7

A Review on Node-Matching Between Networks

Qi Xuan[1], Li Yu[1], Fang Du[2] and Tie-Jun Wu[3]
[1]*Zhejiang University of Technology*
[2]*Johns Hopkins University*
[3]*Zhejiang University*
[1,3]*China*
[2]*USA*

1. Introduction

The relationships between individuals in various systems are always described by networks. Recently, the quick development of computer science makes it possible to study the structures of those super-complex networks in many areas including sociology (Xuan et al., 2009; Xuan, Du & Wu, 2010a), biology (Barabási & Oltvai, 2004; Eguíluz et al., 2005), physics (Dorogovtsev et al., 2008; Rozenfeld et al., 2010), etc., by the tools in graph theory. Interestingly, it was revealed that many of these complex networks in various areas present several similar topological properties, such as small-world (Watts & Strogatz, 1998), scale-free (Barabási & Albert, 1999), self-similarity (Motter et al., 2003), symmetry (Xiao et al., 2008), etc. In order to explain these properties, a large number of models have been proposed (Barabási & Albert, 1999; Li & Chen, 2003; Mossa et al., 2002; Watts & Strogatz, 1998; Xiao et al., 2008; Xuan, Du, Wu & Chen, 2010; Xuan et al., 2006; 2007; 2008). However, most of current researches still focus on understanding the relationships between individuals in a single system, while the inter-system relationships are always ignored.

One of such inter-system relationships may be caused by the fact that an individual may be active in different systems with different identities (Xuan & Wu, 2009), and this type inter-system relationships may further lead to the similar structures of different complex networks. For instance, an ancient protein may evolve into various homologous proteins in different species, a concept may be expressed by different words in different languages, and a person may be active in different communication networks with different identities represented by telephone numbers (Onnela et al., 2007) and email addresses (Newman et al., 2002), etc. Therefore, revealing the different identities of an individual in several different systems has practical significance in many areas (Xuan, Du & Wu, 2010b), e.g., revealing homogeneous proteins, auto-translating languages, inter-network filtrating information, and so on. Through describing complex systems by networks, these different tasks can be transferred to a common node-matching problem between different complex networks, and thus can be solved in the same framework.

However, since many real-world complex networks are always highly symmetric (Xiao et al., 2008), i.e., there are always large numbers of nodes sharing the same neighbors in a network, it seems quite difficult to distinguish them in one network only by comparing their topological properties (Costa et al., 2007), such as degrees, clustering coefficient and

so on, not to mention matching them between different complex networks. Fortunately, the researchers of different areas can use their own dedicated methods, such as chemical (Cootes et al., 2007; Kelley et al., 2003), semantic (Giunchiglia & Shvaiko, 2004) 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. Such extra information are certainly very useful in solving the node-matching problem between complex networks. Based on these findings, we first introduced two kinds of co-evolving models (Xuan, Du & Wu, 2010b; Xuan & Wu, 2009) to create interacting networks, which can help better understand the co-evolution of different systems. Such co-evolution results in some structural similarity between complex networks, which made it possible to design node-matching algorithms by adopting the structural information. With the reason that the selection of the pairwise matched nodes revealed a priori by the dedicated methods is somewhat controllable, we then proposed several revealed matched nodes selecting strategies to improve the performances of node-matching algorithms. Finally, based on the similarities between nodes of different networks calculated by their connections to several pairs of preliminarily revealed matched nodes, we provided three different node-matching algorithms, including the classical optimal matching algorithm (Kuhn, 2005; Munkres, 1957; Xuan & Wu, 2009) in graph theory, one-to-one and one-to-many iterative node-matching algorithms (Du et al., 2010; Xuan, Du & Wu, 2010b) to solve artificial and real-world node-matching problems.

This chapter will review the overall process that we defined and solved the node-matching problems between different networks. In the next section, the node-matching problem is defined, and two co-evolving network models as well as a real-world node-matching data set are introduced. In Section 3, several revealed matched nodes selecting strategies are provided in order to improve the performances of the subsequent node-matching algorithms. Then in Section 4, the similarities between nodes of different networks are defined and several node-matching algorithms are introduced and the experiments are implemented. Finally, the chapter is concluded in Section 5.

2. Definitions and data sets

2.1 Definitions

The node matching problem between two different networks are described as follows (Xuan, Du & Wu, 2010b; Xuan & Wu, 2009): the two networks under study are denoted by $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, where $V_i = v_1^i, \ldots, v_M^i$ and E_i represent the node set and the link set of network $i(i = 1, 2)$, respectively. Assume that there are $M(M \leqslant \min\{N_1, N_2\})$ pairs of matched nodes $v_j^1 \leftrightarrow v_j^2$ defined by $\{v_1^i, \ldots, v_M^i\} \subseteq V_i(i = 1, 2)$, while $P_r(P_r < M)$ pairs of them have been already revealed, named as revealed matched nodes and denoted by $\{v_1^i, \ldots, v_{P_r}^i\} \subset V_i(i = 1, 2)$. Then the problem is: can we design a method to find the other $M - P_r$ pairs of matched nodes in these two distinct networks by using the structural information of G_1 and G_2 and the revealed matched nodes? If we can design such a method and finally $P_c(Pc \leqslant M - Pr)$ pairs of them are revealed correctly, the matching precision ϕ then can be calculated by

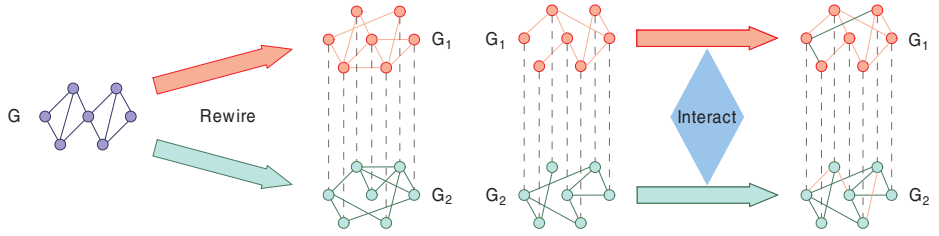
$$\phi = \frac{P_c}{M - P_r}. \tag{1}$$

2.2 Co-evolution network models

In order to better understand the interactions between different systems and test the subsequent node matching algorithms, two co-evolution network models need to be first introduced, where the parameters are set to be $N_1 = N_2 = M = N$ for convenience. Generally, there are two ways to create a pair of interactional networks, as is shown in Fig. 1 (a) and (b), respectively, both of which may work in reality. Inspired by the evolution of organisms, the first way is that the pair of interactional networks G_1 and G_2 are evolved from a common original network; in other words, they are derived from the same network (obtained by some model) through random rewiring processes. And the other way is that the pair of interactional networks are derived from two independent networks by a random interacting process composed of the following two steps (Xuan & Wu, 2009):

- **Networks initialization:** Two networks G_1 and G_2 with N nodes respectively are created 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.
- **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).

Here, the second way will be adopted to create pairs of tested artificial interactional networks.



(a) Rewiring model (b) Interacting model

Fig. 1. Two ways to create a pair of interactional networks (Xuan & Wu, 2009). (a) The pair of interactional networks G_1 and G_2 are derived from the same original network through random rewiring. The corresponding nodes are matched and connected by brown dashed lines. (b) The pair of interactional networks G_1 and G_2 are derived from a pair of independent networks by random interacting, i.e., two non-linked nodes in the network G_1 are connected by a green line with probability η_2 if their corresponding matched nodes in G_2 were linked while two non-linked nodes in G_2 are connected by a red line with probability η_1 if their corresponding matched nodes in G_1 were linked. η_1 and η_2 are named as interactional degree.

2.3 Real-world interactional networks

In reality, when two strangers chat with each other for some reason, e.g., demand of business, common interests, curiosity, warmheart, etc., they may be friends one day in the future if they enjoy with each other, in other words, the chat network may influence the evolution of the friendship network. On the other hand, there is also 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. Therefore, chat network and friend network can be considered as a pair of real-world interactional networks, which can be figured on a quite large scale by advanced communication technologies and thus used to test the subsequent node matching algorithms.

As an example, we collected the communication records and the contact lists in a week from the database of *Alibaba trademanager* (an instant messenger (IM) mainly used for electronic commerce). We mainly focus on 14,800 employees of the *Alibaba* company and construct the chat network G_1 and the friendship network G_2 among them by these records. The two networks were then preprocessed by the following two steps (Du et al., 2010; Xuan, Du & Wu, 2010b):

- **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.
- **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 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 the first step.

After the preprocessing, both the 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. Their basic topological properties, such as the number of nodes N, the average degree $\langle k \rangle$, the average clustering coefficient $\langle C \rangle$, and the average shortest path length $\langle L \rangle$ are presented in Table 1.

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

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* database (Du et al., 2010; Xuan, Du & Wu, 2010b).

3. Revealed matched nodes selecting strategies

Since the interactional networks under study are usually not completely identical (Xuan & Wu, 2009), it seems unpractical to match nodes between different networks just by their local structural properties. As a result, a few pairs of matched nodes would be better revealed as references before the node-matching algorithms are implemented.

Recent studies on real-world networks reveals that many of them have similar heterogeneous structure characterized by a power-law degree distribution (Barabási, 2009; Barrat et al., 2004; Eguíluz et al., 2005; Xuan et al., 2009). This property, first modeled by Barabási and Albert (BA) (Barabási & Albert, 1999), indicates that the connection of a heterogeneous network highly depends on hub nodes with quite large degrees, i.e., once these hub nodes are attacked,

the average shortest path length of the network will increase quickly (Albert et al., 2000; Crucitti et al., 2004; Motter & Lai, 2002), as a result, the communication efficiency of the network will be largely weakened. For the node matching problem introduced here, we proved that (Xuan & Wu, 2009) such hub nodes can provide more structural information than those normal nodes and thus are more suitable to be revealed matched nodes. Based on the interactional model introduced in Fig. 1 (b), denoting the degree of v_i^1 by d_i^1 and the degree of v_j^2 by d_j^2, if they are randomly selected as a pair of matched nodes, then, averagely speaking, there are $d_i^1 d_j^2/N$ other pairs of matched nodes around them before the interaction. And after the interaction, the degree of v_i^1 and that of v_j^2 can be calculated by Eq. (2) and Eq. (3) respectively,

$$\widetilde{d_i^1} = d_i^1 + d_j^2(1 - \frac{d_i^1}{N})\eta_2, \tag{2}$$

$$\widetilde{d_j^2} = d_j^2 + d_i^1(1 - \frac{d_j^2}{N})\eta_1. \tag{3}$$

And the number of pairs of other matched nodes around the matched nodes v_i^1 and v_j^2 after the interaction can be calculated by Eq. (4),

$$F_{ij} = d_j^2(1 - \frac{d_i^1}{N})\eta_2 + d_i^1(1 - \frac{d_j^2}{N})\eta_1 + \frac{d_i^1 d_j^2}{N}. \tag{4}$$

Since real-world complex networks always have a very huge number of nodes and a relatively small average degree, Eq. (2)-Eq. (4) can be further simplified to Eq. (5)-Eq. (7) respectively,

$$\widetilde{d_i^1} \approx d_i^1 + \eta_2 d_j^2, \tag{5}$$

$$\widetilde{d_j^2} \approx d_j^2 + \eta_1 d_i^1, \tag{6}$$

$$F_{ij} \approx \eta_1 d_i^1 + \eta_2 d_j^2. \tag{7}$$

Then we get Eq. (8) as

$$F_{ij} \approx \begin{cases} \frac{\eta_1(1-\eta_2)}{1-\eta_1\eta_2}\widetilde{d_i^1} + \frac{\eta_2(1-\eta_1)}{1-\eta_1\eta_2}\widetilde{d_j^2}, & \eta_1\eta_2 < 1; \\ \frac{1}{2}\widetilde{d_i^1} + \frac{1}{2}\widetilde{d_j^2}, & \eta_1\eta_2 = 1. \end{cases} \tag{8}$$

With the reason that the matched nodes are supposed unknown beforehand in reality, it seems unpractical to sort all the pairs of matched nodes by F_{ij} in descending order in order to improve the final matching precision ϕ, although larger F_{ij} corresponds to more pairs of unrevealed matched nodes around a pair of revealed matched nodes v_i^1 and v_j^2. Fortunately, Eq. (8) suggests a substitute way, i.e., selecting nodes with larger degree in the reference network, revealing their matched nodes in the other network by some dedicated methods, then these pairs of matched nodes are set to the revealed matched nodes.

3.1 Large degree priority strategies

Based on this principle, we proposed large degree priority strategies (Xuan & Wu, 2009) for the optimal node matching algorithm, as described by

- **Large Degree Priority in G_1 (LDP1)**: G_1 is selected as the reference network, where the nodes are sorted by their degrees in descending order, and the top P_r of them as well as their matched nodes in G_2 are selected as the revealed matched nodes.
- **Large Degree Priority in G_2 (LDP2)**: G_2 is selected as the reference network, where the nodes are sorted by their degree in descending order, and the top P_r of them as well as their matched nodes in G_1 are selected as the revealed matched nodes.

But which of them can bring higher matching precision? Can we answer this question just by comparing the structural properties (in particular, the degree sequences) of the two interactional networks? Without loss of generality, for a pair of interactional networks, suppose G_1 has larger average degree than G_2, i.e., $\langle \widetilde{d^1} \rangle > \langle \widetilde{d^2} \rangle$. Multiply Eq. (5) by η_1 and minus Eq. (6), we get

$$\eta_1 \langle \widetilde{d^1} \rangle - \langle \widetilde{d^2} \rangle = (\eta_1 \eta_2 - 1) \langle d^2 \rangle. \tag{9}$$

Since $\eta_1 \eta_2 \leqslant 1$, the value of η_1 can be roughly estimated by

$$\eta_1 \leqslant \frac{\langle \widetilde{d^2} \rangle}{\langle \widetilde{d^1} \rangle}, \tag{10}$$

while the value of η_2 cannot be estimated just by comparing the structural properties of the interactional networks. Suppose that the nodes are sorted by their degrees in descending order, denote by $R^i (i = 1, 2)$ the set of top P_r nodes in G_i, then from Eq. (8), we can see that more structural information may be provided when G_2 is selected as the reference network, if it is satisfied that

$$\eta_1 \sum_{v_i^1 \in R^1} \widetilde{d_i^1} + (1 - \eta_1) P_r \langle \widetilde{d^2} \rangle < \eta_1 P_r \langle \widetilde{d^1} \rangle + (1 - \eta_1) \sum_{v_i^2 \in R^2} \widetilde{d_i^2}, \tag{11}$$

which is equivalent to

$$\eta_1 \Big(\sum_{v_i^1 \in R^1} \widetilde{d_i^1} - P_r \langle \widetilde{d^1} \rangle \Big) + (1 - \eta_1) \Big(P_r \langle \widetilde{d^2} \rangle - \sum_{v_i^2 \in R^2} \widetilde{d_i^2} \Big) < 0. \tag{12}$$

Because it is always satisfied that

$$\sum_{v_i^1 \in R^1} \widetilde{d_i^1} \geqslant P_r \langle \widetilde{d^1} \rangle, \sum_{v_i^2 \in R^2} \widetilde{d_i^2} \geqslant P_r \langle \widetilde{d^2} \rangle, \tag{13}$$

Eq. (12) must be satisfied if we have

$$\frac{\langle \widetilde{d^2} \rangle}{\langle \widetilde{d^1} \rangle} \Big(\sum_{v_i^1 \in R^1} \widetilde{d_i^1} - P_r \langle \widetilde{d^1} \rangle \Big) + \Big(1 - \frac{\langle \widetilde{d^2} \rangle}{\langle \widetilde{d^1} \rangle} \Big) \Big(P_r \langle \widetilde{d^2} \rangle - \sum_{v_i^2 \in R^2} \widetilde{d_i^2} \Big) < 0. \tag{14}$$

where all the parameters are known when two interactional networks are provided. That is, only when Eq. (14) is satisfied, we can say that LDP2 may be superior to LDP1.

3.2 Centralized large degree priority strategies

The above LDP strategies are designed for optimal node-matching algorithms, while for iterative node-matching algorithms, these strategies need to be further modified. Because in this case, 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. Correspondingly, we propose two centralized large degree priority strategies specially for iterative node-matching algorithms (Xuan, Du & Wu, 2010b):

- **Centralized Large Degree Priority in G_1 (CLDP1)**. G_1 is selected as the reference network, where a set R_1 ($|R_1| = P_r$) of nodes are picked up according to their degrees by 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 $V_1 \setminus (U_1 \cup R_1)$) is at least connected to one node in R_1, at each time the nodes in $V_1 \setminus R_1$ 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.
- **Centralized Large Degree Priority in G_2 (CLDP2)**. G_2 is selected as the reference network, where a set R_2 ($|R_2| = P_r$) of nodes are picked up according to their degrees by following process. The node of the largest degree in G_2 is firstly selected as the only member of R_2. Denoting the neighbor set of R_2 as U_2 ($U_2 \cap R_2 = \emptyset$), i.e., each node in U_2 (but none of the nodes in $V_2 \setminus (U_2 \cup R_2)$) is at least connected to one node in R_2, at each time the nodes in $V_2 \setminus R_2$ are sorted by the number of neighbors belonging to U_2 in descending order and the top one is selected to join in R_2. Update R_2 and U_2 and repeat the selecting process until the set R_2 contains exactly P_r nodes. Then the set R_2 of nodes in G_2 as well as their matched nodes in G_1 are selected as the revealed pairwise matched nodes.

4. Node-matching algorithms

4.1 Similarities between nodes of interactional networks

Name	Definition
Common Neighbors (Newman, 2001)	$S_{ij} = n_M(v_i^1, v_j^2)$
Salton Index (Salton & McGill, 1983)	$S_{ij} = \frac{n_M(v_i^1,v_j^2)}{\sqrt{n_L(v_i^1)\times n_L(v_j^2)}}$
Jaccard Index (Jaccard, 1901)	$S_{ij} = \frac{n_M(v_i^1,v_j^2)}{n_L(v_i^1)+n_L(v_j^2)-n_M(v_i^1,v_j^2)}$
Sørensen Index (Sørensen, 1948)	$S_{ij} = \frac{2n_M(v_i^1,v_j^2)}{n_L(v_i^1)+n_L(v_j^2)}$
Hub Promoted Index (Ravasz et al., 2002)	$S_{ij} = \frac{n_M(v_i^1,v_j^2)}{\min\{n_L(v_i^1),n_L(v_j^2)\}}$
Hub Depressed Index (Lü & Zhou, 2011)	$S_{ij} = \frac{n_M(v_i^1,v_j^2)}{\max\{n_L(v_i^1),n_L(v_j^2)\}}$

Table 2. Several definitions of similarities between nodes of interactional networks based on their local structural information.

The similarity between two nodes belonging to different networks can be measured by the number of pairs of revealed matched nodes around them, e.g., the number of common friends they contact with in different communication networks, where a common friend is denoted by a pair of revealed matched nodes in corresponding communication networks. Denote by $n_L(v_i^1)$ and $n_L(v_j^2)$ the numbers of links connected to the node v_i^1 and v_j^2 in the networks G_1 and G_2, respectively, and by $n_M(v_i^1, v_j^2)$ the number of pairs of revealed matched nodes (v_k^1, v_k^2) where v_i^1 and v_j^2 are mutually 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. Then the similarity between v_i^1 and v_j^2 can be calculated by a number of methods (Jaccard, 1901; Lü & Zhou, 2011; Newman, 2001; Ravasz et al., 2002; Salton & McGill, 1983; Sørensen, 1948), as presented in Table. 2. Here, we adopt Jaccard Index to calculate the similarities between nodes of interactional networks.

4.2 Optimal node-matching algorithm

When revealed pairwise matched nodes are selected by LDP strategies, the similarity of each pair of the remaining nodes from different interactional networks can be calculated by Jaccard Index. Then, reviewing the definitions in Section 2.1, the node-matching problem between G_1 and G_2 can be transferred to a maximum matching problem for the bipartite graph $G_b = (U_1, U_2, W)$ where $U_i = \{v_{P_r+1}^i, v_{P_r+2}^i, \ldots, v_N^i\}$ $(i = 1, 2)$, and W denotes the set of links weighted by the similarities between these two groups of nodes. Without loss of generality, under the assumption $N_1 \leqslant N_2$, the task is to find a set of nonadjacent weighted links $\{w_1, w_2, \ldots, w_{N_1-P_r}\}$ to maximize the sum of their weights $\sum_{i=1}^{N_1-P_r} s_i$, which can be solved by the classical KM algorithm (Kuhn, 2005; Munkres, 1957). Note that, although the KM algorithm was developed for the case $N_1 = N_2$, it could be also feasible in the case $N_1 < N_2$ through factitiously adding $N_2 - N_1$ isolated nodes in G_1. For this reason we supposed $N_1 = N_2 = N$ for simplicity.

Since the KM algorithm has relatively high complexity $O(N^3)$, the sizes of the test networks cannot be very large. Here the two interactional networks G_1 and G_2 are both created by the BA model with $N = 100$ nodes and average degree $\langle k \rangle = 8$. Then they interact with each other with different interactional degrees $\eta_1 = 0.9$ and $\eta_2 = 0.1$ by the model shown in Fig. 1 (b). Denote the sample ratio by $\gamma = P_r/N$, the matching results are shown in Fig. 2, where we can see that, in most cases, LDP1 is prior to LDP2. This result is reasonable because when $\eta_1 \gg \eta_2$, Eq. (8) suggests that larger F_{ij} can be expected when select those nodes with large degrees in G_1 and their correspondences in G_2 as the revealed matched nodes. Note that, in this experiment, we set $M = N$ for simplicity, that is, every node in one network has its correspondence in the other network. In reality, M may be smaller than N, i.e., some individuals may be active in only one of the interactional networks. In this case, we need further select $M - P_r$ pairs of matched nodes from $N - P_r$ pairs of matched nodes obtained by the node-matching algorithm. If the value of M is known a priori, we can simply sort $N - P_r$ pairs of matched nodes by their attached similarities, then select the top $M - P_r$ pairs with larger similarities as the final pairs of matched nodes. However, if M is unknown, we have to set a threshold $\theta \in [0, 1)$ beforehand, and those pairs of matched nodes with similarities larger than θ then are selected as the final pairs of matched nodes, which will not be further discussed here. That is, in the following studies, we always set $M = N_1 = N_2 = N$ for simplicity.

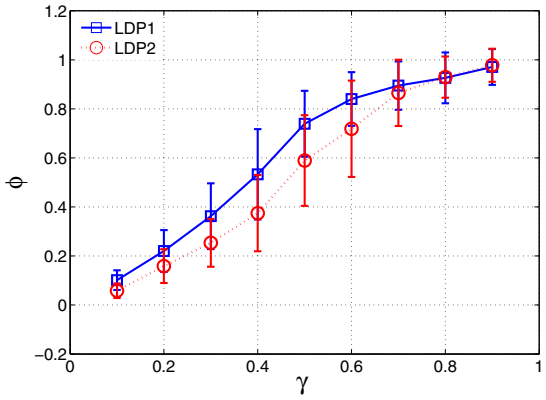


Fig. 2. The matching precision ϕ as the function of the sample ratio γ by adopting the two revealed matched nodes selection strategies, i.e., LDP1 and LDP2, for scale-free networks created by the BA model with $N = 100$ and $\langle k \rangle = 8$ and different interactional degrees $\eta_1 = 0.9$ and $\eta_2 = 0.1$ (Xuan & Wu, 2009). For each γ and each selection strategy, the experiment is implemented on 100 different pairs of scale-free networks, then the average matching precision as well as the error bar is recorded.

4.3 Iterative node-matching algorithm

As we can see in Fig. 2, the optimal node-matching algorithm fails to achieve acceptable results when there are only a relatively small number of pairwise matched nodes revealed beforehand, e.g., in order to achieve a matching precision of 80%, we have to reveal as many as 60% correspondences between nodes of the two networks in advance, which, as well as its long running time, hinders its efficient application in node-matching between real-world networks of quite large size. Based on the CDLP revealed matched nodes selecting strategies and Jaccard similarities between nodes of different networks, the iterative node-matching algorithm is simply composed of the following two steps (Xuan, Du & Wu, 2010b):

- **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, then recalculate the similarities between the remaining nodes, and so forth.
- **Termination.** The iterative process is terminated when all of the nodes in the interactional networks have been matched.

The time complexity of the above node-matching algorithm mainly depends on the recalculation of the similarities. Generally, once a pair of nodes from different networks are matched at $(\tau - 1)$th round, we need to recalculated the similarities of about $k_\tau^1 k_\tau^2$ pairs of nodes mutually connected to that pair of matched nodes at τth round, where k_τ^i $(i = 1, 2)$ represents the degree of the matched node in G_i at $(\tau - 1)$th round. Provided $N_1 = N_2 = M = N$, the running time of the algorithm, denoted by Γ, can be calculated by Eq. (15) statistically,

$$\Gamma \sim E(\sum_{\tau=1}^{N} k_\tau^1 k_\tau^2). \tag{15}$$

If the two networks under study are strongly dependent each other, i.e., extremely G_1 and G_2 are identical and a node in one network only can be matched to the node of equal degree in the other network, Eq. (15) can be replaced by Eq. (16),

$$\Gamma \sim \sum_{\tau=1}^{N} E((k_\tau^1)^2). \tag{16}$$

For scale-free networks generated by the BA model, the degree distribution follows $p(k) \sim k^{-3}$, thus the running time can be simplified by Eq. (17),

$$\Gamma \sim N \int_1^N k^2 k^{-3} dk \sim N \ln N. \tag{17}$$

However, if the two target networks are relatively independent from each other, i.e., a node with large degree in one network can be matched to a node with small degree in the other network, which is more common in reality, Eq. (15) can be approximatively transferred to Eq. (18),

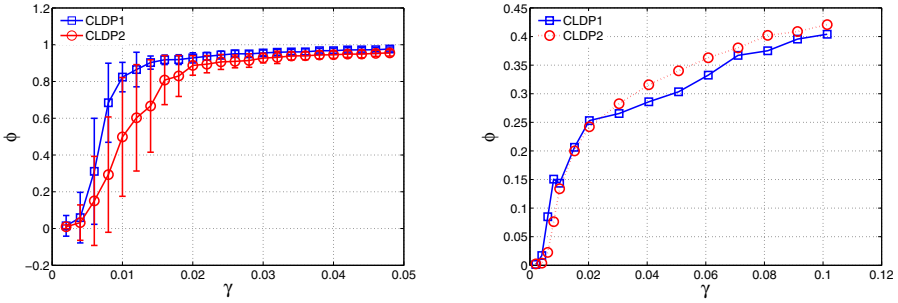
$$\Gamma \sim \sum_{\tau=1}^{N} E(k_\tau^1) E(k_\tau^2) \sim N \langle k^1 \rangle \langle k^2 \rangle, \tag{18}$$

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, Eq. (18) suggests a linear time complexity $O(N)$ of the algorithm (Xuan, Du & Wu, 2010b). Eqs. (17) and (18) mean that the iterative node-matching algorithm has much lower complexity than the optimal node-matching algorithm.

In order to compare to the optimal node-matching algorithm, here we take the same example to test the iterative node-matching algorithm. Since the iterative algorithm is able to solve node-matching problems between networks of quite large size, the two interactional networks G_1 and G_2 here are also created by the BA model with same average degree $\langle k \rangle = 8$, but much larger network size $N = 500$. Then these two networks interact with each other with different interactional degrees $\eta_1 = 0.9$ and $\eta_2 = 0.1$ by the same model shown in Fig. 1 (b). The matching results are show in Fig. 3 (a). At this time, in order to correctly reveal most of matched nodes in the networks (e.g., $\phi \geqslant 80\%$), we only need to have a very small percentage of matched nodes revealed beforehand (1% for CLDP1 and 1.6% for CLDP2), i.e., the iterative node-matching algorithm is far more efficient than the optimal node-matching algorithm on interactional artificial scale-free networks.

However, when we test this iterative node-matching algorithm on the real-world interactional chat network and friendship network introduced in Section 2.3, the matching results, as shown in Fig. 3 (b), are not that satisfactory, i.e. 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. For example, only about 40% matched nodes are revealed correctly, even though there are as many as 10% matched nodes are revealed beforehand. This phenomenon may be caused by the relatively high symmetry of the chat network and the friendship network. Generally, the local symmetry between the two non-linked nodes v_i and v_j in a network is defined by (Xuan, Du & Wu, 2010b)

$$\omega_{ij} = \frac{\chi_{ij}^c}{\chi_{ij}^t}, \tag{19}$$



(a) Matching results on artificial networks (b) Matching results on real-world networks

Fig. 3. The matching precision ϕ as the function of the sample ratio γ by adopting the two revealed matched nodes selection strategies, i.e., CLDP1 and CLDP2, for (a) the interactional scale-free networks created by the BA model with $N = 500$ and $\langle k \rangle = 8$ and different interactional degrees $\eta_1 = 0.9$ and $\eta_2 = 0.1$, and (b) the interactional real-world chat network and friendship network (Xuan, Du & Wu, 2010b). For artificial networks, the experiment is implemented on 100 different pairs of scale-free networks for each γ and each selection strategy, then the average matching precision as well as the error bar is recorded.

where χ_{ij}^c and χ_{ij}^t are the numbers of their common and total neighbors, respectively. If nodes v_i and v_j are connected, release the link and then calculate the symmetry between them following Eq. (8). Since it is impossible to distinguish two nodes v_i and v_j in a network with the symmetry $\chi_{ij} = 1$ (i.e. they share the same neighbors excluding themselves) just by adopting their topological information, those highly symmetric nodes in one network may be wrongly matched to the nodes in the other network with quite a high probability, and thus one-to-one node-matching algorithms may produce poor results in such situations.

4.4 One-to-many iterative node-matching algorithms

In order to overcome the above limitation of one-to-one node-matching algorithms, we proposed one-to-many node matching (Du et al., 2010) through expanding the number of nodes in each matching step. In fact, one-to-many node matching has its practical significance because it can help 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 defined by Eq. (20),

$$v_i^1 \to Q_i^2 = \{v_{i_1}^2, v_{i_2}^2, \ldots, v_{i_M}^2\}, \tag{20}$$

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 including the top M most likely matched nodes of v_i^1 in G_2. It should be noted that here 1-to-M match is just a natural generalization of 1-to-1 match, therefore, Eq. (20) also provides a consistent 1-to-1 match, 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 $i \neq j$. For a 1-to-M matching algorithm, a node v_i^1 in G_1 is considered correctly matched if its real matched node v_i^2 is contained in Q_i^2, i.e., $v_i^2 \in Q_i^2$. Denoting P_M ($P_M \leqslant N - P_r$) as the number of nodes in G_1 that are correctly matched, the matching precision ϕ_M for the 1-to-M node

matching algorithm can be calculated by Eq. (21), and naturally Eq. (22) is always satisfied.

$$\phi_M = \frac{P_M}{N - P_r}, \tag{21}$$

$$\phi_M \geq \phi_{M-1}. \tag{22}$$

Next, we will introduce two different one-to-many iterative node-matching algorithms (Du et al., 2010).

1) **A1: Local mapping**. Since the similarity between each pair of nodes may change as the one-to-one iterative algorithm is implemented step by step, it is possible to correct some initially wrongly matched nodes by recalculating their similarities after the one-to-one node matching algorithm is terminated. This fact leads to the first one-to-many node matching algorithm based on local mapping. In particular, the Algorithm A1 is defined by the following two steps (Du et al., 2010):

- **Iterative 1-to-1 node matching**. $N - P_r$ pairs of nodes, i.e., $v_i^1 \leftrightarrow Q_i^2 = \{v_{i_1}^2\}$ $(i = P_r + 1, P_r + 2, \ldots, N)$, are firstly matched by the iterative 1-to-1 node matching algorithm.
- **Candidate nodes selection**. Denote by X_i^1 the neighbor set of node v_i^1 in G_1, which has a matched node set X_i^2 in G_2 where the nodes are 1-to-1 matched to those in X_i^1, then denote by Y_i^2 the neighbor set of X_i^2, including all the nodes directly connected to those in X_i^2. Based on the definition of similarity, only the similarities between node v_i^1 $(i = P_r + 1, P_r + 2, \ldots, N)$ and the nodes in Y_i^2 can be larger than 0 and thus are recalculated. Then the top $M - 1$ nodes with largest similarities are selected as the candidate corresponding nodes of v_i^1. It should be noted that $v_{i_1}^2$ is not reconsidered here, and if Y_i^2 only contains fewer than $M - 1$ nodes, other $M - 1 - |Y_i^2|$ nodes can be randomly selected from G_2 to be consistent with Eq. (20).

2) **A2: Ensembling**. In the area of machine learning, it is a common way to improve the generalization performance of an algorithm by combining the results of many different predictors (Breiman, 1996; Freund & Schapire, 1997; Krogh & Sollich, 1997; Miyoshi et al., 2005). However, the above iterative one-to-one node matching algorithm is totally deterministic, i.e., for a given pair of target networks and certain revealed matched nodes, the algorithm must produce the same matching result. Therefore, it cannot be directly used for ensemble, and thus a new statistical iterative one-to-one node matching algorithm have to be introduced first, where a pair of newly revealed matched nodes is adopted only with probability $p (p < 1)$ to calculate the similarities between those unrevealed nodes of different networks in the succeeding iterative process. 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 obtained results can be merged into a unique one-to-many matching result by a voting strategy. In particular, the algorithm A2 is defined by the following three steps (Du et al., 2010):

- **Iterative 1-to-1 node matching**. $N - P_r$ pairs of nodes, i.e., $v_i^1 \leftrightarrow Q_i^2 = \{v_{i_1}^2\}$ $(i = P_r + 1, P_r + 2, \ldots, N)$, are firstly matched by the deterministic iterative 1-to-1 node matching algorithm.
- **Implement and vote**. The statistical 1-to-1 node matching algorithm with parameter $p (p < 1)$ is implemented for B $(B \gg M)$ rounds and a group of B different 1-to-1 matching results are obtained. All of the correspondences in G_2 of v_i^1 in G_1 are grouped by a node set Z_i^2

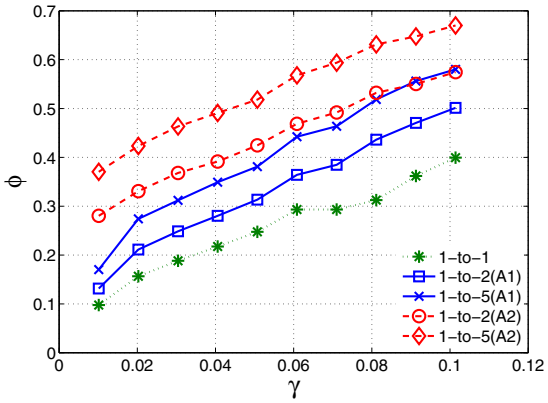


Fig. 4. The matching precision ϕ as the function of the sample ratio γ for $M = 1, 2, 5$ ($M = 1$ means the one-to-one matching result) between the friendship network and the chat network obtained from the database of *Alibaba trademanager* (Du et al., 2010). Here, the chat network is taken as the reference network.

with its size (the number of nodes) satisfying $|Z_i^2| \leq B$. It should be noted that each node in Z_i^2 is attached by a positive integer as its weight representing the times that it is matched to v_i^1 in the total B rounds, and similarly $v_{i_1}^2$ is excluded here.

- **Candidate nodes selection**. The top $M-1$ nodes with largest weights in Z_i^2 are selected as the $M-1$ candidate corresponding nodes of v_i^1. Sometimes, there may be only fewer than $M-1$ nodes in Z_i^2, i.e., $|Z_i^2| \leq M-1$, in such a situation, other $M-1-|Z_i^2|$ nodes can be randomly selected from G_2 to be consistent with Eq. (20).

Similarly, these two one-to-many iterative node matching algorithms are tested on the real-world interactional chat network and friendship network introduced in Section 2.3, and the matching results are shown in Fig. 4. As we can see, both the proposed one-to-many algorithms (especially the random algorithm A2) can significantly improve the matching precision, and thus can be considered to partially overcome the limitation of one-to-one node-matching algorithms.

5. Conclusion

Since an individual may appear in different systems with different identities, many real-world complex systems are considered to be interacted with each other all the time. Revealing these identities of the same individual is a common task in many areas such as sociology, linguistics, biology, etc, by their dedicated methods. When these complex systems are described by networks, this common task can be changed to a node matching problem between different complex networks, and thus can be solved in the framework of graph theory.

In this chapter, we reviewed the overall process to solve such node-matching problems between different networks: We first calculated the similarities between nodes of different networks through their connections to several pairs of preliminarily revealed matched nodes and transferred the node matching problem between two different networks to a maximum

weighted bipartite matching problem; then we proposed several node-matching algorithms to solve such problem. By comparison, the iterative node-matching algorithm has approximately linear complexity and behaves much better than the traditional KM algorithm in graph theory. However, it seems that almost all of the network structure-based one-to-one node-matching algorithms lose their efficiencies when the target networks are highly symmetric, e.g., the iterative node-matching results are not that good on real-world chat network and friendship network obtained from the database of *Alibaba trademanager*. Such limitation can be partially overcome by the proposed one-to-many node-matching algorithms, which mainly focus on quickly narrowing down the searching range, rather than revealing exact one-to-one mapping between nodes of different networks. Meanwhile, we also introduced several degree-based revealed matched nodes selecting strategies for optimal and iterative node-matching algorithms, respectively, in order to further improve the matching results. In the future, more information about individuals and connections may be adopted to create more efficient node-matching algorithms.

6. References

Albert, R., Jeong, H., & Barabási, A.-L. (2000). Error and attack tolerance of complex networks, *Nature* 6794(406): 378–382.

Barabási, A.-L. (2009). Scale-free networks: A decade and beyond, *Science* 325(5939): 412–413.

Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks, *Science* 286(5439): 509–512.

Barabási, A.-L. & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization, *Nature* 5(2): 101–113.

Barrat, A., Barthélemy, M., Satorras, R. P. & Vespignani, A. (2004). The architecture of complex weighted networks, *Proceedings of the National Academy of Sciences U.S.A* 101(11): 3747–3752.

Breiman, L. (1996). Bagging predictors, *Machine Learning* 26(2): 123–140.

Cootes, A. P., Muggleton, S. H. & Sternberg, M. J. E. (2007). The identification of similarities between biological networks: Application to the metabolome and interactome, *Journal of Molecular Biology* 369(4): 1126–1139.

Costa, L. D. F., Rodrigues, F. A., Travieso, G. & Boas, P. R. V. (2007). Characterization of complex networks: A survey of measurements, *Advances in Physics* 56(1): 167–242.

Crucitti, P., Latora, V., Marchiori, M. & Rapisarda, A. (2004). Error and attack tolerance of complex networks, *Physica A: Statistical Mechanics and Its Applications* 340(1-3): 388–394.

Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. (2008). Critical phenomena in complex networks, *Review of Modern Physics* 80(4): 1275–1335.

Du, F., Xuan, Q. & Wu, T.-J. (2010). One-to-many node matching between complex networks, *Advances in Complex Systems* 13(6): 725–739.

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M. & Apkarian, A. V. (2005). Scale-free brain functional networks, *Physical Review Letters* 94(1): 018102.

Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences* 55(1): 119–139.

Giunchiglia, F. & Shvaiko, P. (2004). Semantic matching, *The Knowledge Engineering Review* 18: 265–280.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des alpes et des jura, *Bulletin de la Société Vaudoise des Science Naturelles* 37: 547–579.

Kelley, B. P., Sharan, R., Karp, R. M., Sittler, T., Root, D. E., Stockwell, B. R. & Ideker, T. (2003). Conserved pathways within bacteria and yeast as revealed by global protein network alignment, *Proceedings of the National Academy of Sciences U.S.A* 100(20): 11394–11399.

Krogh, A. & Sollich, P. (1997). Statistical mechanics of ensemble learning, *Physical Review E* 55(1): 811–825.

Kuhn, H. W. (2005). The hungarian method for the assignment problem, *Naval Research Logistics* 52(1): 7–21.

Li, X. & Chen, G. (2003). A local-world evolving network model, *Physica A: Statistical Mechanics and Its Applications* 328(1-2): 274–286.

Lü, L. & Zhou, T. (2011). Link prediction in complex networks: A survey, *Physica A: Statistical Mechanics and Its Applications* 390(6): 1150–1170.

Miyoshi, S., Hara, K. & Okada, M. (2005). Analysis of ensemble learning using simple perceptrons based on online learning theory, *Physical Review E* 71(3): 036116.

Mossa, S., Barthélémy, M., Stanley, H. E. & Amaral, L. A. N. (2002). Truncation of power law behavior in scale-free network models due to information filtering, *Physical Review Letters* 88(13): 138701.

Motter, A. E. & Lai, Y.-C. (2002). Cascade-based attacks on complex networks, *Physical Review E* 66(6): 065102.

Motter, A. E., Nishikawa, T. & Lai, Y.-C. (2003). Large-scale structural organization of social networks, *Physical Review E* 68(3): 036105.

Munkres, J. (1957). Algorithms for the assignment and transportation problems, *Journal of the Society for Industrial and Applied Mathematics* 5(1): 32–38.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks, *Physical Review E* 64(2): 025102.

Newman, M. E. J., Forrest, S. & Balthrop, J. (2002). Email networks and the spread of computer viruses, *Physical Review E* 66(3): 035101.

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J. & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks, *Proceedings of the National Academy of Sciences U.S.A* 104(18): 7332–7336.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks, *Science* 297(5586): 1551–1555.

Rozenfeld, H. D., Song, C. & Makse, H. A. (2010). Small-world to fractal transition in complex networks: A renormalization group approach, *Physical Review Letters* 104(2): 025701.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*, MuGraw-Hill, Auckland.

Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons, *Biologislce Skrifler* 5(4): 1–34.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks, *Nature* 393(6684): 440–442.

Xiao, Y., Xiong, M., Wang, W. & Wang, H. (2008). Emergence of symmetry in complex networks, *Physical Review E* 77(6): 066108.

Xuan, Q., Du, F. & Wu, T.-J. (2009). Empirical analysis of internet telephone network: From user id to phone, *Chaos* 19(2): 023101.

Xuan, Q., Du, F. & Wu, T.-J. (2010a). Partially ordered sets in complex networks, *Journal of Physics A: Mathematical and Theoretical* 43(18): 185001.

Xuan, Q., Du, F. & Wu, T.-J. (2010b). Partially ordered sets in complex networks, *Journal of Physics A: Mathematical and Theoretical* 43(39): 395002.

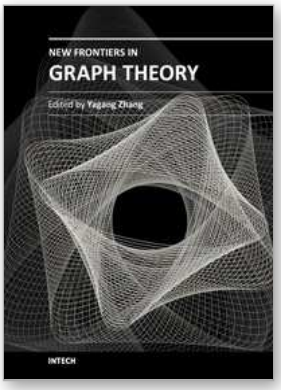
Xuan, Q., Du, F., Wu, T.-J. & Chen, G. (2010). Emergence of heterogeneous structures in chemical reaction-diffusion networks, *Physical Review E* 82(4): 046116.

Xuan, Q., Li, Y. & Wu, T.-J. (2006). Growth model for complex networks with hierarchical and modular structures, *Physical Review E* 73(3): 036105.

Xuan, Q., Li, Y. & Wu, T.-J. (2007). A local-world network model based on inter-node correlation degree, *Physica A: Statistical Mechanics and Its Applications* 378(2): 561–572.

Xuan, Q., Li, Y. & Wu, T.-J. (2008). Does the compelled cooperation determine the structure of a complex network?, *Chinese Physics Letters* 25(2): 363–366.

Xuan, Q. & Wu, T.-J. (2009). Node matching between complex networks, *Physical Review E* 80(2): 026103.



New Frontiers in Graph Theory

Edited by Dr. Yagang Zhang

ISBN 978-953-51-0115-4
Hard cover, 526 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

Nowadays, graph theory is an important analysis tool in mathematics and computer science. Because of the inherent simplicity of graph theory, it can be used to model many different physical and abstract systems such as transportation and communication networks, models for business administration, political science, and psychology and so on. The purpose of this book is not only to present the latest state and development tendencies of graph theory, but to bring the reader far enough along the way to enable him to embark on the research problems of his own. Taking into account the large amount of knowledge about graph theory and practice presented in the book, it has two major parts: theoretical researches and applications. The book is also intended for both graduate and postgraduate students in fields such as mathematics, computer science, system sciences, biology, engineering, cybernetics, and social sciences, and as a reference for software professionals and practitioners.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821