Advances in Complex Systems, Vol. 13, No. 6 (2010) 725–739

DOI: 10.1142/S0219525910002827

World Scientific
www.worldscientific.com

ONE-TO-MANY NODE MATCHING BETWEEN COMPLEX NETWORKS

FANG DU*, QI XUAN† and TIE-JUN WU*

*Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China

†Department of Automation, Zhejiang University of Technology, Hangzhou 310023, China
†crestxq@hotmail.com

Received 22 January 2010
Revised 26 April 2010

Revealing the corresponding identities of the same individual in different systems is a common task in various areas, e.g., criminals inter-network tracking, homologous proteins revealing, ancient words translating, and so on. With the reason that, recently, more and more complex systems are described by networks, this task can also be accomplished by solving a node matching problem among these networks. Revealing one-to-one matching between networks is for sure the best if we can, however, when the target networks are highly symmetric, or an individual has different identities (corresponds to several nodes) in the same network, the exact one-to-one node matching algorithms always lose their effects to obtain acceptable results. In such situations, one-to-many (or many-to-many) node matching algorithms may be more useful. In this paper, we propose two one-to-many node matching algorithms based on local mapping and ensembling, respectively. Although such algorithms may not tell us the exact correspondence of the identities in different systems, they can indeed help us to narrow down the inter-network searching range, and thus are of significance in practical applications. These results have been verified by the matching experiments on pairwise artificial networks and real-world networks.

Keywords: Complex networks; node matching; inter-network searching; ensembling.

1. Introduction

Recently, complex networks are widely used as a tool to describe the relationships among different components or individuals in large-scale complex systems [1–7]. It is believed that the dynamics or the behavior of an individual in a complex network is strongly related to its local structure in the network. Thus, with the reason that the same individual always behaves similarly in different environments, it can be imaged that the individual may have similar local structure in different complex networks [8–10]. This observation provides a chance to reveal nodes representing different identities of the same individual in different networks by comparing their local connections. Such an inter-network searching is very popular in various areas due to the fact that similar information is often organized in different

ways in different networks [11]. For instances, police may try to track the suspects in a real-name telephone network based on the illegal behaviors firstly noticed in anonymous social networks such as instant messaging networks; biologists or linguists may be interested in finding homologous proteins in different protein-protein networks [12] or words representing the similar concepts in language networks [13]; and sometimes, geographical engineers in transportation specialization may have to integrate road networks of the same region based on data coming from various geographic information systems (GIS) [14].

Indeed, many of the inter-network searching problems in different areas can be solved by the dedicated methods specially designed in geometrical [14], semantics [15], or chemical [16, 17] areas, and so on. However, due to high economical or computational cost, it is almost impossible to examine and further compare each pair of nodes among different large-scale complex networks by these methods. Fortunately, it is unnecessary to do so. In fact, the inter-network searching problem can be transferred to a node matching problem [10] provided there is an exact one-to-one mapping between the corresponding nodes in different networks. Then, the problem can be solved by various kinds of well-known matching algorithms, e.g., Hungarian algorithm [18] in the graph theory. The efficiency and the matching precision can be further improved by introducing an iterative mechanism [19]. However, although the iterative one-to-one node matching algorithm performs surprisingly well on artificial interacted scale-free networks, it fails to produce acceptable matching results between interacted real-world networks if there is only a very small number of pairwise revealed matched nodes. The possible reasons may be that real-world networks are always highly symmetric [19, 20]; and sometimes, an individual may have different IDs in the same network, e.g., email alias [21], where the presupposition of one-to-one mapping is no longer valid. This is quite common in co-author networks [22] where the same author may be represented by different nodes just because his name may be spelled in different orders or represented by different abbreviations in different publishing systems.

In such situations, we would better propose one-to-many node matching algorithms which mainly focus on quickly narrowing down the searching range, i.e., revealing several possible candidates, of a target individual in different networks, rather than revealing exact one-to-one mapping as before. This work can be considered as a compromise proposal, and after such a narrowing, those dedicated methods could be much more efficient. In this paper, we will propose two one-to-many node matching algorithms, one (A1) is based on local mapping and the other (A2) is based on the ensembling methods [23]. It should be noted that both algorithms are based on the iterative one-to-one node matching algorithm and thus obtain better matching results than the latter. Generally, we find that A1 behaves better on heterogeneous networks while A2 works better on homogeneous networks. The latter may be attributed to the preference of the ensembling algorithms on homogeneous structures, that is, A2 on homogeneous networks can produce more various one-to-one matching results than on heterogeneous networks and therefore better matching results can be obtained.

The rest of the paper is organized as follows. In Sec. 2, the one-to-many node matching problem is briefly introduced. In Sec. 3, both the local mapping and the ensembling based one-to-many node matching algorithms are proposed. Then, the algorithms are tested on pairwise artificial networks as well as real-world social networks in Sec. 4. The paper is concluded in Sec. 5.

2. One-to-Many Node Matching Problem

Considering N pairs of matched nodes between two networks $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $V_i = \{v_1^i, v_2^i, \ldots, v_{N_i}^i\}$ and E_i represent the node set and the link set of network G_i $(i = 1, 2)$ with $N \leq \min\{N_1, N_2\}$, respectively, one may try to reveal all the other $N - P_r$ pairs of exactly one-to-one matched nodes by utilizing the structural information provided by P_r pairs of aforehand revealed matched nodes, as is shown in Figs. 1(a) and (b). In real-world cases, however, one-to-one node matching may result in a quite low matching precision especially when P_r is very small and therefore needs to be further improved in order to be applied practically. Such low matching precision may be partially attributed to the high degree of local symmetry of these real-world complex networks where many pairs of nodes share a great proportion of common neighbors. In such a situation, one-to-many or further many-to-many node matching may be more appropriate through expanding the number of nodes in each matching step, as is shown in Fig. 1(c). In fact, one-to-many node matching has its practical significance because it can help us to quickly narrow down the searching range of a target individual in different complex systems. Particularly, a 1-to-M algorithm should output $N - P_r$ correspondences, as is defined by Eq. (1),

$$v_i^1 \to Q_i^2 = \{v_{i_1}^2, v_{i_2}^2, \ldots, v_{i_M}^2\}, \tag{1}$$

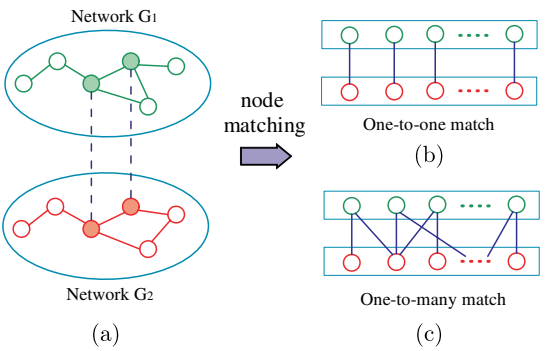


Fig. 1. (Color online) The sketch map for the node matching problem. (a) There are N pairs of matched nodes in networks G_1 and G_2. Initially, P_r pairs of matched nodes are revealed beforehand (connected by the blue dash line). (b) One-to-one match tries to exactly reveal all $N - P_r$ pairs of matched nodes. (c) One-to-many match tries to find several possible correspondences of a target node in the other network.

where v_i^1 $(i = P_r + 1, P_r + 2, \ldots, N)$ is a node in G_1, and Q_i^2 is a node set in G_2 including the top M most likely matched nodes of v_i^1. It should be noted that 1-to-M matching is just a natural generalization of 1-to-1 matching. Therefore, Eq. (1) also includes the case of a consistent 1-to-1 matching, i.e., $v_i^1 \leftrightarrow v_{i_1}^2$, satisfying that $v_{i_1}^2$ and $v_{j_1}^2$ represent two different nodes in G_2 if $i \neq j$, as is represented by Eq. (2),

$$\left| \bigcup_{i=P_r+1}^{i=N} \{v_{i_1}^2\} \right| = N - P_r. \tag{2}$$

For a 1-to-M matching algorithm, a node v_i^1 in G_1 is considered correctly matched if its really matched node v_i^2 is contained in Q_i^2, i.e., $v_i^2 \in Q_i^2$. Suppose P_M $(P_M \leq N - P_r)$ nodes in G_1 are correctly matched, the matching precision ϕ_M of the 1-to-M node matching algorithm can be calculated by Eq. (3),

$$\phi_M = \frac{P_M}{N - P_r}, \tag{3}$$

and naturally Eq. (4) must be satisfied.

$$\phi_M \geq \phi_{M-1}. \tag{4}$$

Based on these definitions, there will be totally $(1 - \phi_1)(N - P_r)$ pairs of nodes wrongly matched by adopting a 1-to-1 node matching algorithm, and $(\phi_M - \phi_1)(N - P_r)$ pairs of them can be corrected by the corresponding 1-to-M node matching algorithm. As a result, the superiority of 1-to-M over 1-to-1 can be evaluated by the error recovery capability (ERC) defined by Eq. (5),

$$\rho_M = \frac{\phi_M - \phi_1}{1 - \phi_1}, \tag{5}$$

standing for the ability of the 1-to-M node matching algorithm to correct the error introduced by the 1-to-1 node matching algorithm.

3. One-to-Many Node Matching Algorithms

In this section, we propose two one-to-many node matching algorithms to improve the matching precision by adopting the result of the iterative one-to-one node matching algorithm as their initial states. The iterative one-to-one node matching algorithm is well established in our another paper [19] and will not be discussed here. At the same time, here, the similarity between two nodes v_i^1 and v_j^2 of different networks is identically defined by Eq. (6),

$$S(v_i^1, v_j^2) = \frac{n_M(v_i^1, v_j^2)}{n_L(v_i^1) + n_L(v_j^2) - n_M(v_i^1, v_j^2)}, \tag{6}$$

where $n_M(v_i^1, v_j^2)$ denotes the number of revealed pairwise matched nodes (v_k^1, v_k^2) to which the nodes v_i^1 and v_j^2 are both connected, 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 $n_L(v_i^1)$ (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).

3.1. *A1: Local mapping*

The iterative one-to-one node matching algorithm has its intrinsic confusion on the similarity defined in Eq. (6) that the similarity between each pair of nodes may change as the algorithm is implemented step by step. Therefore, it seems a little imprudent if we determine the match of nodes just by the values of similarities at that iterative time. Let $n_M^t(v_i^1, v_j^2)$ the number of revealed pairwise matched nodes to which the nodes v_i^1 and v_j^2 are both connected at iterative time t and $S^t(v_i^1, v_j^2)$ the similarity between them. Then, as more and more pairwise matched nodes are revealed, $n_M^t(v_i^1, v_j^2)$ will increase as the algorithm proceeds, i.e., Eq. (7) must be satisfied.

$$t_1 \leq t_2 \Rightarrow S^{t_1}(v_i^1, v_j^2) \leq S^{t_2}(v_i^1, v_j^2). \tag{7}$$

In other words, some wrongly matched nodes can be probably corrected by recalculating the similarity between nodes of different networks after the one-to-one node matching algorithm is terminated, which inspires us to propose the first one-to-many node matching algorithm based on local mapping. Formally, the Algorithm A1 is defined by following two steps.

(1) **Iterative 1-to-1 node matching.** The iterative 1-to-1 node matching algorithm is carried out firstly, and $N - P_r$ pairs of nodes, i.e., $v_i^1 \leftrightarrow v_{i_1}^2$ ($i = P_r + 1, P_r + 2, \ldots, N$), are matched when the algorithm terminates, as is presented in Eq. (8).

$$v_i^1 \rightarrow Q_i^2 = \{v_{i_1}^2\}, \quad i = P_r + 1, P_r + 2, \ldots, N. \tag{8}$$

(2) **Corresponding node candidates selection.** In order to establish a 1-to-M match, as is defined by Eq. (1), we should still select other $M - 1$ nodes from G_2 as the correspondences of v_i^1 after the 1-to-1 matching result has been obtained. Each node v_i^1 in G_1 has a neighbor set X_i^1 including all the nodes directly connected to v_i^1, X_i^1 then has a matched node set X_i^2 in G_2 representing the nodes 1-to-1 matched to the nodes in X_i^1. Similarly, the neighbor set of X_i^2, including all the nodes, except $v_{i_1}^2$, directly connected to at least one node in X_i^2, is denoted by Y_i^2. Recalculating the similarity between each pair of nodes of different networks is quite time-consuming and unnecessary. In fact, based on the definition of similarity (Eq. (6)) in this paper, only the similarities between node v_i^1 ($i = P_r + 1, P_r + 2, \ldots, N$) and the nodes in Y_i^2 are larger than 0 and need to be recalculated by Eq. (6). The top $M-1$ nodes with largest similarities in Y_i^2 as well as $v_{i_1}^2$ are selected as the M corresponding node candidates of v_i^1. It should be noted that, if Y_i^2 only contains fewer than $M - 1$ nodes, other $M - 1 - |Y_i^2|$ corresponding node candidates can be randomly selected from G_2 to fit the definition in Eq. (1). A simple example is shown in Fig. 2.

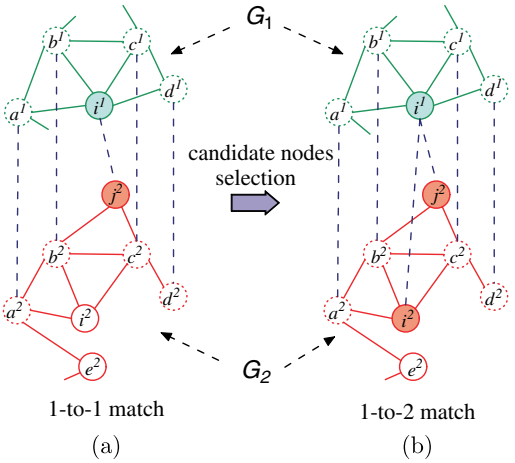


Fig. 2. (Color online) The sketch map for the corresponding node candidates selection in the Algorithm A1. (a) The 1-to-1 match is established where node i^1 in G_1 is wrongly matched to node j^2 in G_2 linked by a blue dash line. In G_1, i^1 has a neighbor set $X_i^1 = \{a^1, b^1, c^1, d^1\}$, X_i^1 has a matched set $X_i^2 = \{a^2, b^2, c^2, d^2\}$ in G_2, and X_i^2 then also has a neighbor set $Y_i^2 = \{a^2, b^2, c^2, d^2, e^2, i^2\}$ in G_2. (b) Then the similarity between i^1 and each node in Y_i^2 is calculated. The top $M-1$ nodes with largest similarities in Y_i^2 as well as j^2 are selected as the M corresponding node candidates of i^1. Here $M = 2$, then i^2 (the really matched node of i_1) and j_2 are selected as the corresponding node candidates.

The above one-to-many node matching algorithm suggests that the correctly matched neighbors may help a node to be recovered from a wrong match. Naturally, the ERC of the one-to-many node matching algorithm is closely related to the initial matching precision of the one-to-one node matching algorithm.

3.2. *A2: Ensembling*

Here, we try to introduce another one-to-many node matching algorithm based on ensembling methods which bears a close resemblance to the ensemble learning [23] with the goal of constructing a collection of individual predictors that are diverse and yet accurate. This idea of improving the generalization performance by combining the results of many different predictors has been investigated extensively both in theory and in practice [24–27]. Two of the most popular techniques for constructing ensembles are the Adaboost family of algorithms (i.e., Boosting) [24] and the bootstrap aggregation (i.e., Bagging) [25]. Generally, the latter is relatively robust to outliers and noise [28], and will be more suitable in the context of node matching problem between networks where the one-to-one correspondence is blurred due to high symmetry or multiple identities.

However, the traditional iterative one-to-one node matching algorithm is totally deterministic, that is, for a given pair of target networks and certain initial states, the algorithm must produce the same matching result. Therefore, it cannot be directly used for ensemble. In order to produce an ensemble of one-to-one matching

results, we should introduce a new statistical iterative one-to-one node matching algorithm first. In the deterministic iterative one-to-one node matching algorithm, each pair of revealed matched nodes is considered in calculating the similarities between unrevealed nodes of different networks in the posterior iterative process, which is not the case in the statistical iterative one-to-one node matching algorithm, where there is a probability γ $(0 < \gamma < 1)$ to determine whether or not a pair of newly revealed matched nodes should be considered in the succeeding iterative process. That is, only with probability γ, a pair of newly revealed matched nodes is adopted to calculate the similarities between those unrevealed nodes of different networks.

Then a group of different one-to-one matching results can be obtained by implementing such a statistical iterative one-to-one node matching algorithm for several rounds, and the results can be merged into a unique one-to-many matching result by a voting strategy. The Algorithm A2 is defined by following three steps.

(1) **Deterministic iterative 1-to-1 node matching.** The deterministic iterative 1-to-1 node matching algorithm is carried out firstly, and $N - P_r$ pairs of nodes, i.e., $v_i^1 \leftrightarrow v_{i_1}^2$ $(i = P_r + 1, P_r + 2, \ldots, N)$, are matched when the algorithm terminates, as is presented in Eq. (8).

(2) **Ensembling and voting.** The statistical 1-to-1 node matching algorithm is implemented for B $(B \gg M)$ rounds and an ensemble of B different 1-to-1 matching results is obtained. All of the correspondences, except $v_{i_1}^2$, in G_2 of v_i^1 in G_1 are grouped by a node set Z_i^2 with its size (the number of nodes) satisfying $1 \leq |Z_i^2| \leq B$. It should be noted that each node in Z_i^2 is attached by a positive integer as its weight representing the number of times that it is matched to v_i^1 in the totally B rounds, which can be regarded as a voting process.

(3) **Corresponding node candidates selection.** The top $M - 1$ nodes with the largest weights in Z_i^2 as well as $v_{i_1}^2$ are selected as the M corresponding node candidates of v_i^1. If Z_i^2 only contains fewer than $M - 1$ nodes, other $M - 1 - |Z_i^2|$ corresponding node candidates are randomly selected from G_2 to fit the definition in Eq. (1).

4. Matching Experiments

In this section, the matching results of the one-to-one $(M = 1)$ iterative node matching algorithm proposed by Xuan *et al.* [19] and the expanded one-to-many $(M > 1)$ node matching algorithms proposed in this paper will be all recorded for comparison.

4.1. *Tests on artificial networks*

In order to test the one-to-many matching algorithms, we will here adopt the same model [10] to create pairs of artificially matched networks. In the model,

two networks G_1 and G_2 with N nodes respectively are created by the same rule, in which all the nodes are randomly one-to-one matched (the parameters are set to be $N_1 = N_2 = N$ for convenience). Then G_1 and G_2 interact with each other by link coping, that is, two unlinked nodes in G_2 (or G_1) are connected with probability η_1 (or η_2) if their matched nodes in G_1 (or G_2) are linked initially. In this paper, the initial networks G_1 and G_2 are respectively set to be scale-free networks [2], local-world networks [30–32], regular networks and small-world networks [1] due to their typicality in the complex network research.

In all the experiments, the basic parameters are set to be the same, i.e., $N = 1000$ and $\eta_1 = \eta_2 = 0.3$. Particularly, in the scale-free network experiment, the initial networks G_1 and G_2 are both generated by the BA model [2]: 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 different nodes which are selected by the preferential attachment (PA) rule, i.e., nodes are selected with a probability proportional to their degree. Here, the parameters are set to be $m = m_0 = 5$. In the local-world network experiment, the initial networks G_1 and G_2 are both generated by the LW model proposed by Xuan *et al.* [32]: the above PA rule is applied in a predefined local world (a set of nodes closest to the target node and the size is denoted by l) of a newly added node v_t, m ($m < l$) of which are selected according to the PA rule and connected to v_t. Here the parameters are set to be $m = 5$ and $l = 10$. The resultant network has a relatively high average clustering coefficient [32]. In the regular network experiment, G_1 and G_2 are regular networks in which every node has an identical degree of $k = 10$. In the small-world network experiment, G_1 and G_2 are both generated by the model proposed by Watts and Strogatz (WS) [1], where randomness is introduced into a regular (RG) network through a rewiring process (each link is rewired with a probability p). Here the rewiring probability is set to $p = 0.1$.

The one-to-many node matching strategy provides a compromise proposal to make a balance between the initial matching precision and the follow-up searching range. Although it is obvious that we can obtain higher matching precision just by expanding the number of correspondences in the other network of a target node, as is presented in Eq. (4), we still need to reveal where and when the one-to-many node matching algorithms behave best and therefore achieve acceptable results with as small number of correspondences as possible. In each experiment, the Algorithms A1 and A2 are implemented in 100 different pairs of matched networks for each $P_r \in \{10, 20, \ldots, 300\}$, and the matching results for $M = 1, 2$ ($M = 1$ means the one-to-one matching result) and different types of networks including the BA networks, the LW networks, the WS networks, and the RG networks at different P_r are shown in Figs. 3(a)–(d) respectively. For Algorithm A2, the parameters are set to be $\gamma = 0.9$ and $B = 20$. It is shown that both the algorithms behave best in the maximum gradient zone of the one-to-one node matching precision for all of the experiments. This phenomenon is quite reasonable because the two one-to-many node matching algorithms proposed in this paper are both based on the one-to-one

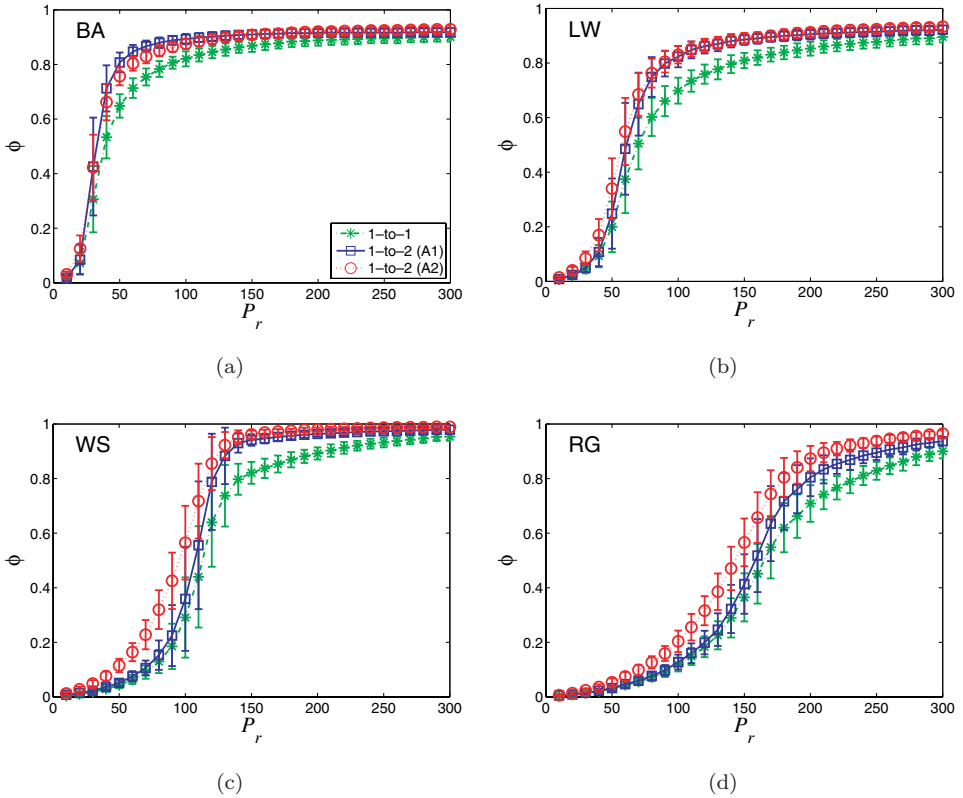


Fig. 3. (Color online) The matching results for $M = 1, 2$ ($M = 1$ means the one-to-one matching result) and different types of networks including (a) the BA networks with parameters $m = m_0 = 5$, (b) the LW networks with parameters $m = m_0 = 5$ and $l = 10$, (c) the WS networks with rewiring probability $p = 0.1$, and (d) the RG networks at different number of pairs of revealed matched nodes P_r. Each tested network has $N = 1000$ nodes and the interactive parameters are set to $\eta_1 = \eta_2 = 0.3$. All of these parameters are set the same in Figs. 4 and 6.

node matching algorithm, therefore, it is not strange that they will consequently fail to achieve satisfactory results when the one-to-one node matching precision is close to 0. Besides, when the one-to-one node matching precision is close to 1, i.e., only a small portion of pairs of nodes are wrongly matched, naturally there is a small chance for the one-to-many node matching algorithms to further improve the matching result. Besides, by comparison, the superiority of A2 over the one-to-one algorithm is more obvious in homogeneous networks such as WS networks and RG networks, which may be attributed to the preference of the ensembling algorithms on homogeneous structures which can provide more various one-to-one matching results and therefore one-to-many matching results of higher quality can be produced.

Moreover, the detailed relationships between the ERCs of the two 1-to-2 node matching algorithms and the 1-to-1 node matching precision are shown in

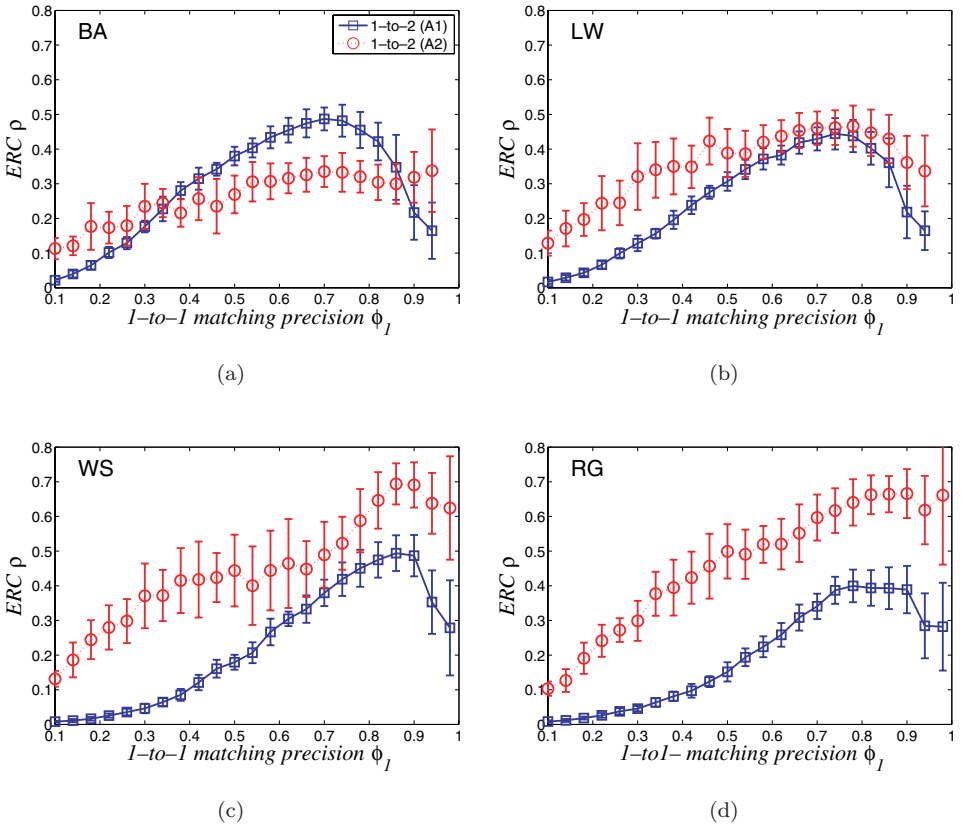


Fig. 4. (Color online) The relationships between the ERCs of the two 1-to-2 node matching algorithms and the 1-to-1 node matching precision for different types of networks including (a) the BA networks, (b) the LW networks, (c) the WS networks, and (d) the RG networks.

Figs. 4(a)–(d) respectively for the four different types of networks (results for more values of M can be found in Fig. 5), where we can see that, for A1, the relationship presents a clear unimodal pattern in each experiment, i.e., ERC reaches its maximum value at about $\phi_1 \approx 0.7$ for the BA networks and LW networks, which further right shifts to $\phi_1 \approx 0.8$ for the WS networks and RG networks. This phenomenon can be well explained by introducing the aggregation of the real matched nodes. Here, a pair of real matched nodes v_i^1 and v_i^2 are considered aggregated with each other if their similarity $S(v_i^1, v_i^2) > 0$ and they are wrongly matched by the one-to-one node matching algorithm. Obviously, only those pairs of aggregated real matched nodes can be recovered by A1. Therefore, the average aggregation, defined by Eq. (9), can be considered as an upper bound of the ERC by adopting A1.

$$\langle \pi \rangle = \frac{\sum_{i=P_r+1}^{i=N} c(v_i^1, v_i^2)}{(1-\phi_1)(N-P_r)}, \tag{9}$$

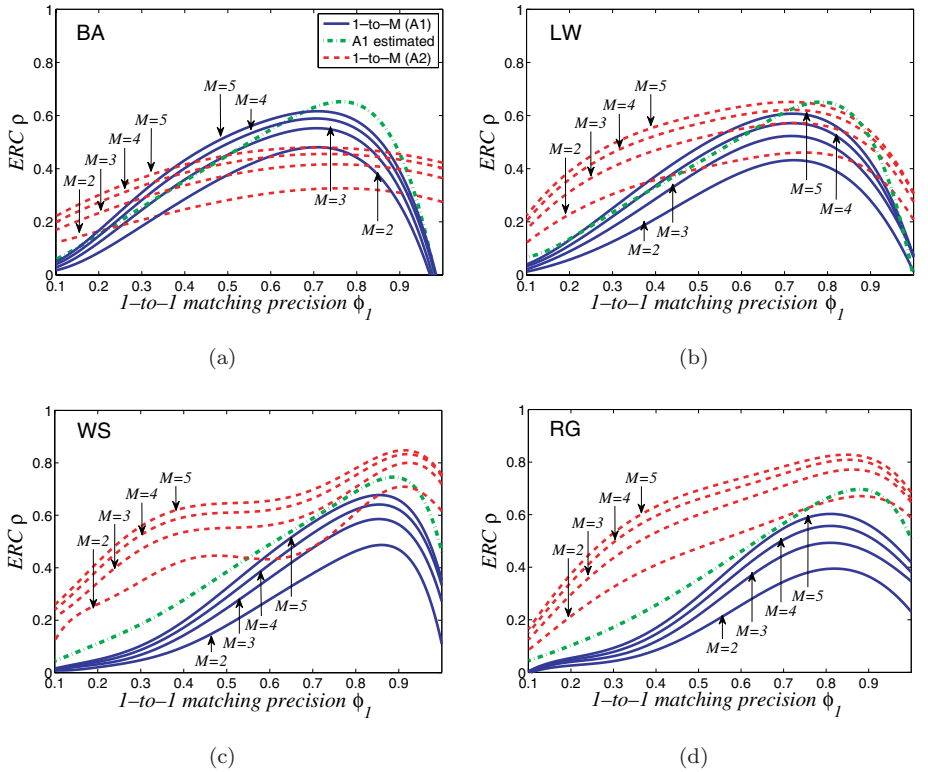


Fig. 5. (Color online) The relationships between the real ERCs of the two 1-to-M node matching algorithms as well as the estimated ERC of A1 (green dash-dot line) and the 1-to-1 node matching precision for $M = 2, 3, 4, 5$ and different types of networks including (a) the BA networks, (b) the LW networks, (c) the WS networks, and (d) the RG networks. Here, the error bars are hidden in order to provide clearer images.

where $c(v_i^1, v_i^2)$ is a sign function, i.e., $c(v_i^1, v_i^2) = 1$ if $S(v_i^1, v_i^2) > 0$ and $c(v_i^1, v_i^2) = 0$ otherwise. Particularly, the relationship between the average aggregation $\langle \pi \rangle$ and the one-to-one node matching precision ϕ_1 for the four different types of networks are shown in Fig. 6, where we can see that such relationships present the similar unimodal pattern and similar one-to-one node matching precisions at which ERCs reach their maximum value as those shown in Fig. 4. The decreasing trend of $\langle \pi \rangle$ when $\phi_1 > 0.8$ suggests that there are indeed a small portion of matched nodes really very hard to be revealed by the current matching algorithms, which is mainly attributed to their quite high local symmetry [19].

Furthermore, considering that a node with more correctly matched neighbors may recover from a wrong match with a higher probability, the ERC can be further estimated by Eq. (10),

$$\tilde{\rho} = \frac{\sum_{i=P_r+1}^{i=N} c(v_i^1, v_i^2) n_c(v_i^1)/k(v_i^1)}{(1-\phi_1)(N-P_r)}, \tag{10}$$

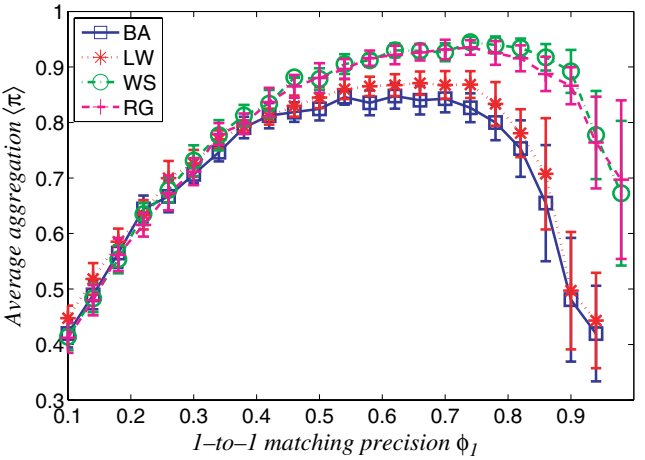


Fig. 6. (Color online) The relationship between the average aggregation $\langle\pi\rangle$ and the one-to-one node matching precision ϕ_1 for different types of networks including (a) the BA networks, (b) the LW networks, (c) the WS networks, and (d) the RG networks.

where $n_c(v_i^1)$ denotes the number of correctly matched neighbors of v_i^1 and $k(v_i^1)$ denotes the total number of its neighbors. The estimated ERC $\tilde{\rho}$ as well as the real ERC for various $M = 2, 3, 4, 5$ and the different types of networks are plotted in Figs. 5(a)–(d) respectively. It should be noted that the error bars are hidden here in order to provide clearer images. The results shown in Fig. 5 are basically consistent with those shown in Figs. 3 and 4.

4.2. *Tests on real-world networks*

In this section, the Algorithms A1 and A2 will be tested on a pair of real-world networks collected from the database of *Alibaba trademanager* [29] which is an instant messenger (IM) mainly used for electronic commerce. We mainly focus on 14,800 employees of the Alibaba company and construct the friendship network G_1 and the chat network G_2 among them based on their contact lists and communication records in a week. The two networks are then preprocessed by the following two steps.

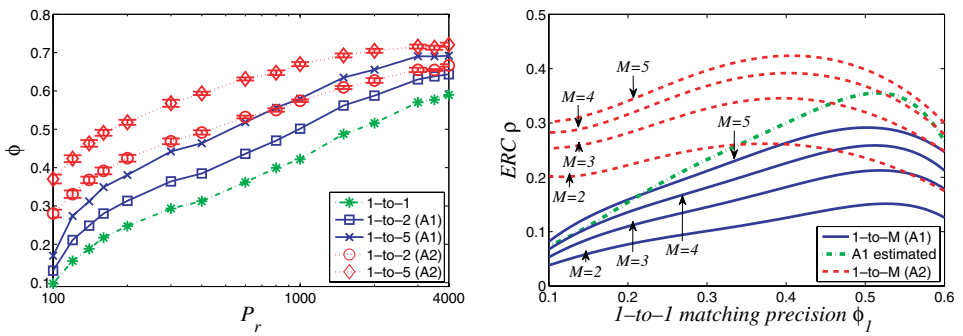
(1) **Extract the Giant Cluster (GC).** Extract the GCs of G_1 and G_2, denoted by $G_1^g = (V_1^g, E_1^g)$ and $G_2^g = (V_2^g, E_2^g)$ where V_i^g and E_i^g represent the node set and the link set of the GC G_i^g respectively.
(2) **Calculate the intersection.** A pair of matched nodes in the networks correspond to the same 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 $G_1^c = (V^c, E_1^c)$ and $G_2^c = (V^c, E_2^c)$ where $E_i^c \subseteq E_i^g$ represents the set of links among the 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 networks G_1 and G_2 have $N = 9{,}859$ 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. Their basic topological properties, such as the number of vertices N, the average degree $\langle k \rangle$, the average clustering coefficient $\langle C \rangle$, the average shortest path length $\langle L \rangle$ are presented in Table 1.

The matching results for $M = 1, 2, 5$ are shown in Fig. 7 as well as the relationships between the ERCs of the two 1-to-M node matching algorithms and the 1-to-1 node matching precision for $M = 2, 3, 4, 5$ are shown in Fig. 7(b). For Algorithm A2, the parameters are set to $\gamma = 0.9$ and $B = 20$. It is shown that both the one-to-many node matching algorithms proposed in this paper take obvious effects on recovering from matching errors caused by the one-to-one node matching algorithm. And by comparison, A2 behaves better than A1. These results are quite similar to those presented in the experiments on LW networks, WS networks, and RG networks, all of which have significantly large average clustering coefficients. Therefore we guess that the clustering property may be another factor that influences the error

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

Networks	N	$\langle k \rangle$	$\langle C \rangle$	$\langle L \rangle$
Chat	9,859	39.4	0.218	3.37
Friendship	9,859	172	0.313	2.55



(a) (b)

Fig. 7. (Color online) (a) The matching results for $M = 1, 2, 5$ ($M = 1$ means the one-to-one matching result) between the friendship network and the chat network obtained from the database of *Alibaba trademanager*. (b) The relationships between the real ERCs of the two 1-to-M node matching algorithms as well as the estimated ERC of A1 (green dash-dot line) and the 1-to-1 node matching precision for $M = 2, 3, 4, 5$.

recovering ability of the one-to-many node matching algorithms, which needs to be further validated in the future.

5. Conclusions

With the reason that one-to-one matching algorithms fail to reach acceptable results when the target networks are highly symmetric or the correspondences of nodes are not unique in different networks, we expanded the concept and proposed two different one-to-many node matching algorithms in this paper. Such algorithms have its practical significance because it can help us to reduce the inter-network searching range and thus make those dedicated methods more efficiently. The one-to-many node matching algorithms can be qualified by their ERCs from the basic one-to-one node matching algorithm. The matching experiments on pairwise artificial networks and real-world networks show that both the proposed Algorithms A1 and A2 have quite high ERCs, and A2 behaves especially well for pairwise WS networks and RG networks which share similar homogeneous structure. Besides, more naturally, it seems that the ERCs of one-to-many algorithms are strongly correlated to the matching precision obtained by the basic one-to-one node matching algorithm. In the future, the ensembling based one-to-many node matching algorithm can be further improved provided that more various one-to-one matching results can be obtained by different one-to-one node matching algorithms.

Acknowledgments

We would like to thank all the members of our research group in the Department of Control Science and Engineering, Zhejiang University at Yuquan campus, for the valuable discussion about the ideas presented in this paper. This work has been supported by China Postdoctoral Science Foundation (Grant No. 20080441256).

References

[1] Watts, D. J. and Strogatz, S. H., Collective dynamics of "small-world" networks, *Nature* **393** (1998) 440.
[2] Barabási, A.-L. and Albert, R., Emergence of scaling in random networks, *Science* **286** (1999) 509.
[3] Yook, S.-H., Jeong, H. and Barabási, A.-L., Modeling the Internet's large-scale topology, *Proc. Natl. Acad. Sci.* **99** (2002) 13382.
[4] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabási, A.-L., Hierarchical organization of modularity in metabolic networks, *Science* **297** (2002) 1551.
[5] Serrano, M. Á. and Boguñá, M., Topology of the world trade web, *Phys. Rev. E* **68** (2003) 015101(R).
[6] Barabási, A.-L. and Oltvai, Z. N., Network biology: Understanding the cell's functional organization, *Nature* **5** (2004) 101.
[7] Motter, A. E., Nishikawa, T. and Lai, Y.-C., Large-scale structural organization of social networks, *Phys. Rev. E* **68** (2003) 036105.

[8] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. and Ideker, T., Conserved patterns of protein interaction in multiple species, *Proc. Natl. Acad. Sci.* **102** (2006) 1974.
[9] Cancho, R. F., Solé, R. V. and Köhler, R., Patterns in syntactic dependency networks, *Phys. Rev. E* **69** (2004) 051915.
[10] Xuan, Q. and Wu, T.-J., Node matching between complex networks, *Phys. Rev. E* **80** (2009) 026103.
[11] Kurant, M. and Thiran, P., Layered complex networks, *Phys. Rev. Lett.* **96** (2006) 138701.
[12] Jeong, H., Mason, S., Barabási, A.-L. and Oltvai, Z. N., Lethality and centrality in protein networks, *Nature* **411** (2001) 41.
[13] Cancho, R. F. and Solé, R. V., The small world of human language, *Proc. R. Soc. Lond. B* **268** (2001) 2261.
[14] Xiong, D., A three-stage computational approach to network matching, *Trans. Res. C* **8** (2000) 71–89.
[15] Giunchiglia, F. and Shvaiko, P., Semantic matching, *The Knowledge Engineering Review* **18** (2004) 265–280.
[16] Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. and Ideker, T., Conserved pathways within bacteria and yeast as revealed by global protein network alignment, *Proc. Natl. Acad. Sci.* **100** (2003) 11394.
[17] Cootes, A. P., Muggleton, S. H. and Sternberg, M. J. E., The identification of similarities between biological Networks: Application to the metabolome and interactome, *J. Mol. Biol.* **369** (2007) 1126.
[18] Kuhn, H. W., The Hungarian method for the assignment problem, *Naval Res. Logist. Quart.* **2** (1995) 88–97.
[19] Xuan, Q., Du, F. and Wu, T.-J., Iterative node matching between complex networks, *J. Phys. A: Math. Theor.* **43** (2010) 395002.
[20] Xiao, Y., Xiong, M., Wang, W. and Wang, H., Emergence of symmetriy in complex networks, *Phys. Rev. E* **77** (2008) 066108.
[21] Holzer, R., Malin, B. and Sweeney, L., Email alias detection using social network analysis, *Proc. ACM SIGKDD* (2005).
[22] Newman, M. E. J., Coauthorship networks and patterns of scientific collaboration, *Proc. Natl. Acad. Sci.* **101** (2004) 5200.
[23] Dietterich, T. G., Ensemble methods in machine learning, *First International Workshop, MCS2000* (Cagliari, Italy) (2000).
[24] Freund, Y. and Shapire, R. E., A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.* **55** (1997) 119.
[25] Breiman, L., Bagging predictors, *Mach. Learn.* **26** (1996) 123–140.
[26] Krogh, A. and Sollich, P., Statistical mechanics of ensemble learning, *Phys. Rev. E* **55** (1997) 811.
[27] Miyoshi, S., Hara, K. and Okada, M., Analysis of ensemble learning using simple perceptrons based on online learning theory, *Phys. Rev. E* **71** (2005) 036116.
[28] Breiman, L., Random forests, *Mach. Learn.* **45** (2001) 5–32.
[29] Alibaba Corporation, http://trademanager.alibaba.com/ (2009).
[30] Mossa, S., Barthélémy, M., Stanley, H. E. and Amaral, L. A. N., Truncation of power law behavior in scale-free network models due to information filtering, *Phys. Rev. Lett.* **88** (2002) 138701.
[31] Li, X. and Chen, G., A local-world evolving network model, *Physica A* **328** (2003) 274.
[32] Xuan, Q., Li, Y. and Wu, T.-J., A local-world network model based on inter-node correlation degree, *Physica A* **378** (2007) 561.