An Integer Programming Approach and Visual Analysis for Detecting Hierarchical Community Structures in Social Networks

Chun-Cheng Lin[a,*], Jia-Rong Kang[a], Jyun-Yu Chen[a]

[a]*Department of Industrial Engineering and Management, National Chiao Tung University, Hsinchu 300, Taiwan*

Abstract

Detecting community structures in social networks is a very important task in social network analysis as these community structures explain relationships among individuals and can be used to predict social behavior. The relationship among subcommunities in each community can further be identified as hierarchical community structures, in which each super node at each hierarchical level represents a nested structure of communities or nodes. Most previous studies attempting to detect hierarchical community structures focused on new metaheuristics that are computationally efficient but do not guarantee the optimal community partition. As a result, this work applies a novel integer programming (IP) approach to detect hierarchical community structures in social networks. This approach has flexible community capacity limits, does not limit the community numbers at different levels, and maximizes a quality measure for hierarchical community partition. The proposed IP approach can use existing software solvers to detect hierarchical community structures without implementing an algorithm. Visual analysis of experimental results shows that the proposed model with different settings for level numbers can analyze reasonable and sophisticated hierarchical community structures, such that the relationships between communities at different levels can be elucidated clearly.

[☆]A preliminary work was presented at the 2013 IEEE Conference on Industrial Engineering and Engineering Management (IEEM 2013), Dec. 10-13, 2013, Bangkok, Thailand.

[*]Corresponding author. Phone: +886-3-5731758. Fax: +886-3-5729101.

Email address: cclin321@nctu.edu.tw (Chun-Cheng Lin)

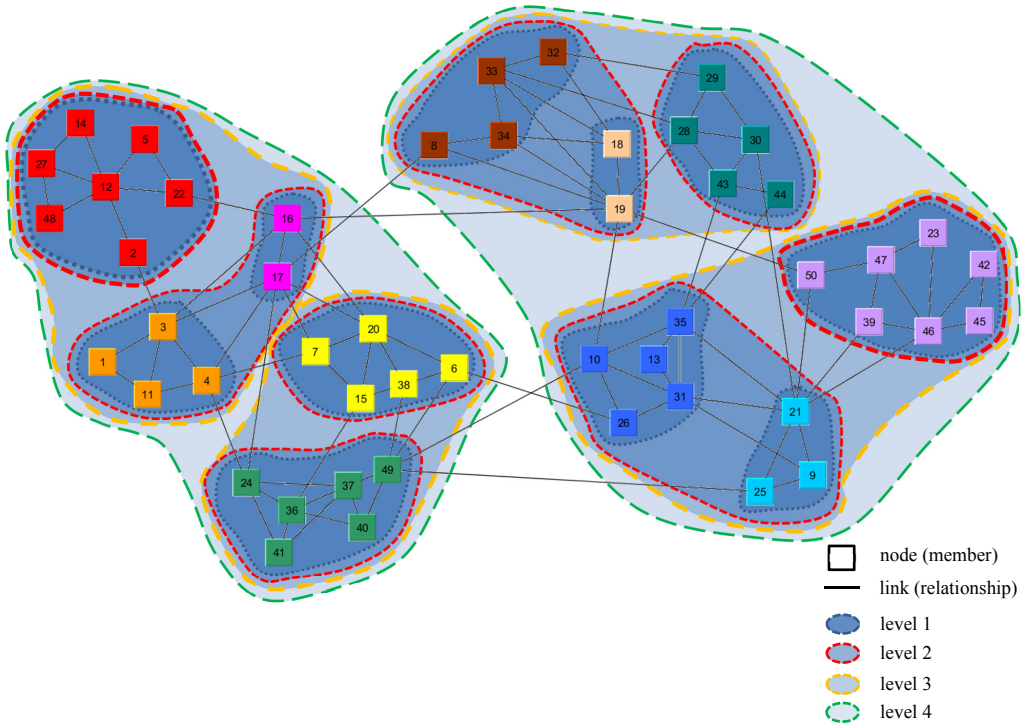
Keywords: Social network, community detection, hierarchical community structure, integer programming, visual analysis.

1. Introduction

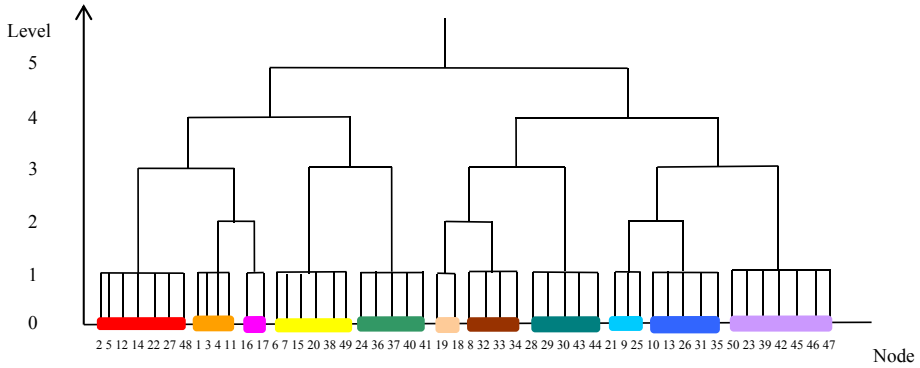
A social network can be regarded as a graph in which each individual is represented as a node, and the social relationship between two individuals is represented as a link between the two corresponding nodes [48]. Figure 1(a) shows an example of nodes and their connections. Recent social network analysis has become increasingly popular, as it helps elucidate the social behavior and backgrounds of individuals [2, 41, 46]. Some previous studies applied the cluster analysis to social networks, in which the nodes within each cluster are strongly connected with each other (*i.e.*, interactions within the same cluster are strong), and the links between two different clusters are connected weakly [10, 20, 47]. Extended from cluster analysis, the individuals with similar backgrounds or interests interact frequently and generally gather together to form one or several communities. A structure with multiple communities for the social network is called a *community structure* [6, 7, 27], *e.g.*, nodes with the same color that belong to the same community, which is encircled by blue dots (Figure 1(a)). Hence, differing from previous studies using cluster analysis, the degree of similarity is regarded as the evaluation criterion for detecting/partitioning community structures in social or complex networks [28].

Recent works further investigated the hierarchical relationships in community structures. Each community may have subcommunities based on the strong similarities among individuals, such that the hierarchical relationship forms a nested structure called a *hierarchical community structure* [5, 11, 35] (Figure 1(a)), and the hierarchical relationship is represented by a tree-like dendrogram (Figure 1(b)), in which each community is represented by a node and the hierarchical relationship is represented as a tree. The hierarchical community structure provides information about a community partition and the hierarchical level division for large-scale complex social networks, such that sophisticated social behaviors and interactive relationships among each node can be realized. Hence, this hierarchical community structure has garnered considerable attention and has been applied in both science and engineering.

Most previous works were lacking in three important ways. First, to the best of our knowledge, no mathematical programming methods have been



(a)



(b)

Figure 1: The hierarchical community structure of the Facebook network instance detected by the proposed IP approach with the setting of 6 hierarchical levels. (a) Drawing the detected structure, and (b) dendrogram of the detected structure.

developed to detect hierarchical community structures. Most previous works designed heuristics or metaheuristics to generate approximate solutions for hierarchical community detection in social or complex networks [11, 32, 33, 36, 37, 45]. Although those approaches are computationally efficient, they cannot guarantee that exact optimal solutions will be obtained. Notably, [21, 43] developed the mathematical programming methods for detecting community structures without any hierarchy.

Second, most previous works (*e.g.*, [32, 36, 37]) did not apply flexible community capacity limits for different hierarchical levels. In a hierarchical community structure, the community capacity limit markedly affects the quality of detecting the community structure at each hierarchical level. For larger networks, when the community capacity limit is too small, a hierarchical community structure with too many similar communities at the same hierarchical level is typically detected; for small networks, when the community capacity limit is too large, a community structure with different community sizes will be detected. Although some studies (*e.g.*, [43]) used the community capacity limit to avoid excessively large differences between communities, the community capacity limit was fixed for all hierarchical levels.

Third, many works applied an upper bound for the number of communities at each hierarchical level. Generally, the number of communities at different hierarchical levels differs, and, hence, most works (*e.g.*, [11, 33, 45]) set an upper bound for the number of communities at each hierarchical level according to experience. However, an upper bound that is too large or too small tends to generate unreasonable community structures, adversely affecting the structure's quality.

This work applies a novel integer programming (IP) approach to detect hierarchical community structures in social networks. This approach has flexible community capacity limits for different hierarchical levels and the number of communities at each hierarchical level is not restricted, such that reasonable hierarchical community structures are detected efficiently and effectively according to a predetermined number of hierarchical levels. The hierarchical community structure facilitates observation of the information for the community partition and the hierarchical level divisions. Additionally, the proposed approach overcomes the three main limitations. Finally, visual analysis of detected hierarchical community structures for three real social network instances helps elucidate detection results. Comparison with approaches in previous studies shows that the proposed approach performs better.

The remainder of this paper is organized as follows. Section 2 reviews related studies, including those focused on detecting community structures and hierarchical community structures. Section 3 describes the proposed IP approach. Section 4 analyzes experimental results for two benchmark network instances as well as a real social network instance collected from the Facebook website. Section 5 gives conclusion and directions for future work.

2. Related Works

This section discusses the detection of community structures and hierarchical community structures in social networks or complex networks. In addition, the similarity or quality measures that assess community structures are also examined.

2.1. Detection of Community Structures

Various methods for detecting community structures based on conventional graph clustering methods have been proposed, including the hierarchical clustering algorithm [12, 8, 25], graph partitioning [1], k-means clustering [23], particle swarm optimization [3], and bee colony optimization [18]. Subsequently, novel methods for detecting community structures were proposed. For example, [43] proposed an IP method that applied a community capacity limit and minimal difference for numbers of nodes in different communities. In [13], a weight was assigned to each node in a network, and then the conventional k-means clustering method was applied to classify nodes into k communities. A nonlinear programming model was established in [21], which also adopted the Lagrangian method to reduce time complexity of computing the model. In [40], a novel algorithm was proposed to detect dynamic communities in social network, such that the effect of noisy data was eliminated, and the real community structure and abnormal events were discovered. In [46], a method with supervised learning mechanism was proposed to incorporate prior information into community structure detection.

Metaheuristics have been recently employed to detect community structures. In [9], particle swarm optimization was applied to detect community structures, and experimental results showed that the isolated nodes can be detected with increased ease. A genetic algorithm (GA) developed in [31] to solve the community detection problem with a single-objective function based on community score; the quality of the detected community structures was better than that of the previous GA approaches. A GA approach was

also proposed in [15] to solve the community structure detection problem with a multi-objective function, improving the single-objective function designed in [31]. The k-means algorithm was first adopted by [22] to find an initial solution, which considered link length between nodes in a network, and then applied the simulated annealing algorithm to detect community structures. Experimental results showed that aside from visualizing the detected community structure, the most important node in each community was also found. In [16], ant colony optimization algorithm was applied to detect community structures, in which each ant represented a network node; each ant then iteratively selected whether to join a community; and finally, a community structure was found after each ant determined its community.

To assess the quality of detected community structures, various similarity or quality measures were applied, including internal density [19], normalized mutual information (NMI) [6], and function-modularity intensity [38]. To minimize difference among different communities, [19] used the proposed *internal density* to compute the density of links in each community, and attempted to balance the link densities of all communities. Based on the theory of probability and statistics, [6] proposed the NMI, which measures the length of the correlation between each pair of communities in the detected community structure. As the NMI value increases, the quality of the community structure increases.

Aside from these measures for detected community structures, the *modularity metric* (Q value) [26, 28] has been a very popular quality measure. The Q value is the sum of differences of the fraction of all links within each community minus the expected value of the same quantity in a network in which nodes have the same degrees but links are placed randomly, which is computed as follows:

$$Q = \sum_{k=1}^{K} \left(\frac{L_k}{L} - (\frac{D_k}{2L})^2 \right) \tag{1}$$

where K is the number of communities; L is the number of links in a network; L_k is the number of the links in the k-th community, *i.e.*, the two end nodes of each link are allocated to the k-th community; D_k is the sum of the degrees of nodes in the k-th community. Such a notation is derived from assessing the differences in the expected values of the link ratios before and after community detection. When the Q value is 0, nodes in the network are distributed at random partitions and no obvious community structure exists; when the Q value approaches 1, the network has a strong and obvious

community structure.

2.2. Detection of Hierarchical Community Structures in Social Networks

Many studies found that the detected community structure may have sophisticated relations, including hierarchical community relationships [5, 11, 35] and overlapping community relationships [29, 30]. In hierarchical community structures, a hierarchical relationship exists among communities, *i.e.*, a community can be divided further into multiple subcommunities. On the other hand, community structures overlap, some nodes in the network may belong to more than one community simultaneously, and the overlapping nodes may be classified into multiple communities. Since this work focuses on only hierarchical community structures, only major studies on these structures are reviewed.

Aside from information about distributing nodes to different communities, hierarchical community structures can further provide information on community distribution at different hierarchical levels to assist in identifying the inner structure of a network. Hence, hierarchical community structures have been detected in many complex networks. As well, numerous studies (*e.g.*, [32, 36, 4, 34, 14]) have proposed approaches to detect hierarchical structures in complex networks, and applied their approaches to real social networks. Notably, [32] improved the agglomerative algorithm and objective function, and applied a two-stage algorithm to detect hierarchical community structures; the first stage improved objective function to determine the number of communities in the network; and then the second stage determined the total number of hierarchical levels. An algorithm that repeatedly merges nodes according to a similarity matrix and a probability matrix was developed by [36], and the algorithm detected hierarchical community structures for a karate club network and a network of college football games. A generic model was developed by [4] to generate an arbitrary hierarchical structure for a random graph, which can provide a mathematically principled means to learn about the hierarchical organizations of real-world graphs. Their model was also applied to the karate club network and college football games network, and the results showed that their model is suitable for small graphs. A model to detect the hierarchical community structure for social networks was developed by [42]. In this model, the link between two members indicates that the two members exchanged at least one letter. Their model aimed to determine the conditions under which the average length of a message chain connecting a randomly selected sender to a random addressee is

small. Boettcher and Percus [34] constructed a hierarchical network model that combines two generic properties of hierarchical organization—scale-free and a high degree of clustering—and then analyzed the relationships among actors in the same Hollywood movies. A procedure was developed by [14] for detecting hierarchical communities in a real social network based on the communication relationship of emails in a medium-sized university. In the email network, each email address represented a node, and the communication between two nodes represented a link. By analyzing the email network, accurate and nonintuitive description of the flow of information within human organizations was obtained.

3. Proposed IP Approach for Detecting Hierarchical Community Structures

This section describes the problem of detecting hierarchical community structures, and then proposes the IP approach.

3.1. Problem Description

Consider a social network that consists of nodes and links, in which each node represents an individual in the social network, and each link represents the social relationship between the two end node individuals. Given such a social network (Figure 2(a)) and the number of hierarchical levels (*e.g.*, 4 levels are predetermined (Figure 2)(b)), the problem considered is to establish an IP model for detecting a hierarchical community structure for a social network (*e.g.*, Figures 2(b) and 2(c)) according to the number of its hierarchical levels.

3.2. The Proposed IP Approach

The proposed IP approach has the following three main attributes:

- *Flexible community capacity limits for different hierarchical levels*: Regardless of the number of levels and network size, previous studies assumed a fixed community capacity limit, *i.e.*, this limit restricts the fixed maximal number of links allocated to each community, such that the number of links in each community is increasingly uniform. However, if the applied fixed community capacity limit is too small for a large network, almost all detected communities at the same hierarchical level have similar numbers of links and therefore look very similar. Conversely, if a fixed community capacity limit that is too large is applied

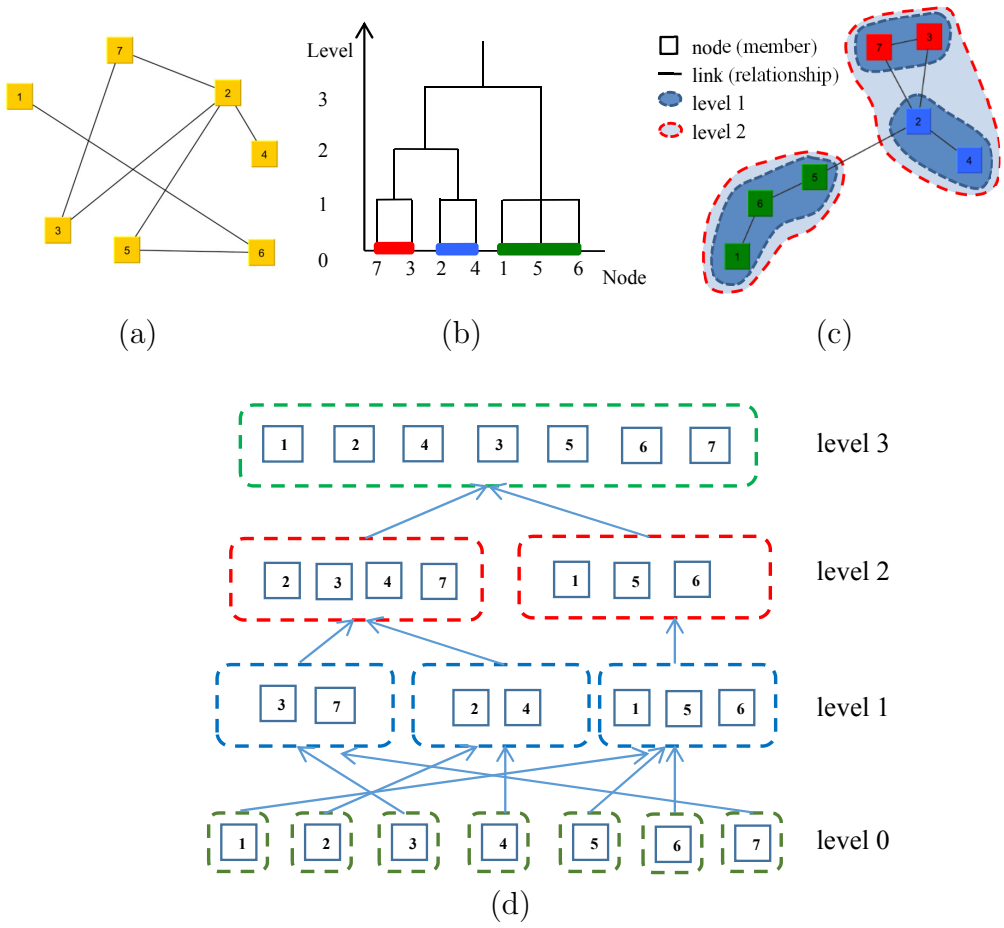


Figure 2: (a) A simple network instance with 7 nodes and 7 links. (b) The dendrogram of the detected hierarchical community structure for the network instance. (c) The drawing of the detected hierarchical community structure. (d) A construction feature of the IP model for detecting the hierarchical community structure of this simple network.

to a small network, the detected communities at the same hierarchical level could be excessively different, and some of those communities may contain too few nodes. Therefore, we assume flexible community capacity limits for different hierarchical levels, and these limits depend on number of levels and network size.

- *No upper bound for the number of communities at each hierarchical level*: Previous work assumed an upper bound for the number of com-

munities at each hierarchical level. When restricted to this bound, some communities at a certain hierarchical level cannot be detected, and furthermore, the community detection at the adjacent higher level is adversely affected. The proposed IP approach does not have this bound; therefore any number of communities, within reason, can be detected.

- *Allow the user to predetermine the total number of hierarchical levels*: By predetermining the number of hierarchical levels, the community structures with different numbers of hierarchical levels and different community constitutions at each hierarchical level could be detected by the proposed IP model, and an increased number of insights into the target network instance can be observed.

According to the scale of the target network instance, the network is classified into $T+1$ hierarchical levels, labeled $0, 1, ..., T$ from lowest to highest (Figure 2(b)). Note that T is predetermined by the user according to the network scale. In our IP approach, the number of communities at level 0 equals the number of nodes, and each community at level 0 contains only one node. The IP approach can be characterized as a method of allocating nodes to super communities at each level in a bottom-up fashion, even though the IP solver may not perform in this order. Let L be the number of links in the network. For each $t = 1, 2, ..., T$, the community capacity limit at level t is set as $L/(t/T)$, and then communities or nodes are allocated to different super communities under this limit, while the Q value at this level is maximized. Finally, all communities at level $T-1$ are merged into a single super community at level T.

In the following, the IP model is presented in detail. Notations used in this model are as follows:

- Parameters:

N Number of nodes.
L Number of links.
T Number of hierarchical levels minus one.
K Number of communities including dummy communities. Let $K = N \cdot (T + 1)$, *i.e.*, each level has N communities. The idea of such a setting is explained as follows. Each of all N nodes is a community at level 0, and all the N communities at level 0 are classified into the N communities at level 1 (in which some N communities doe not include a node, and are therefore dummy communities); this process then continues until all communities are allocated to a single community at level T.
d_n Number of degrees of node n.

- Indices:
  n, e Node index ($n, e = 1, 2, ..., N$).
  l Link index ($l = 1, 2, ..., L$).
  m, k Community index ($m, k = 1, 2, ..., K$).
  t Hierarchical level index ($t = 0, 1, ..., T$).

- Variables:
  $Y^t_{n,k}$ The variable of determining whether node n is assigned to the k-th community at level t, *i.e.*,
  $$Y^t_{n,k} = \begin{cases} 1, & \text{assigned;} \\ 0, & \text{otherwise.} \end{cases}$$
  $X^t_{l,k}$ The variable of determining whether link l is assigned to the k-th community at level t, red*i.e.*,
  $$X^t_{l,k} = \begin{cases} 1, & \text{assigned;} \\ 0, & \text{otherwise.} \end{cases}$$
  L^t_k The total number of links in the k-th community at level t.
  D^t_k The total sum of the degrees of the nodes in the k-th community at level t.

The objective function for detecting hierarchical community structures in this work differs from that in [28], which focused on detecting the community structures at only one level. To detect community structures with multiple levels, the objective function is obtained by tailoring the Q value in (1) as

follows:

$$Q^H = \sum_{t=1}^{T} \sum_{k=1}^{N} \left(\frac{L_k^t}{L} - (\frac{D_k^t}{2L})^2 \right)$$

Different from the Q value, the objective function Q^H additionally introduces the index t of hierarchical level, *i.e.*, the Q values of community structures for all levels are summed as the Q^H value to evaluate the quality of community structures for all levels. Consequently, the IP model is as follows:

$$\text{Maximize} \quad Q^H = \sum_{t=1}^{T} \sum_{k=1}^{N} \left(\frac{L_k^t}{L} - (\frac{D_k^t}{2L})^2 \right) \tag{2}$$

subject to

$$\sum_{k=1}^{K} Y_{n,k}^t = 1, \forall n, t \tag{3}$$

$$\sum_{n=1}^{N} Y_{n,k}^0 = 1, \forall k \tag{4}$$

$$2 \cdot X_{l,k}^t \leq Y_{n,k}^t + Y_{e,k}^t, \forall l = \{n, e\}, k, t \tag{5}$$

$$L_k^t = \sum_{l=1}^{L} X_{l,k}^t, \forall k, t \tag{6}$$

$$0 \leq L_k^t \leq \frac{L}{(T/t)}, \forall k, t \in \{1, 2, ..., T\} \tag{7}$$

$$Y_{n,k}^t + Y_{e,k}^t \leq Y_{n,m}^{t+1} + Y_{e,m}^{t+1}, \forall n \neq e, k, m, t \in \{0, 1..., T-1\} \tag{8}$$

$$D_k^t = \sum_{n=1}^{N} d_n \cdot Y_{n,k}^t, \forall k, t \tag{9}$$

$$\sum_{n=1}^{N} Y_{n,1}^T = N \tag{10}$$

This IP model is explained as follows. Objective (2) maximizes the objective function Q^H. As the value of Q^H increases, the quality of communities at all levels increases (*i.e.*, the nodes within each community are connected densely with each other, while links between two different communities are connected sparsely), and *vice versa.* Constraint (3) ensures that each node

at each level must be allocated to a community. Constraint (4) ensures that each node at level 0 forms a community with only one node.

Constraint (5) ensures that each link l with end nodes n and e at each level t is connected with nodes n and e, and if both the two nodes are allocated to the k-th community at level t (*i.e.*, $Y^t_{n,k} = Y^t_{e,k} = 1$), then link l is also allocated to this community (*i.e.*, $X^t_{l,k} = 1$). Consider the example in Figure 3, in which link l is connected with nodes e and n at level t; link l' is connected with nodes n and e'. For link l, since its two end nodes e and n are located in community k_1 (*i.e.*, $Y^t_{e,k_1} = Y^t_{n,k_1} = 1$), Constraint (5) (*i.e.*, $2 \cdot X^t_{l,k_1} \leq 1 + 1$) and the maximized objective function ensure that link l is allocated to community k_1 (*i.e.*, $X^t_{l,k} = 1$). For link l', since its two end nodes n and e' are located in two different communities k_1 and k_2, respectively (*i.e.*, $Y^t_{n,k_1} = Y^t_{e',k_2} = 1$ but $Y^t_{n,k_2} = 0$), Constraint (5) (*i.e.*, $2 \cdot X^t_{l',k_2} \leq Y^t_{n,k_2} + Y^t_{e',k_2} = 0 + 1$) ensures that link l' cannot be allocated to community k_2 (*i.e.*, $X^t_{l',k_2} = 0$).

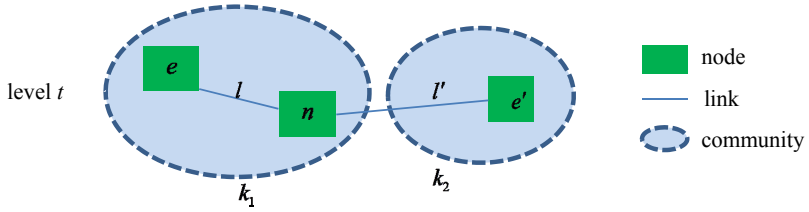


Figure 3: Illustration of Constraint (5).

Constraint (6) computes the total number of links contained in the k-th community at level t. Constraint (7) computes the community capacity limit at each level according to the number of levels; notably, capacity limit $L/(t/T)$ at level t is flexible as it is adjusted according to the number of levels t of concern and the number of levels minus one (*i.e.*, T). Constraint (8) ensures that if nodes n and e are assigned to the k-th community at level t, then they must be assigned to some community m at the adjacent higher level $t + 1$. Consider the example of constructing a hierarchical community structure in Figure 2(d). From the lowest level to the highest level of the hierarchy, if two nodes belong to the same community at a level, then they also belong to the same community at the adjacent higher level. For instance, nodes 1 and 5 are allocated to the 3rd community at level 1, they are allocated to the 2nd community at level 2 (*i.e.*, $Y^1_{1,3} + Y^1_{5,3} \leq Y^2_{1,2} + Y^2_{5,2}$), and they are

allocated to the 1st community at level 3 (*i.e.*, $Y^2_{1,2} + Y^2_{5,2} \leq Y^3_{1,1} + Y^3_{5,1}$).

Constraint (9) computes the total sum of the degrees of all nodes in the k-th community at level t; Constraint (10) ensures that all nodes at the highest level must be allocated to the community indexed by 1.

Generally, decision variables in an IP model are determined simultaneously; however, to elucidate the proposed IP approach, decision variables for detecting hierarchical community structures in this IP approach can be characterized as being determined from the bottom upward (*e.g.*, see the bottom-up construction in Figure 2(d) for detecting the hierarchical community structure for the network in Figure 2(a)). Initially, each node is understood as a community at level 0. The hierarchical community structure can be regarded as being constructed by allocating each community at each level to a community in its adjacent higher level. Notably, the community capacity limit for different hierarchical levels differs according to Constraint (7). A similar allocation process continues from the lowest level to the highest until all communities are allocated to a single community at the top level.

4. Experimental Design and Results

To assess the performance of the proposed IP approach, its performance on three real social network instances is assessed experimentally. The proposed IP model is solved using the IBM ILOG CPLEX Optimizer on the Microsoft Windows 7 Enterprise 64-bits platform using an Intel Core$^{\text{TM}}$2 Quad CPU Q9550 @ 2.83GHz with 8 GB memory. Drawings of the detected community structures are generated by yEd graph editor[1].

4.1. Network for Zachary's Karate Club

The first network instance, the well-known benchmark social network, represents the relationships among the members in a karate club, which was observed by Zachary from 1970–1972 [44]. The network graph has 34 members (nodes) and 78 links; each link represents a social relationship between the two end nodes of a link.

Figure 4(a) shows the dendrogram of the hierarchical community structure for this network instance detected by the proposed IP approach. The number of community levels is set at 4, and the Q^H value for the detected

[1]The yEd graph editor is obtained via the following URL:
http://www.yworks.com/en/products_yed_about.html

structure is 0.802. In the drawing of the detected structure (Figure 4(b)), nodes of the same color belong to the same community at level 1; each community at level 2 is encircled by red dashes; and all the nodes are grouped together at level 3 but are not marked.

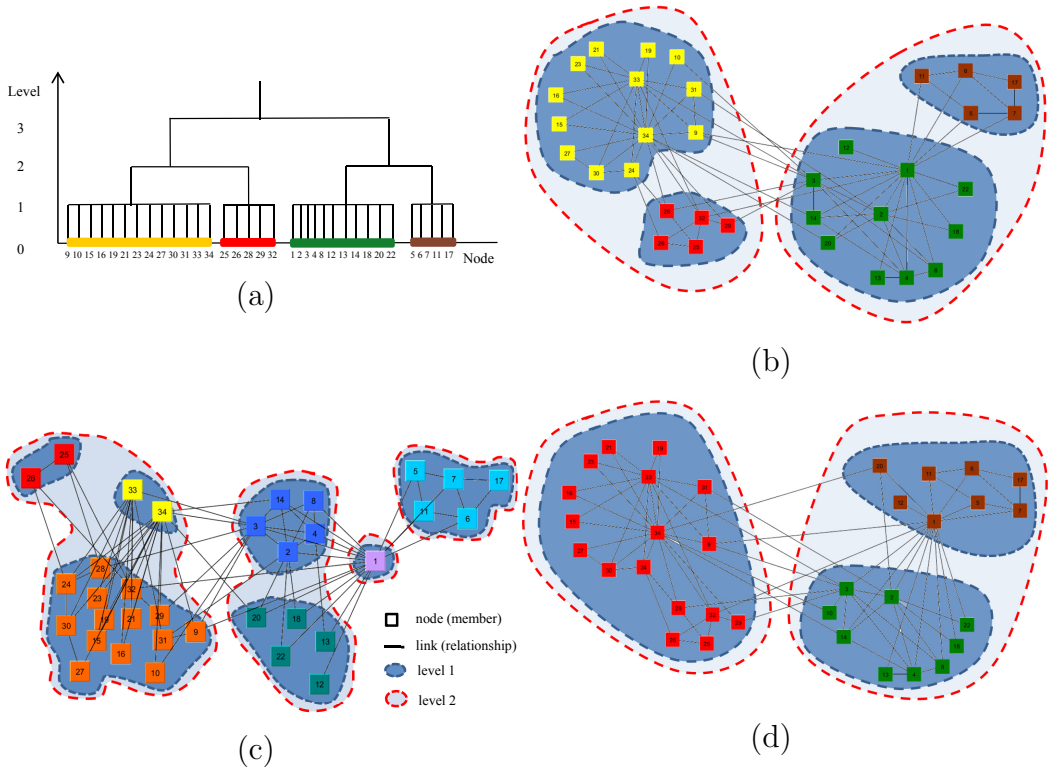


Figure 4: (a) Dendrogram of the detected hierarchical community structure for Zachary's karate club network. (b) Drawing of our detected structure. (c) Drawing of the structure detected by [17]. (d) Drawing of the structure detected by [24].

Nodes in the two super communities at level 2 are connected strongly with nodes 1 and 34, respectively (Figures 4(a) and 4(b)). In the real Zachary's karate club, node 1 is the coach of the club, and node 34 is the owner of the club. Hence, the detection result is the same as those in most previous works in that karate club members are divided into two groups that have a close relationship with the coach or a close relationship with the owner. Differing from detection results in previous works, the detection result in this work (Figure 4(b)) also finds that each community at level 2 (encircled

by red dashes) is divided into two subcommunities at level 1, where the key member belongs to the larger subcommunity.

The detection result for Zachary's karate club (Figure 4(b)) by the proposed approach is compared with the detection result in [17] (Figure 4(c)) and the detection result in [24] (Figure 4(d)). The Q^H values are as follows: this work, 0.802; [17], 0.411; and [24], 0.752 (Table 1). The total Q^H value (*i.e.*, 0.802) and the Q^H value at each level (*i.e.*, 0.417 at level 1 and 0.384 at level 2) in this work are largest. Hence, we conclude that the structure detected by the proposed approach preforms best in terms of the total Q^H value and the Q^H value at each level.

Table 1: Comparison of detection results for Zachary' karate club network.

Method	Level	Community number							Q^H at	Q^H
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	each level	
Our	$t=1$	0.066	0.147	0.049	0.155	–	–	–	0.417	0.802
	$t=2$	0.199	0.186	–	–	–	–	–	0.384	
	$t=3$	0.000	–	–	–	–	–	–	0.000	
[17]	$t=1$	0.011	-0.019	-0.022	0.068	-0.004	-0.008	0.066	0.093	0.411
	$t=2$	0.173	0.087	-0.008	0.066	–	–	–	0.318	
	$t=3$	0.000	–	–	–	–	–	–	0.000	
[24]	$t=1$	0.186	0.094	0.101	–	–	–	–	0.381	0.752
	$t=2$	0.186	0.186	–	–	–	–	–	0.372	
	$t=3$	0.000	–	–	–	–	–	–	0.000	

In a visual comparison (Figure 4(c)), node 1 (coach) is isolated from the other nodes (Q^H value of the community with node 1 at level 1 as well as level 2 is -0.08, less than that of any community in the detection result in this work and that in [24]), such that members who have close social relationships with the coach cannot be identified. The other key member, node 34 (club owner), has the same problem: node 34 is isolated from the other nodes in the leftmost community (community encircled with red dashes) (Figure 4(c)). Additionally, the scenario in which nodes 33 and 34 belong to the same community at level 1 (Figure 4(c)) (where the Q^H value of the community with nodes 33 and 34 at level 1 is -0.02) is strange because no link exists between the two nodes. Consequently, the detected hierarchical community structure for Zachary's karate club network by the proposed approach provides more reasonable information than that detected by the approach in [17].

To the best of our knowledge, [24] is one of the most recent attempts to detect the hierarchical community structure for Zachary's karate club network. A comparison of detection results (Figure 4(b) and Figure 4(d)) indicates that both meet the requirement for the real network: two communities are centered at nodes 1 and 34, respectively, at level 2 (*i.e.*, the two red-dashed communities in each figure). In addition, both detection results divide the community centered at node 1 (*i.e.*, the right community encircled with red dashes) into two subcommunities; however, these two subcommunities have different members (green nodes and brown nodes (Figures 4(b) and 4(d))). The total Q^H values of the two subcommunities at level 1 by the proposed approach and that in [24] are 2.13 (= 0.066 + 0.147) and 0.195 (= 0.101 + 0.094), respectively. Therefore, the detected hierarchical community structure for the two subcommunities performs better than that in [24] in terms of Q^H value. Moreover, although the method in [24] cannot detect the partition of the community centered at node 34 (*i.e.*, the left community encircled by red dashes in Figure 4(d)), the proposed approach finds the subcommunity of members close to node 34 (*i.e.*, yellow nodes in Figure 4(b)).

4.2. Network for College Football Games

The second network instance, that for the college football games in Division IA in the US in Fall 2000 [27], includes 115 teams from 12 conferences, in which each conference has 8–12 teams, and teams from the same conference have more chance to compete. In this network instance, each node represents a team, and a link exists between two nodes if there was ever a football game between the two teams represented by the two nodes.

Figures 5(a) and 5(b) show the hierarchical community structure detected by the proposed IP approach. The Q^H value for the detected structure is 1.176. This detection result is compared with that detected by [39] (Figure 5(c)), which is one of the most recent detection results for this network. Their Q^H values are also compared (Table 2). At level 1, both methods detected the same 12 communities at level 1 (Q^H at level 1 for both results is 0.598), which are similar to the division of the 12 conferences of the college football games but have more precise relationship on competition among teams, which has been shown in [39].

The primary difference between this detection result and the previous detection result in [39] is the community partition at level 2 (Figures 5(b) and 5(c)); in this study, five communities exist at level 2 (Q^H value at level 2 is 0.577); and in the previous study, only three communities exist (Q^H

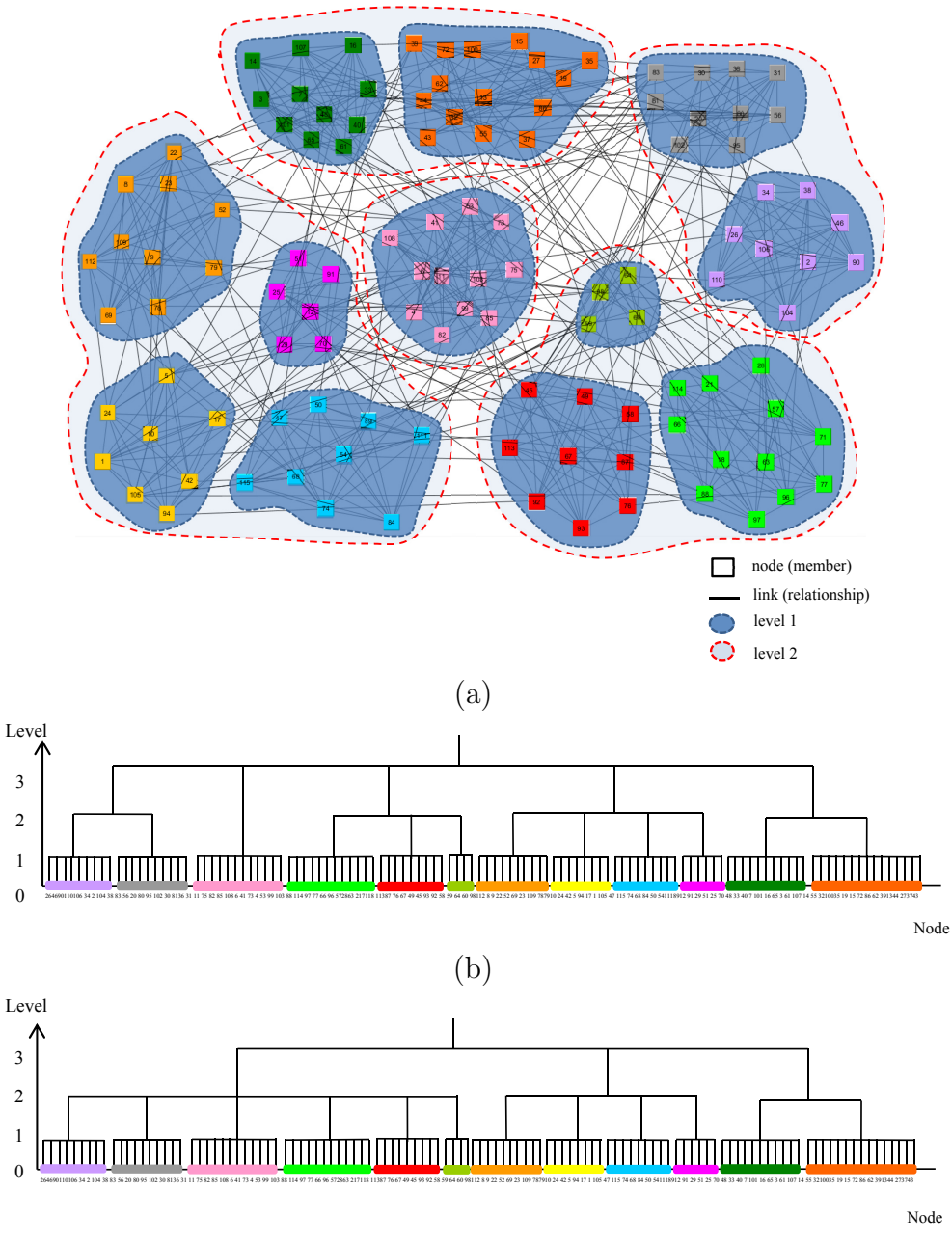


Figure 5: (a) Drawing of the detected hierarchical community structure for the network of college football games. (b) Dendrogram of the detected structure by the proposed approach. (c) Dendrogram of the structure detected in [39].

Table 2: Comparison of the results detected by the proposed approach and the method in [39] for the network of college football games.

Method	Level	Community number												Q^H at each level	Q^H
		$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$		
Our	$t=1$	0.052	0.009	0.067	0.043	0.078	0.062	0.067	0.022	0.055	0.041	0.051	0.052	0.598	1.176
	$t=2$	0.131	0.158	0.067	0.102	0.120	–	–	–	–	–	–	–	0.577	
	$t=3$	0.000	–	–	–	–	–	–	–	–	–	–	–	0.000	
[39]	$t=1$	0.052	0.009	0.067	0.043	0.078	0.062	0.067	0.022	0.055	0.041	0.051	0.052	0.598	1.074
	$t=2$	0.131	0.640	0.181	–	–	–	–	–	–	–	–	–	0.476	
	$t=3$	0.000	–	–	–	–	–	–	–	–	–	–	–	0.000	

value at level 2 is 0.476). That is, the first three communities at level 2 (Figure 5(b)) are merged into one (Figure 5(c)). Three communities at level 2 (*i.e.*, the central and the two rightmost communities encircled by red dashes) have fewer inter-community links between them than inner-community links (Figure 5(a)). Hence, the three communities are reasonable but are not detected by the previous approach. In addition, the proposed IP approach yields a higher Q^H value at level 2 (Table 2). Therefore, we conclude that the detection result obtained by the proposed approach is more precise than that by the approach in [39].

4.3. The Facebook Social Network

The third network instance uses the data collected from the Facebook website. Each user can communicate with friends on his/her friend list via some interactive behaviors, including sending messages and photos, posting on walls, and recording a "Like".

The Facebook network instance in this work addresses the friend lists of 50 graduate students who interact frequently. The students are studying for their M.S. degrees in Industrial Engineering (IE), Statistics, or Power Mechanical Engineering (PME) at National Chiao Tung University (NCTU), National Tsing Hua University (NTHU), or National University of Kaohsiung (NUK). Most of these students obtained their bachelor degrees in IE at NUK and are now studying for their M.S. degrees in IE at NCTU. Since IE programs at NCTU and NTHU are two of the best IE programs in Taiwan, and their campuses are adjacent geographically, students studying at the two universities often interact. In addition, since the IE and PME departments

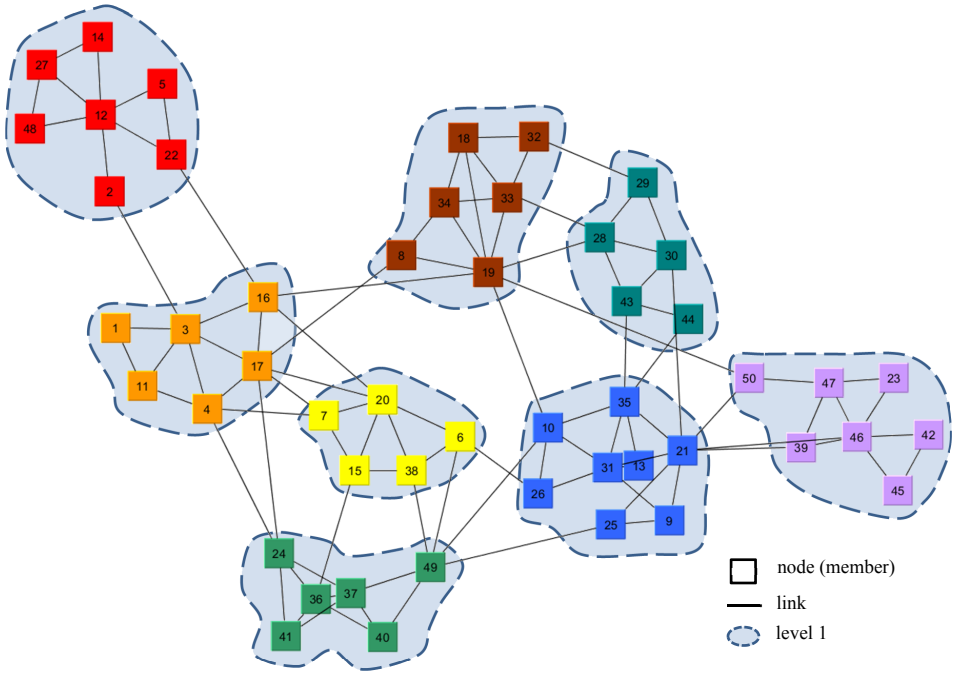
are located in the same building at NTHU, these students also have numerous interactions. Another portion of students studied for their bachelor degrees in Statistics at NUK and are studying for their M.S. degrees in Statistics at NCTU. These students interact frequently with those who have a bachelor degree in IE at NCTU because of their shared academic history.

In the Facebook network instance, each node represents a graduate student, and two students are linked when one of the two students has posted a message on the wall of the other Facebook friend. After selecting 50 nodes according to wall posting and friend lists, this network instance has 100 links. In the following, different numbers of hierarchical levels are used to analyze the detected hierarchical community structures.

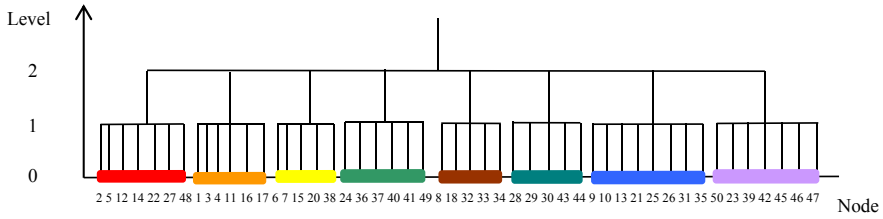
First, the number of hierarchical levels is set to 3. The detected community structure (Figure 6) has a Q^H value of 0.56865. The nodes at level 1 are divided into 8 communities, and the nodes in the same community are connected densely, while the links between communities are sparse (Figure 6). In the real social relationships on the network, nodes in red, orange, brown, and dark-green represent students majoring in IE; the nodes in yellow and light-green represent those majoring in Statistics; and the blue and purple nodes represent those majoring in PME.

The number of hierarchical levels is further set to 4. The detected community structure (Figure 7) has a Q^H value of 1.1485. Each of the 7 communities at level 1 is strongly connected. At level 2, the communities in orange, yellow, and brown form a super community; the communities in blue and purple form another; and the remaining two communities, that in red and that in green, are not merged with any other community. In real social relationships on this network, the four communities at level 2 (encircled by red dashes) correspond to the universities where these students received their bachelor degrees. Most nodes in the largest community at level 2 (*i.e.*, orange, yellow, and brown nodes) graduated from NUK; most red and green nodes graduated from NCTU; most of the nodes in in blue and purple colors graduated from NTHU.

Furthermore, the roles of some individuals in the network are discussed as follows. Node 19, a female graduate student studying for an M.S. degree in IE at NTHU, had been playing a leadership role since she was studying for her bachelor degree at NUK. Hence, the detected community structure shows that although she belongs to the brown community and has only one link to each of the yellow and orange communities based on the information at level 1, she has closer relations to the two communities at level 2. In

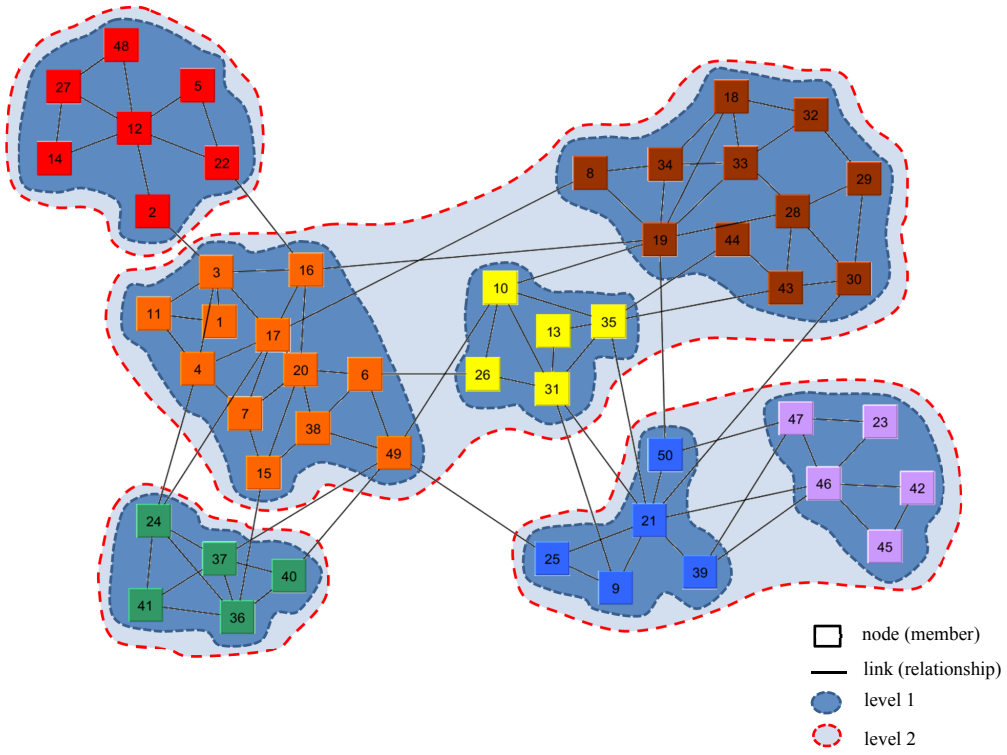


(a)

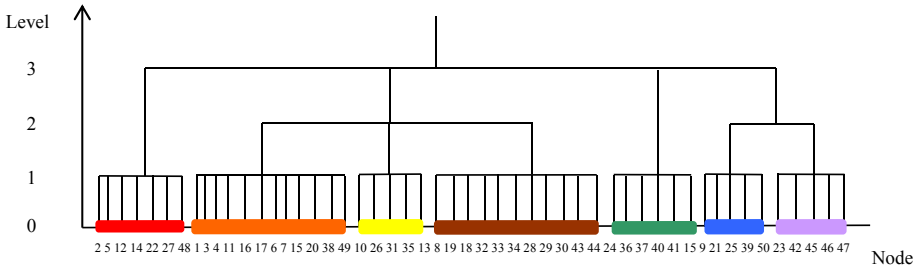


(b)

Figure 6: The hierarchical community structure of the Facebook network instance detected by the proposed IP approach with the setting of 3 hierarchical levels. (a) Drawing the detected structure, and (b) dendrogram of the detected structure.



(a)



(b)

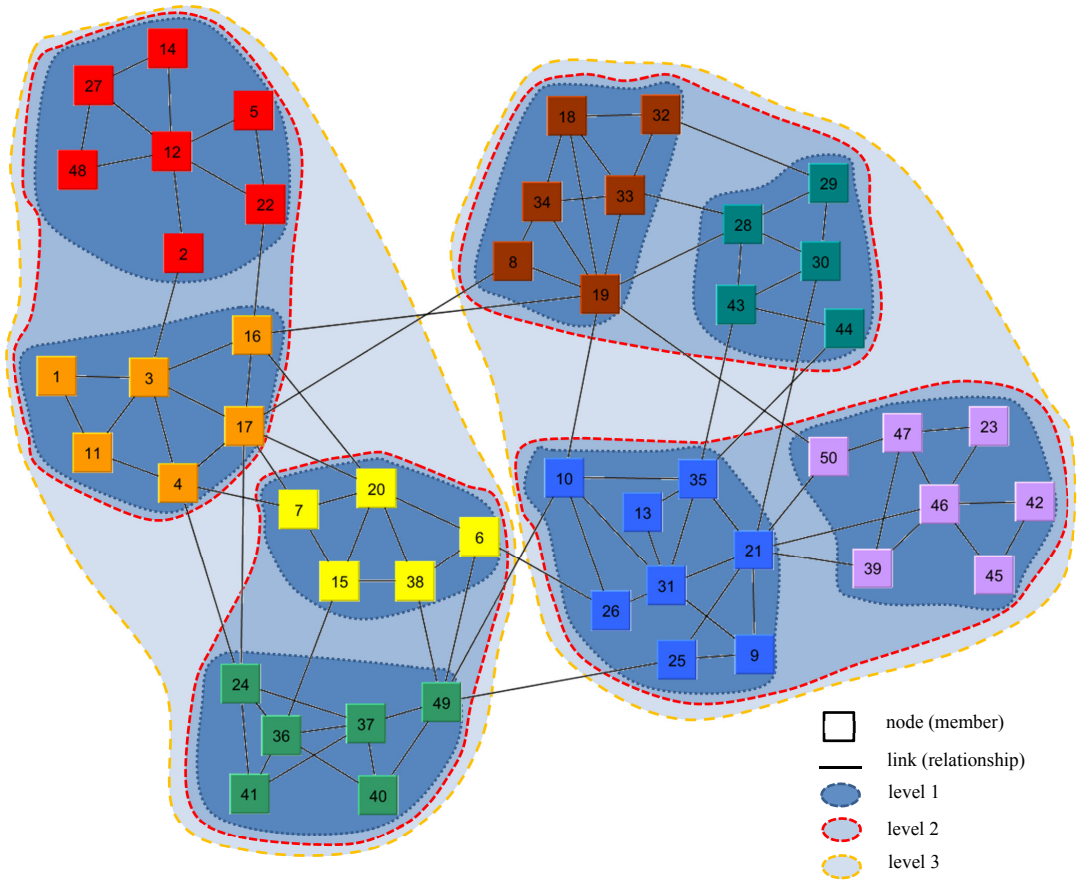
Figure 7: The hierarchical community structure of the Facebook network instance detected by the proposed IP approach with the setting of 4 hierarchical levels. (a) Drawing the detected structure, and (b) dendrogram of the detected structure.

addition, since the IE and PME departments at NTHU are located in the same building, she also interacts with a student in the PME department (*i.e.*, purple and blue nodes). In reality, node 21 is a male graduate student studying for his M.S. degree in PME at NTHU who enjoys interacting with others in his neighborhood. Node 21 has more connections with nodes at the same university (NTHU) and the same department (PME) than with nodes in other departments (Figure 7).

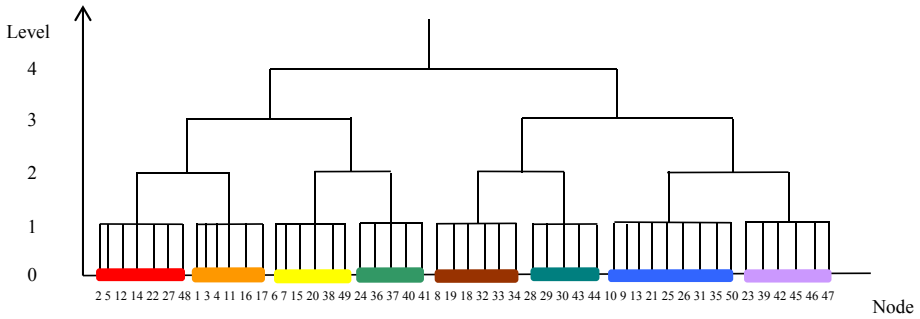
To gain additional insights into the network, the number of hierarchical levels is set to 5. The Q^H value is 1.6861 (Figure 8). The major difference from previous results is that all the nodes are divided into two super communities at level 3. In reality, the two super communities comprise NCTU graduates (left) and NTHU graduates (right). Furthermore, each of these two super communities at level 3 is divided into two communities at level 2 (*i.e.*, the four communities encircled by red dashes). In reality, the four communities at level 2 correspond to four departments, and are called department communities in the following. At level 1, the department communities at level 2 are further divided into multiple subcommunities. Note that the subcommunities formed of red and orange nodes belong to the same super community at level 2. In reality, the red and orange nodes are supervised by two different professors, but interact frequently because the two professors cooperate academically.

When the number of hierarchical levels is set to 6, the Q^H value is 2.2588 (Figure 1). The 8 communities at level 2 (Figure 1(b)) are the same as those at level 1 (Figure 8(b)). That is, different from the detection result in Figure 8, more small communities are found at level 1 (*e.g.*, pink community with nodes 16 and 17, the beige community with nodes 18 and 19, and the azure community with nodes 9, 21, and 25) (Figure 1). In reality, the nodes in the same community at level 1 are very close friends.

Since the detection result with 6 levels has more communities at level 1, close relations among individuals are found. Hence, it is of interest to set more levels to look into more insights. The detection result with the setting of 7 levels (Figures 9 and 10(a)) has a Q^H value of 2.8629. The communities at levels 2 and 3 (Figure 10(a)) are similar to those at level 2 (Figure 1(b)), but are merged differently (*e.g.*, the pink community (nodes 16 and 17) and the beige community (nodes 18 and 19) are merged with the others at level 2, but the azure community (nodes 9, 21, and 25) is merged with the others at level 3). By reality and a comparison of the dendrograms in Figures 1(b) and 10(a), the relationship between the pink and beige communities is closer



(a)



(b)

Figure 8: The hierarchical community structure of the Facebook network instance detected by the proposed IP approach with the setting of 5 hierarchical levels. (a) Drawing the detected structure, and (b) dendrogram of the detected structure.

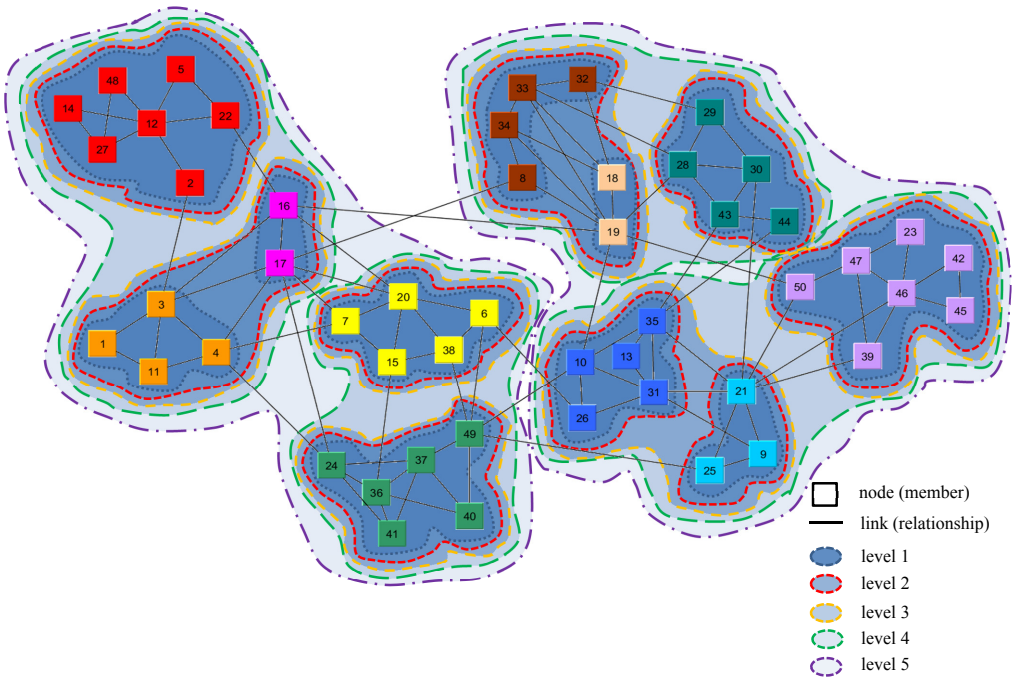


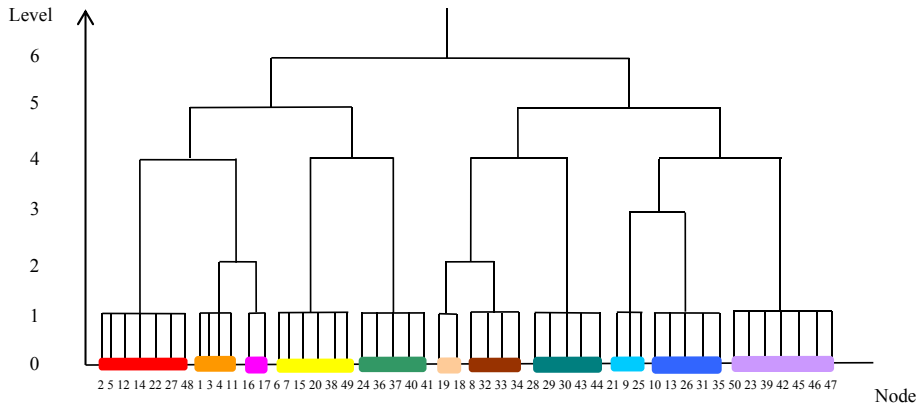
Figure 9: The hierarchical community structure for the Facebook network instance detected by our IP approach with the setting of 7 hierarchical levels.

than the azure community.

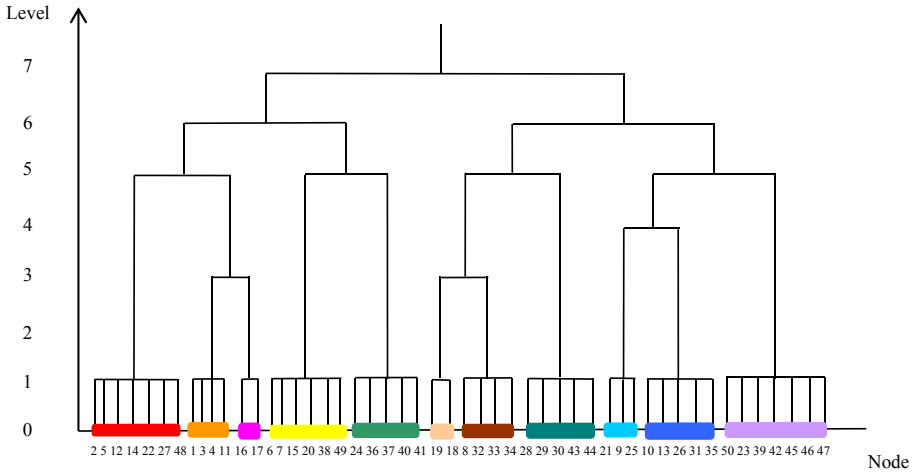
Since the detection results with 6 and 7 hierarchical levels are similar, 8 hierarchical levels are used (Figure 10(b)), yielding a Q^H value of 3.4656. The detection result in Figure 10(b) is roughly the same as that in Figure 10(a) except that the community originally merged at level 2 (Figure 10(a)) is merged at level 3 (Figure 10(b)). The detection result with 7 levels is best, as the other detection results do no differ significantly.

5. Conclusion and Future Work

As no previous works have proposed a mathematical programming approach that detects hierarchical community structures in social networks, this IP approach is novel. This approach allows a user to employ an existing software solver without implementing an algorithm. Different from metaheuristics that do not guarantee the optimal solution, the IP approach is simple and efficient for solving moderately-sized problems, and guarantees



(a)



(b)

Figure 10: The dendrograms of the hierarchical community structures of the Facebook network instance detected by the proposed approach with the settings of (a) 7 hierarchical levels and (b) 8 hierarchical levels.

solution optimality. In addition, the proposed approach uses flexible community capacity limits for different hierarchical levels according to problem scale. To assess the performance of the proposed approach, three network instances were experimentally tested. The detected hierarchical community structure for the first network instance, the karate club, by the IP method provides more information than that detected by previous methods. Experimental results for the second network instance, the network for college football games, reasonably displays competitive conditions for teams, and provides more information than that by previous methods. The last network instance is from the friend lists of real students on the Facebook website. Comprehensive analysis of this case successfully identified complex social behavior and personal backgrounds in this real social network.

In the future work, a mathematical programming method that can detect hierarchical community structures in dynamic social networks (*i.e.*, nodes/edges may be added or deleted at different times) will prove useful. Furthermore, a mathematical programming method that detects a community structure with hierarchical and overlapping relations simultaneously is also an interesting direction. Metaheuristic algorithms are always popular approaches for handling larger problem instances, even though solution optimality cannot be guaranteed.

Acknowledgements

The authors thank the anonymous referees for comments that improved the content as well as the presentation of this paper. This work has been supported in part by NSC 102-2219-E-009-013 and NSC 101-2628-E-009-025-MY3, Taiwan.

References

[1] S. Boettcher, A. G. Percus, Extremal optimization for graph partitioning, Physical Review E 64 (2001) Article ID: 026114.

[2] S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, Network analysis in the social sciences, Science 323 (5916) (2009) 892–895.

[3] Q. Cai, M. Gong, L. Ma, S. Ruan, F. Yuan, L. Jiao, Greedy discrete particle swarm optimization for large-scale social network clustering, Information Sciences, Available online 7 October 2014.

[4] A. Clauset, C. Moore, M. E. J. Newman, Structural inference of hierarchies in networks, Statistical Network Analysis: Models, Issues, and New Directions 4503 (2007) 1–13.

[5] A. Clauset, C. Moore, M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks, Nature 453 (7191) (2008) 98–101.

[6] L. Danon, J. Duch, A. Diaz-Guilera, A. Arenas, Comparing community structure identification, Journal of Statistical Mechanics: Theory and Experiment 29 (9) (2005) P09008.

[7] P. De Meo, E. Ferrara, G. Fiumara, A. Provetti, Enhancing community detection using a network weighting strategy, Information Sciences 222 (2013) 648–668.

[8] W. Didimo, F. Montecchiani, Fast layout computation of clustered networks: Algorithmic advances and experimental analysis, Information Sciences 260 (2014) 185–199.

[9] X. Duan, C. Wang, X. Liu, Y. Lin, Web community detection model using particle swarm optimization, in: Proceedings of IEEE Congress on Evolutionary Computation (CEC 2008), IEEE Press, 2008, pp. 1074–1079.

[10] V. Estivill-Castro, Why so many clustering algorithms: A position paper, ACM SIGKDD Explorations Newsletter 4 (1) (2002) 65–75.

[11] S. Eum, U. Osaka, S. Arakawa, M. Murata, A new approach for discovering and quantifying hierarchical structure of complex networks, in: Proceedings of 4th International Conference on Autonomic and Autonomous Systems (ICAS 2008), IEEE Press, 2008, pp. 182–187.

[12] D. Fisher, J. Artif, Iterative optimization and simplification of hierarchical clusterings, Journal of Artificial Intelligence Research 4 (1996) 147–180.

[13] Z. Gao, N. Jin, Detecting community structure in complex networks based on k-means clustering and data field theory, in: Proceedings of 20th Chinese Control and Decision Conference (CCDC 2008), IEEE Press, 2008, pp. 4411–4416.

[14] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Physical Review E 68 (6) (2003) 065103.

[15] A. I. Hafez, N. I. Ghali, A. E. Hassanien, A. A. Fahmy, Genetic algorithms for community detection in social networks, in: Proceedings of 12th International Conference on Intelligent Systems Design and Applications (ISDA 2012), IEEE Press, 2012, pp. 460–465.

[16] D. He, J. Liu, D. Liu, D. Jin, Z. Jia, Ant colony optimization for community detection in large-scale complex networks, in: Proceedings of 7th International Conference on Natural Computation (ICNC 2011), vol. 2, IEEE Press, 2011, pp. 1151–1155.

[17] T. Herlau, M. Mørup, M. N. Schmidt, L. K. Hansen, Detecting hierarchical structure in networks, in: Proceedings of 3rd International Workshop on Cognitive Information Processing (CIP 2012), IEEE Press, 2012, pp. 1–6.

[18] R. J. Kuo, Y. D. Huang, C. C. Lin, Y. H. Wu, F. E. Zulvia, Automatic kernel clustering with bee colony optimization algorithm, Information Sciences 283 (2014) 107–122.

[19] J. Leskovec, K. Lang, M. Mahoney, Empirical comparison of algorithms for network community detection, in: Proceedings of the 19th International Conference on World Wide Web (WWW 2010), ACM Press, 2010, pp. 631–640.

[20] T. C. Li, L. Zhao, Uncovering overlapping cluster structures via stochastic competitive learning, Information Sciences 247 (2013) 40–61.

[21] W. Li, Revealing network communities with a nonlinear programming method, Information Sciences 229 (2013) 18–28.

[22] J. Liu, T. Liu, Detecting community structure in complex networks using simulated annealing with k-means algorithms, Physical A: Statistical Mechanics and its Applications 389 (11) (2010) 2300–2309.

[23] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, 1967, pp. 281–297.

[24] C. Mu, Y. Liu, Y. Liu, J. Wu, L. Jiao, Two-stage algorithm using influence coefficient for detecting the hierarchical, non-overlapping and overlapping community structure, Physica A: Statistical Mechanics and its Applications 408 (2014) 47–61.

[25] M. C. V. Nascimento, Community detection in networks via a spectral heuristic based on the clustering coefficient, Discrete Applied Mathematics 176 (2014) 89–99.

[26] M. E. J. Newman, Mixing patterns in networks, Physical Review E 67 (2) (2003) 026126.

[27] M. E. J. Newman, M. Girvan, Community structure in social and biological networks, Proceedings of the National Academy Sciences of the United States of America 99 (12) (2002) 7821–7826.

[28] M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69 (2) (2004) Article ID: 026113.

[29] N. Nguyen, T. Dinh, D. Nguyen, M. Thai, Overlapping community structures and their detection on social networks, in: Proceedings of 3rd IEEE International Conference on Social Computing (SOCIALCOM 2011), IEEE Press, 2011, pp. 35–40.

[30] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (2005) 814–818.

[31] C. Pizzuti, GA-Net: A genetic algorithm for community detection in social networks, in: Proceedings of 10th International Conference on Parallel Problem Solving from Nature (PPSN 2008), vol. 5199 of Lecture Notes in Computer Science, 2008, pp. 1081–1090.

[32] P. Pons, M. Latapy, Post-processing hierarchical community structures: Quality improvements and multi-scale view, Theoretical Computer Science 412 (8) (2011) 892–900.

[33] K. Ragab, N. Kaji, K. Anwar, Y. Horikoshi, H. Kuriyama, K. Mori, A novel hierarchical community architecture with end-to-end delay awareness for communication delay enhancement, in: Proceedings of International Symposium on Applications and the Internet (SAINT 2004), IEEE Press, 2004, pp. 43–49.

[34] E. Ravasz, A. L. Barabási, Hierarchical organization in complex networks, Physical Review E 67 (2) (2003) 026112.

[35] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A. Barabasi1, Hierarchical organization of modularity in metabolic networks, Science 297 (5586) (2002) 1551–1555.

[36] C. Shi, J. Zhang, L. Shi, Y. Cai, B. Wu, A novel algorithm for hierarchical community structure detection in complex networks, in: Proceedings of the 6th international conference on Advanced Data Mining and Applications (ADMA 2010), vol. 6440 of Lecture Notes in Computer Science, 2010, Part I, pp. 557–564.

[37] M. Stabeler, C. Lee, G. Williamson, P. Cunningham, Using hierarchical community structure to improve community-based message routing, in: Proceedings of ICWSM Workshop on Social Mobile Web Workshop (SMW 2011), AAAI Press, 2011, pp. 40–44.

[38] P. G. Sun, L. Gao, Y. Yang, Maximizing modularity intensity for community partition and evolution, Information Sciences 236 (2013) 83–92.

[39] L. Šubelj, M. Bajec, Group detection in complex networks: An algorithm and comparison of the state of the art, Physica A: Statistical Mechanics and its Applications 397 (2014) 144–156.

[40] L. Wang, J. Wang, Y. Bi, W. Wu, W. Xu, B. Lian, Noise-tolerance community detection and evolution in dynamic social networks, Journal of Combinatorial Optimization 28 (3) (2014) 600–612.

[41] S. Wasserman, K. Faust, Social network analysis in the social and behavioral sciences, in: Social Network Analysis: Methods and Applications, 1994, pp. 3–27.

[42] D. J. Watts, P. S. Dodds, M. E. J. Newman, Identity and search in social networks, Science 296 (5571) (2002) 1302–1305.

[43] G. Xu, S. Tsoka, L. G. Papageorgiou, Finding community structures in complex networks using mixed integer optimization, European Physical Journal B 60 (2007) 231–239.

[44] W. W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33 (4) (1977) 452–473.

[45] S. Zhang, Hierarchical modular structure in gene coexpression networks, in: Proceedings of IEEE 6th International Conference on Systems Biology (ISB 2012), IEEE Press, 2012, pp. 118–124.

[46] Z. Y. Zhang, Community structure detection in complex networks with partial background information, EPL 101 (4), 48005 (6 pp.).

[47] M. Y. Zhou, Z. Zhuo, S. Cai, Z. Fu, Community structure revealed by phase locking, Chaos 24 (3), article No. 0300128 (7 pp.).

[48] T. Zhu, B. Wang, B. Wu, C. Zhu, Maximizing the spread of influence ranking in social networks, Information Sciences 278 (2014) 535–544.