencing a whistle drop in the early stage, $S(\tau)$ climbs up all along until the matching process is close to termination, and this trend is especially distinct when $P_r \geqslant 3$ for CLDP1 and $P_r \geqslant 5$ for CLDP2.

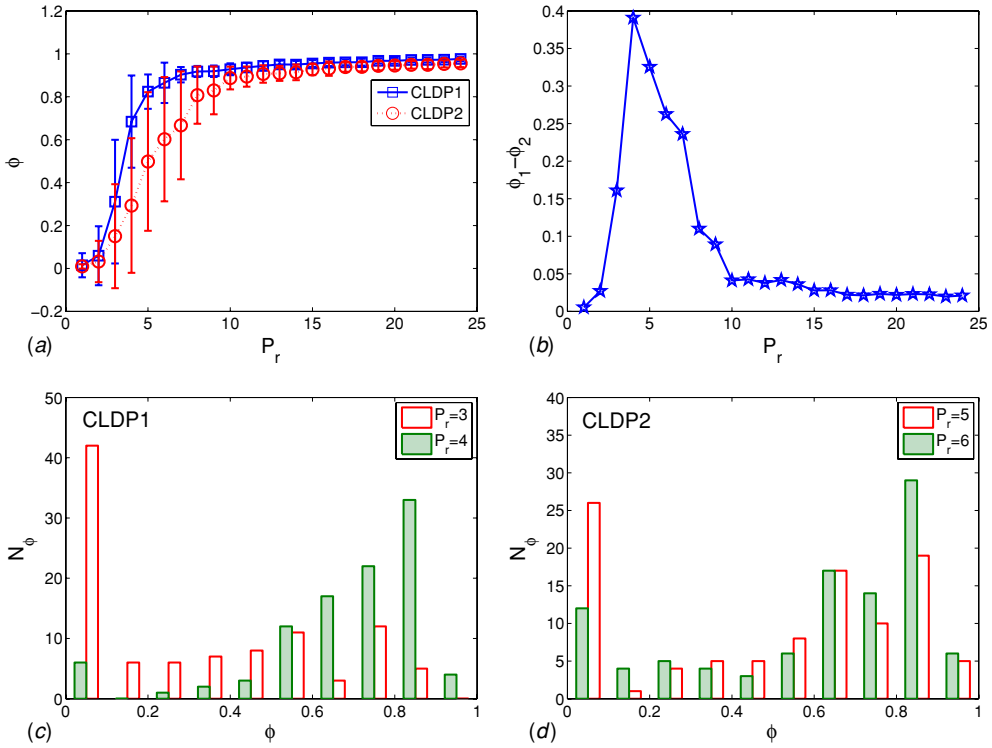

Figure 7. In the experiment, there are totally 100 different pairs of tested scale-free networks generated by the BA model with the same parameters $m = m_0 = 4$ and $N = 500$. (a) The matching precision ϕ as functions of P_r for different revealed pairwise matched nodes' selection strategies CLDP1 and CLDP2. (b) The difference between ϕ_1 and ϕ_2 as a function of P_r, where ϕ_1 denotes the average matching precision obtained by CLDP1 and ϕ_2 represents that obtained by CLDP2. (c) Denoting N_ϕ as the number of pairwise tested scale-free networks with the final matching precision in $[\phi-\delta, \phi+\delta)$, when $\delta = 0.05$, the relationship between N_ϕ and ϕ for CLDP1 is plotted. (d) The relationships between N_ϕ and ϕ for CLDP2. In the two bar figures, it could be found that the distribution of the matching precision ϕ tends to be polarized but not centralized around the average matching precision, i.e. in most cases, a much higher ($\phi \geqslant 0.8$) or a much lower ($\phi < 0.1$) matching precision could be obtained by the iterative node-matching algorithm.

The rebounding phenomenon of $S(\tau)$ is reasonable. Taking the networks shown in figure 3, for example, those pairwise nodes only mutually connected to the revealed pairwise matched nodes were firstly matched by the algorithm due to their large similarities, such as the pairwise matched nodes $v_1^1 \leftrightarrow v_1^2$ and $v_2^1 \leftrightarrow v_2^2$. As the algorithm further proceeded, the pairwise matched nodes such as $v_3^1 \leftrightarrow v_3^2$, $v_4^1 \leftrightarrow v_4^2$ and $v_5^1 \leftrightarrow v_5^2$ with larger degree and also mutually connected to the revealed pairwise matched nodes were revealed, in this period, the maximal similarity $s(\tau)$ steadily decreased. Meanwhile, the pairwise matched nodes with larger degree could provide more information for those remaining unmatched nodes, which may cause a rising of the maximal similarity $s(\tau)$, such as the pairwise matched nodes $v_6^1 \leftrightarrow v_6^2$ and $v_7^1 \leftrightarrow v_7^2$, etc. It should be noted that, since a pair of nodes are matched by their maximal similarity at each round, in some cases, there may be a small number of pairwise nodes with extra small similarities left until the algorithm is close to termination, which will certainly cause another descending of $s(\tau)$, as shown in figure 8.

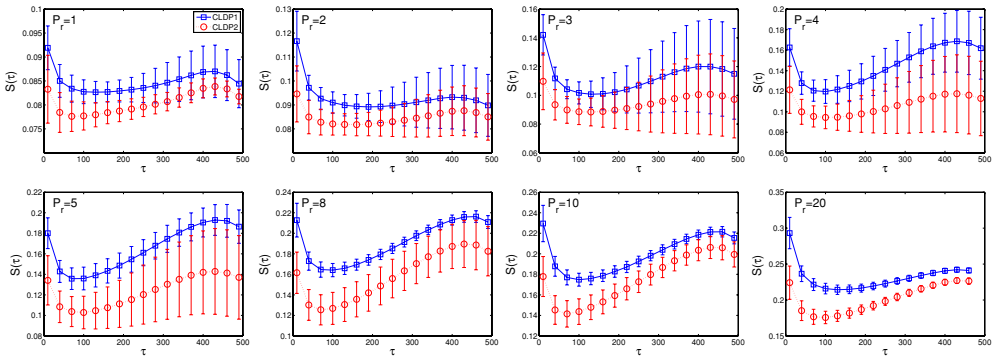

Figure 8. The average similarity $S(\tau)$ of the pairwise matched nodes in the first τ rounds as functions of τ for different revealed pairwise matched node selection strategies CLDP1 and CLDP2 and various $P_r = 1, 2, 3, 4, 5, 8, 10, 20$. Generally, more pairs of revealed matched nodes result in larger similarity of the posterior pairwise matched nodes obtained by the iterative node-matching algorithm. Meanwhile, like the matching precision $\phi(\tau)$, after experiencing a whistle drop in the early stage, $S(\tau)$ climbs up all along until the matching process is close to termination, and this trend is especially distinct when $P_r \geqslant 3$ for CLDP1 and $P_r \geqslant 5$ for CLDP2.

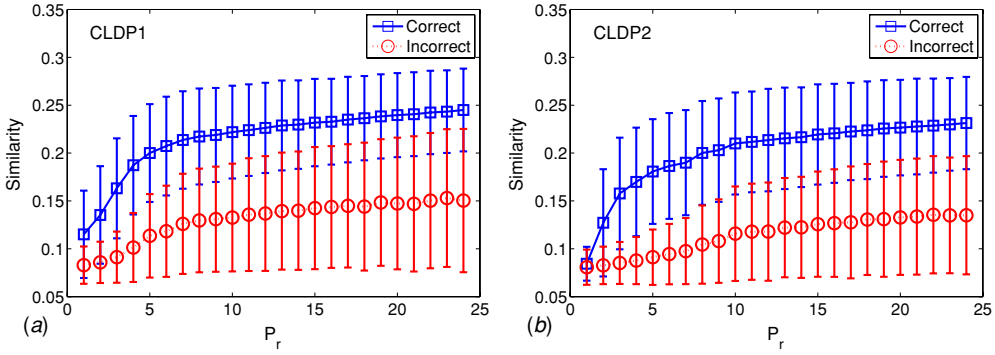

Figure 9. (*a*) The similarities of correctly pairwise matched nodes and those of incorrectly pairwise matched nodes obtained by the iterative algorithm for (*a*) CLDP1 and (*b*) CLDP2, and various values of $P_r = 1 \sim 24$. It seems that there is a distinct difference between the similarities of correctly matched nodes and those of incorrectly matched nodes.

Because $S(\tau)$ fluctuates all along, it seems imprudent to provide a static threshold $\theta \in (0, 1]$ and terminate the algorithm once each pair of unmatched nodes has the similarity smaller than θ. For example, in the experiment, although most of the correctly pairwise matched nodes have the similarity larger than 0.12 for CLDP1 and $P_r = 5$, as shown in figure 9(*a*), in most cases, the algorithm will be terminated prematurely before $\tau = 30$ with an average matching precision of those pairwise matched nodes equal to 0.54 if the threshold is set to be $\theta = 0.12$. By contrast, the average matching precision is equal to 0.82 if there is no threshold so that all of the 495 pairs of nodes are matched successively. However, a threshold of similarity may be useful if it is only used to determine which pairs of matched nodes are correct and which pairs are not afterward because there seems a distinct difference between the similarities of correctly pairwise matched nodes and those of incorrectly pairwise matched nodes, as shown in figure 9. For example, in the same experiment, more than 90%

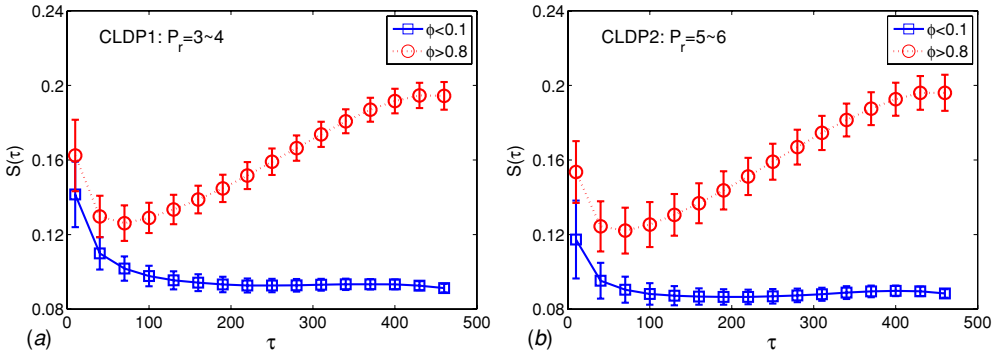

Figure 10. (*a*) In the experiment, there are totally 100 different pairs of tested scale-free networks generated by the BA model with the same parameters $m = m_0 = 4$ and $N = 500$ for CLDP1 and $P_r = 3 \sim 4$. Two groups of pairwise networks are selected by their matching precision ϕ, i.e. one group with $\phi < 0.1$ and the other group with $\phi > 0.8$. The average similarity $S(\tau)$ of the pairwise matched nodes in the first τ rounds as functions of τ for these two groups of pairwise networks are plotted respectively. (*b*) The average similarity $S(\tau)$ of the pairwise matched nodes in the first τ rounds as functions of τ for the two groups ($\phi < 0.1$ and $\phi > 0.8$) of pairwise networks generated by the BA model with the same parameters $m = m_0 = 4$ and $N = 500$ for CLDP2 and $P_r = 5 \sim 6$.

(>82%) pairwise matched nodes with similarity larger than 0.12 are revealed correctly for CLDP1 and $P_r = 5$. Moreover, figures 5 and 8 present an intuitive image that a remarkable recovering of $S(\tau)$ may indicate a relatively high matching precision, which is also validated by figure 10 where it could be found that there is always a remarkable recovering of $S(\tau)$ for those pairwise networks with the final matching precision $\phi > 0.8$ but it is not the case for the pairwise networks with the matching precision $\phi < 0.1$.

4. Test on real-world complex networks

For most of us, a familiar pair of interactional networks are the friendship network and the chat network. On one hand, in most cases, there is a natural trend that one prefers to chat with his friends or acquaintances rather than strangers, i.e. the friendship network determines the chat network to a certain extent. On the other hand, once two strangers chat with each other for some reason (e.g. common interests, curiosity, warmheart, etc), they may be friends one day in the future if they enjoy with each other, i.e. the chat network can influence the evolution of the friendship network. Therefore, one can say that the friendship network and the chat network interact with each other all the time. Fortunately, the advanced communication technologies developed these years make it possible to figure both the friendship network and the chat network on a quite large scale.

Here, we take the *Alibaba trademanager* [28] for example. Alibaba trademanager is one of the most successful Electronic Commerce Platform in China on which people can buy and sell products or services over the Internet. Each Alibaba user has a contact list on the UI of the trademanager showing his closest partners or friends, based on which the friendship network could be constructed, where nodes denote Alibaba users and two nodes are linked if one of the corresponding users appears on the contact list of the other. At the same time, one can search a potential supplier or customer through the web and make an unofficial conversation with him. The communication records then can be used to construct the chat network where two

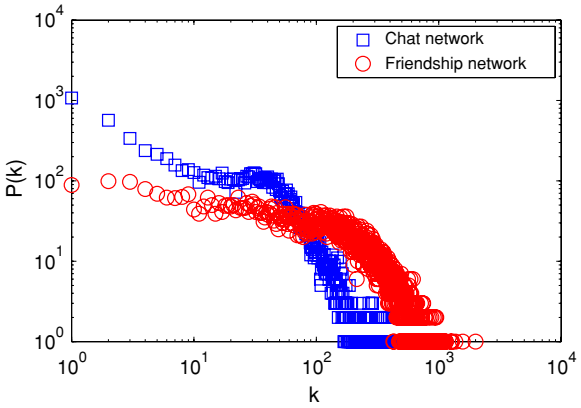

Figure 11. The degree distributions for the chat network and the friendship network obtained from *Alibaba trademanager*.

nodes are linked if there is at least a communication record between the corresponding users. We mainly focus on 14 800 employees of the Alibaba company and construct the friendship network and the chat network among them based on their contact lists and communication records in a week. Denoting the chat network by G_1 and the friendship network by G_2, they are preprocessed by the following steps.

(1) *Extract the giant cluster (GC).* Extract the GCs of G_1 and G_2, denoted by $G_1^g = \left(V_1^g, E_1^g\right)$ and $G_2^g = \left(V_2^g, E_2^g\right)$, where V_i^g and E_i^g represent the node set and the link set of the GC G_i^g, respectively.

(2) *Calculate the intersection.* A pair of matched nodes in the networks correspond to an Alibaba user. Select those users appearing in both the G_1^g and G_2^g, denoted by $V^c = V_1^g \cap V_2^g$, and get the sub-networks $G_1^c = \left(V^c, E_1^c\right)$ and $G_2^c = \left(V^c, E_2^c\right)$ where $E_i^c \subseteq E_i^g$ represents the set of links between nodes in V^c. Set $G_1 = G_1^c$ and $G_2 = G_2^c$ and terminate the preprocessing if both the networks G_1^c and G_2^c are connected, otherwise turn to step (1).

After the preprocessing, both the connected networks G_1 and G_2 have 9859 nodes and are one-to-one matched, i.e. each node in G_1 has a matched node in G_2 and vice versa. Moreover, if there is a link between two nodes in G_1, we can find a link between their matched nodes in G_2 with probability 80.8%, and the probability is 18.4% from G_2 to G_1, which means $\eta_1 > \eta_2$ for these two networks. Their degree distributions are shown in figure 11, and their basic topological properties, including the number of vertices N, the average degree $\langle k \rangle$, the average clustering coefficient $\langle C \rangle$, the average shortest path length $\langle L \rangle$ and the average symmetry $\langle \omega \rangle$ are presented in table 1. Here the average symmetry $\langle \omega \rangle$ of a network is defined by equation (10):

$$\langle \omega \rangle = \frac{\sum_{i=1}^{N} \max_j(\omega_{ij})}{N}. \tag{10}$$

By comparison, the average symmetries of the artificial network G_1 and G_2 generated by the BA model are $\langle \omega \rangle = 0.19 \pm 0.04$ and $\langle \omega \rangle = 0.15 \pm 0.03$, respectively.

The matching precision $\phi(\tau)$ in the first τ rounds as functions of τ between the chat network and the friendship network for different revealed pairwise matched node selection strategies CLDP1 and CLDP2 and various $P_r = 100, 200, 300, 400, 500, 800, 1000, 2000$ are

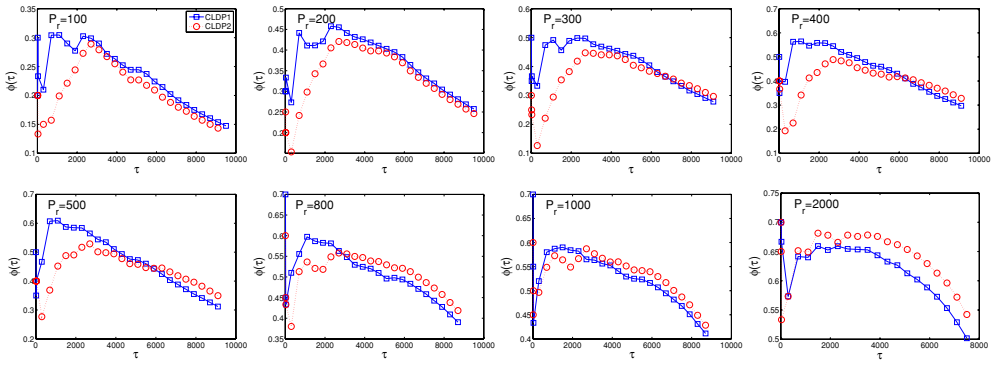

Figure 12. The matching precision $\phi(\tau)$ in the first τ rounds as functions of τ for different revealed pairwise matched node selection strategies CLDP1 and CLDP2 and various $P_r = 100, 200, 300, 400, 500, 800, 1000, 2000$. Different from figure 3, here it could be found that $\phi(\tau)$ drops steadily after it comes through its peak at about $\tau = 2000$.

Table 1. The basic properties, i.e. the number of vertices N, the average degree $\langle k \rangle$, the average clustering coefficient $\langle C \rangle$, the average shortest path length $\langle L \rangle$ and the average symmetry $\langle \omega \rangle$ for the chat network and the friendship network obtained from *Alibaba trademanager*.

Networks	N	$\langle k \rangle$	$\langle C \rangle$	$\langle L \rangle$	$\langle \omega \rangle$
Chat	9859	39.4	0.218	3.37	0.321
Friendship	9859	172	0.313	2.55	0.331

shown in figure 12. In this figure, one can see that $\phi(\tau)$ drops steadily after it comes through its peak at about $\tau = 2000$, which is a little different from the results on the pairwise interactional scale-free networks generated by the BA model, as shown in figure 5, where $\phi(\tau)$ climbs up all along until the matching process is close to termination. As a result, the final matching precision between the pair of real-world networks is much lower than that between the artificial networks generated by the BA model when adopting the same proportion of pairwise revealed matched nodes. This phenomenon may be caused by the relatively high symmetry of the chat network and the friendship network. For example, the matching precision between a fully connected network G_1 ($\langle \omega \rangle = 1$) and a whatever network G_2 must be close to zero because, in such a situation, equation (11) must be satisfied for each group of nodes $\{v_i^1, v_j^1, v_k^2\}$:

$$S(v_i^1, v_k^2) = S(v_j^1, v_k^2). \tag{11}$$

Besides, in figure 13, one can find that CLDP2 behaves better than CLDP1 when $P_r > 200$ even though $\eta_1 > \eta_2$, which may be attributed to the much larger average degree of the friendship network (G_2) over that of the chat network (G_1) [26].

At each time, a pair of nodes belonging to different networks are matched due to their maximum similarity at that round. Denoting $P(s > S)$ as the number of pairwise matched nodes with their matching similarity larger than S, where $P_c(s > S)$ pairs of them are matched correctly, the matching precision for those pairwise matched nodes with similarity $s > S$ can be calculated by equation (12):

$$\phi(s > S) = \frac{P_c(s > S)}{P(s > S)}. \tag{12}$$

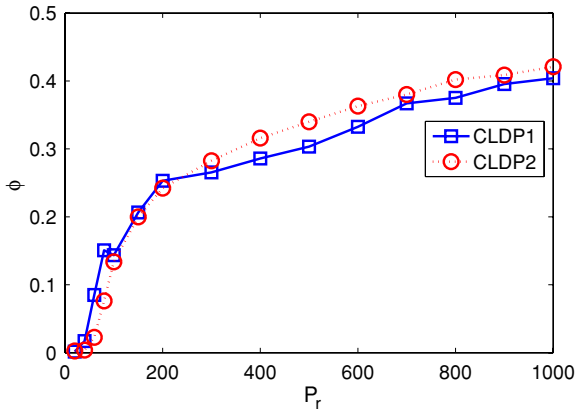

Figure 13. The matching precision ϕ as functions of P_r for different revealed pairwise matched node selection strategies CLDP1 and CLDP2. It can be found that CLDP2 is prior to CLDP1 when $P_r > 200$ even though $\eta_1 > \eta_2$.

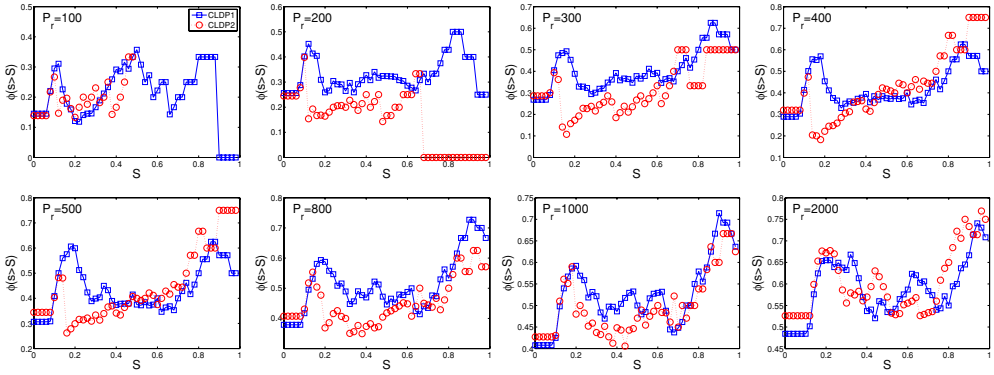

Figure 14. The matching precision $\phi(s > S)$ for the pairwise matched nodes with similarity $s > S$ as functions of S for different revealed pairwise matched nodes' selection strategies CLDP1 and CLDP2 and various $P_r = 100, 200, 300, 400, 500, 800, 1000, 2000$. It can be found that, in most cases, $\phi(s > S)$ experiences a clear decrease after it attains its local maximum value at some $S \in (0.1, 0.2)$ although it increases again when $S > 0.8$.

Generally, a pair of nodes with larger similarity are more likely to be matched correctly, e.g. in figure 9, one can see that the similarities of correctly pairwise matched nodes are always larger than those of incorrectly pairwise matched nodes. Therefore, it may be expected that $\phi(s > S)$ is a monotonic increasing function of S. That is indeed true for the tested interactional scale-free networks generated by the BA model. However, for the pair of real-world networks, it behaves a little different, i.e. in most cases, $\phi(s > S)$ experiences a clear decrease after it attains its local maximum value at some $S \in (0.1, 0.2)$ although it increases again when $S > 0.8$, as shown in figure 14. Such phenomenon may also be attributed to the high symmetry of the chat network and the friendship network. Because in such a situation, according to equations (2) and (8), the pairwise nodes of small degree always possess relatively large similarity as well as high symmetry, as a result, with a quite high probability, they will be wrongly matched in the early stage of the iterative matching process. That is, high similarity does not always

mean high matching precision, i.e. $\phi(s > S)$ may experience a local decrease in the iterative matching process.

5. Conclusions

Through calculating the similarities between pairwise nodes of different networks by their local topological properties, the node-matching problem between different networks can be transferred to a classical maximum weighted bipartite matching problem and thus can be solved by some well-known optimal matching algorithms in graph theory. However, the relatively high time complexity $O(N^3)$ and the poor matching results hinder their applications in matching large-scale real-world networks. In order to resolve the node-matching problem more efficiently, in this paper, we proposed a novel iterative node-matching algorithm with approximately linear running time $O(N)$ as well as two improved revealed pairwise matched nodes' selection strategies. Simulation shows that, compared with the optimal node-matching algorithm, the iterative algorithm can produce much better matching results especially when there are only a small number of pairwise matched nodes revealed beforehand.

However, almost all of the network structure-based one-to-one node-matching algorithms will lose their efficiency when the target networks are highly symmetric, although whether those pairwise symmetric nodes were wrongly matched or not will not influence the posterior matching process. Naturally, the symmetry of a network can be reduced by differentiating the links in terms of their attached weights. Therefore, the iterative node-matching algorithm proposed here can be further improved when considering weighted networks. Moreover, it is also very interesting and challenging to develop many-to-many node-matching algorithms which may be very useful in certain cases where one-to-one matching is not necessary, because the effect of the symmetry of the target networks can be reduced when adopting many-to-many node-matching algorithms. All of these belong to our future work.

Acknowledgment

This work has been supported by China Postdoctoral Science Foundation (grant no 20080441256).

References

[1] Albert R and Barabási A-L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47
[2] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175
[3] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabási A-L 2002 Hierarchical organization of modularity in metabolic networks *Science* **297** 1551
[4] Barabási A-L and Oltvai Z N 2004 Network biology: understanding the cell's functional organization *Nature* **5** 101
[5] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A and Tyers M 2006 BioGRID: a general repository for interaction datasets *Nucl. Acids Res.* **34** D535
[6] Cancho R F and Solé R V 2001 The small world of human language *Proc. R. Soc.* B **268** 2261
[7] Cancho R F, Solé R V and Köhler R 2004 Patterns in syntactic dependency networks *Phys. Rev.* E **69** 051915
[8] Amancio D R, Antiqueira L, Pardo T A S, Costa L F, Oliveira O N and Nunes M G V 2008 Complex networks analysis of manual and machine translations *Int. J. Mod. Phys.* C **19** 583
[9] Lakon C M, Ennett S T and Norton E C 2006 Mechanisms through which drug, sex partner, and friendship network characteristics relate to risky needle use among high risk youth and young adults *Soc. Sci. Med.* **63** 2489

[10] Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J and Barabási A-L 2007 Structure and tie strengths in mobile communication networks *Proc. Natl Acad. Sci.* **104** 7332
[11] Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati A A and Joshi A 2008 Social ties and their relevance to churn in mobile telecom networks *EDBT '08 (Nantes, France)*
[12] Newman M E J, Forrest S and Balthrop J 2002 Email networks and the spread of computer viruses *Phys. Rev.* E **66** 035101
[13] A-Hasan N and Adamic L A 2007 Expressing social relationships on the blog through links and comments *ICWSM'07 (Colorado, USA)*
[14] Furukawa T, Ishizuka M, Matsuo Y, Ohmukai I and Uchiyama K 2007 Analyzing reading behavior by blog mining *AAAI '07 (Vancouver, Canada)*
[15] Costa L D F, Rodrigues F A, Traviewso G and Boas P R V 2007 Characterization of complex networks: a survey of measurements *Adv. Phys.* **56** 167
[16] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440
[17] Barabási A-L and Albert R 1999 Emergence of scaling in Random networks *Science* **286** 509
[18] Motter A E, Nishikawa T and Lai Y-C 2003 Large-scale structural organization of social networks *Phys. Rev.* E **68** 036105
[19] Xiao Y, Xiong M, Wang W and Wang H 2008 Emergence of symmetry in complex networks *Phys. Rev.* E **77** 066108
[20] Mossa S, Barthélémy M, Stanley H E and Amaral L A N 2002 Truncation of power law behavior in scale-free network models due to information filtering *Phys. Rev. Lett.* **88** 138701
[21] Xuan Q, Li Y and Wu T-J 2006 Growth model for complex networks with hierarchical and modular structures *Phys. Rev.* E **73** 036105
[22] Li X and Chen G 2003 A local-world evolving network model *Physica* A **328** 274
[23] Kurant M and Thiran P 2006 Layered complex networks *Phys. Rev. Lett.* **96** 138701
[24] Buldyrev S V, Parshani R, Paul G, Stanley H E and Havlin S 2010 Catastrophic cascade of failures in interdependent networks *Nature* **464** 1025
[25] Xuan Q, Du F and Wu T-J 2009 Empirical analysis of Internet telephone network: from user ID to phone *Chaos* **19** 023101
[26] Xuan Q and Wu T-J 2009 Node matching between complex networks *Phys. Rev.* E **80** 026103
[27] West D B 2004 *Introduction to Graph Theory* 2nd edn (Beijing: China Machine)
[28] http://trademanager.alibaba.com/