Knowledge-Based Systems 87 (2015) 3–15

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Detecting hierarchical structure of community members in social networks

Fengjiao Chen, Kan Li *

Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China

ARTICLE INFO

Article history:
Received 26 October 2014
Received in revised form 19 May 2015
Accepted 28 May 2015
Available online 3 June 2015

Keywords:
Social network
Community detection
Hierarchical structure
Random walk
Linear regression

ABSTRACT

Current methods often predefine fixed roles of members and only detect fixed hierarchy structures that are not consistent with real-world communities; methods with hand-crafted thresholds bring difficulties in real applications, while choosing the community corresponding to the maximal belonging coefficient for each node results in a single boundary and neglects the multi-resolution of communities. In order to solve the limitations above, we propose a novel structure to dig finer information by partitioning the members into several levels according to their belonging coefficients. We call this novel structure Hierarchical Structure of Members (HSM) and discuss its properties in continuity, comparability, consistency and stability which reveal the multi-resolution of community as well as the intra-relations among members. We propose a two-phrase method, Random Walk and Linear Regression (RWLR), to detect HSM. The method measures the belonging coefficients of members by random walk and then divides the members into multiple segments by linear regression. Experiments show that members in the same level hold the same properties and HSM reveals multi-resolution of community. Besides, the comparison in benchmarks shows the efficiency in community detection. Finally, we apply HSM to analyze social networks, including visualization of community structures in large social networks and interactive recommendations in Amazon network.

1. Introduction

Many social networks display communities, groups of vertices with a higher-than-average density of edges connecting them. Community structure is fundamental for uncovering the links between structure and function in complex networks and for practical applications in many disciplines [1,17]. Members in a community play different roles, such as cores, bridges and fringes, which demonstrates the inner hierarchy of a community. Core members have large influence on other members, while bridge members help to keep communication with outside. Since the boundary of a community in social network is often fuzzy [8], a community can have different sizes from different views. Modeling a community by multiple boundaries is more suitable for real social networks, and we call such model as multi-resolution of a community.

Existing community detection methods often predefine fixed roles of members, resulting in a fixed hierarchical structure of community members [16,22,26]. However, communities in the real world usually have a variety of structures rather than a specific one. Take the relation of employees inside a company as an example, the hierarchy structure reflects the leadership of the company. Some companies have chairmen, department managers and their staffs. While other companies have chairmen, department managers, team leaders and team members. The hierarchy structures are different. In addition, current work of detecting multi-resolution of a community mainly follows two trends. One is to set hand-crafted thresholds to filter out the multi-resolution of community [18]. The choice of thresholds is not determined by algorithms, which brings difficulties in real applications. The other one is to choose the community corresponding to the maximal belonging coefficient for each node, which is easy to result in a single boundary. In this case, we need to propose a new method to detect the roles of community members as well as the multi-resolution of community.

To address these problems, we propose a new concept, called hierarchical structure of members (HSM), which describes a community by a 'level' structure according to the belonging coefficients. The belonging coefficients reflect the strength of relation between a member and its community. Members with similar belonging coefficients are in the same level. In the first level,

* Corresponding author.
E-mail addresses: cfjmonkey@hotmail.com (F. Chen), likan@bit.edu.cn (K. Li).

http://dx.doi.org/10.1016/j.knosys.2015.05.026
0950-7051/

members have the highest belonging coefficients, while those in the last level have the lowest belonging coefficients. The former ones are regarded as core members and the latter ones are treated as marginal members. From the first level to the last level, one can form the multi-resolution of community from the seed community to the whole network.

HSM is different from the hierarchical structure of communities (HSC) which gains attentions in recent years [4,10,11,14,30,31]. HSC describes nested community structures and shows the relations of communities, while HSM describes the levels of members and shows the relations of nodes.

An ideal HSM detection approach should maintain continuity, comparability, consistency and stability at the same time (see Section 3), which becomes the most challenging part of this problem. Besides, the automatic determination of the number of levels requires consideration as well. In this paper, we propose the RWLR (Random Walk and Linear Regression) method to solve these problems. It measures the belonging coefficients by random walk from a seed community and divides the members into levels by linear regression on the sorted sequence of belonging coefficients. Our method achieves a good performance on the benchmark datasets as well as on the real-world networks, which demonstrates the usefulness of HSM.

The rest of the paper is organized as follows. Section 2 reviews some related work on belonging coefficients measurement and hierarchical community detection. Section 3 defines the problem in a more formal way and then illustrates the RWLR method with a community detection framework. Section 4 shows the experimental results and the statistical analysis. Section 5 discusses the advantages and the limitation of RWLR method. Finally, we draw the conclusions in Section 6.

2. Related work

According to the belonging coefficients, nodes are assigned into different levels in HSM. Nodes play different roles in the community. There are some works that divide members into predefined roles. Nepusz et al. [22] detected the fuzzy communities and recognized three kinds of nodes ('outlier', 'bridge' and 'regular') according to the belonging coefficients. Huang et al. [11] expanded communities locally from all the nodes to get the overlapping hierarchical structure of communities (HSC) and then separated the homeless nodes as 'hub' or 'outlier'. Stanoev [26] used dynamic process to reveal the fuzzy communities and assigned nodes to one of the three roles, i.e. 'leader', 'follower' and 'proxy'. Leskovec [16] and Guimera [9] focused on two kinds of members, core members and peripheral members. However, roles in these methods are predefined, which limits the generalization ability of HSM.

The most relevant to ours is the work by Havemann [10] which aims to detect HSC. They calculated all the proper values of parameter alpha which represents stable community structures to improve the Fitness function proposed by Lancichinetti et al. [14]. Although the multi-resolution of communities can be produced, their method cannot guarantee the consistency of belonging coefficients in each level.

Measurement of belonging coefficients has been studied for years, and their corresponding methods are often called fuzzy or overlapping methods. Liu [18] extended the modularity to fuzzy modularity based on a random walk process. Psorakis et al. [23] utilized a Bayesian nonnegative matrix factorization (NMF) model to assign the participation scores of nodes. Nepusz et al. [22] considered the fuzzy community detection as a constrained optimization problem which minimized the difference between adjacency matrix. The similarity matrix is generated by belonging coefficients of nodes. Zhang et al. [34] mapped network nodes to Euclidean space based on a generalized modularity and applied fuzzy c-means to obtain a soft assignment. Steve et al. [7] extended the label and propagation dynamic process to fuzzy community detection. Similarly, based on information dynamic process, Xie et al. [29] defined the membership strength as the probability of observing a label in a node's memory.

Although there are many measurements, they are oriented to nodes, i.e. the sum of belonging coefficients of a node to all communities is equal to 1. Since the normalization is independently for each node, these measurements are not suitable for comparison among nodes. Unlike the methods above, another measurements are oriented to the community, where the sum of all nodes' belonging coefficients to a community is equal to 1 [25]. But, the belonging coefficients are dependent on the resolution of community, which is complex to get multi-resolution of community. Other generative models have the same constraints [20].

In addition, current work of detecting multi-resolution of a community mainly follows two trends. One is to set hand-crafted thresholds to filter out the multi-resolution of community, which brings difficulties in real applications. The other one is to choose the community corresponding to the maximal belonging coefficient for each node, which is easy to result in a single boundary. In this case, we need to propose a new method to detect the multi-resolution of community.

In this paper, we propose a method called RWLR to detect HSM. First, we measure the belonging coefficients oriented to communities by random walk and sort them in descending order. Second, we divide the order of belonging coefficients into multiple segments by linear regression and then get the HSM, where belonging coefficients are consistent in each level and the community structure is stable.

3. Method

3.1. Problem formulation

To simplify the problem, we mainly focus on the undirected, unweighted and simple network. The hierarchical structure of members for community C is defined as

$$Hier(C) = \{level_i\}, \quad i = 1, 2, \ldots, K, \tag{1}$$

where K is the number of levels. Each $level_i$ is a subset of members in community C and satisfies constraints below

$$\bigcup_i level_i = S_C, \tag{2}$$

$$level_i \cap level_j = \varnothing, \quad \forall i \neq j, \tag{3}$$

$$Max\{BC(\alpha), \alpha \in level_i\} < Min\{BC(\beta), \beta \in level_j\}, \quad \forall i > j, \tag{4}$$

where S_C is the set of members in the community C and $BC(\alpha)$ indicates the belonging coefficient of node α. $level_1$ has the highest belonging coefficient where the members are called core members; $level_K$ has the lowest belonging coefficient where the members are called marginal members.

Besides the definition of basic structure, HSM should hold several properties.

- Continuity: Nodes in any top levels will construct a connected component. Formally, for any $level_k$, each pair of node $\alpha, \beta \in \bigcup_{i=1}^{k} level_i$, exists a path $\alpha, p_1, \ldots, p_r, \beta$, where $p_i \in \bigcup_{i=1}^{k} level_i$.
- Comparability: The belonging coefficients of nodes to a community should be comparable. In other words, the belonging coefficients of a node to all communities should not be normalized independently.

- Consistency: The belonging coefficients of nodes in the same level are similar, while they have big differences in different levels. We evaluate the consistency in a level by mapping the sorted sequence of belonging coefficients to a line, because nodes in the same level have small differences relative to other levels, and they tend to be on a line. The formulation of consistency ε is defined in Eq. (5).

$$\varepsilon = 1 \Big/ \sum_{k=1}^{K} \sum_{i=n_{k-1}+1}^{n_k} (y(i) - L_i)^2, \tag{5}$$

$$y(i) = wi + b, \tag{6}$$

where K is the number of levels in HSM; n_k is the breakpoint of the kth level; $n_0 = 0$; and L_i is the ith largest belonging coefficient.

- Stability: Given a level k, top k levels can form the stable community. Stable community means that removing any node inside the community or adding any node outside will lower down the quality of community. The quality of community is evaluated by Eq. (7).

$$FL_k = Fitness(C_k) = \frac{d_{in}}{d_{in} + d_{out}}, \tag{7}$$

$$C_k = \bigcup_{k'=1}^{k} level_{k'}, \tag{8}$$

where C_k is the community composed of the top k levels, d_{in} and d_{out} are the internal and external degrees of the nodes in the community respectively. The Fitness function is proposed by [24] which is widely used.

In order to detect the HSM, we propose a method, called RWLR, as a part of community detection framework.

3.2. Community detection framework

The target for traditional community detection is to generate a partition of the graph, i.e. graph partition. As shown in Fig. 1, it is the framework that widely used in many community detection algorithms [3,6,14,15,21,27]. For a graph, the seeds of communities are found firstly. We choose the node with the lowest degree as the first seed. Other seeds are recursively chosen with the lowest degree outside any communities. Next, each seed is extended to construct a community. Finally, the communities are collected and form the graph partition. Some practical techniques can improve the performance, such as selecting little fraction of seeds inside the community and merging similar communities.

In this paper, we focus on the 'Measure BC' and 'Detect Levels'. We detect not only the community structure but also the hierarchical structure of members. Our method consists of two parts: measuring belonging coefficients (RW) and dividing community hierarchy (LR).

3.3. Measurement of belonging coefficient

Belonging coefficients describe how close the nodes connect to the community. Nodes with the largest belonging coefficients correspond to core members. We will find breakpoints to divide nodes into levels. Nodes whose belonging coefficients are larger than a breakpoint's will construct a resolution of the community. As a community, these nodes should form a connected component, i.e. *Continuity*.

Besides, we prefer the community-oriented measurement, which hold the *Comparability*, to node-oriented measurement. In

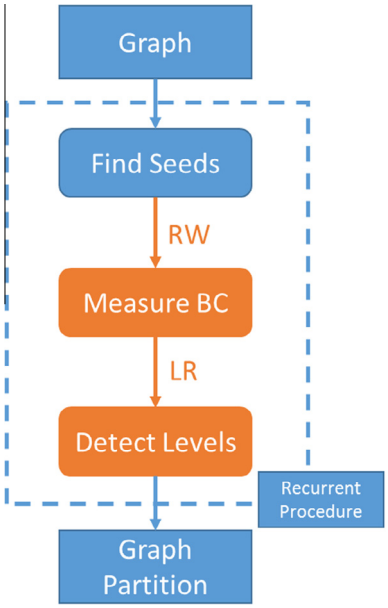


Fig. 1. Community detection framework.

node-oriented measurements, the belonging coefficient of a node to all communities are normalized to 1, which are not comparable (as discussed in Section 2) and restrain the belonging coefficients of overlapping nodes. For example, in Fig. 2a, the toy network has two communities separated by colors. Overlapping node 4 belongs to both two communities. Since node 5,6,7 only belong to one community, their belonging coefficients to the pink community are equal to 1. Since the belonging coefficients of node 4 to the blue community is larger than 0, the belonging coefficients to the pink community must be smaller than 1. However, connecting to all the nodes in the pink community, node 4 has much closer connection to this community and should have higher belonging coefficients to the pink community than node 5,6,7, which is not available for the measurements that oriented to nodes. That's why our measurement should be oriented to communities especially for the detection of overlapping community.

In order to satisfy the continuity and comparability, we design a community-oriented measurement based on random walk which is useful in detecting community [2,5,12]. At each step, a walker is on one node and moves to another chosen randomly and uniformly among its neighbors. The probability of arriving a node is decreasing while walking. This process guarantees that there's a path with decreasing probabilities for any node to the source node (see proof in Appendix). It means that nodes with probability above any positive threshold can construct a connected component.

Let us give a formal description. At each step, the transition probability from node i to node j is given by

$$P_{ij} = \frac{A_{ij}}{d_i}, \tag{9}$$

where $A_{ij} = 1$, if there's a link between i and j; otherwise, $A_{ij} = 0$. d_i is the degree of the node i. The corresponding matrix P is called the transition matrix. Given a source node s, the probability walking from node i to node s with t steps is defined as

$$q_s^t(i) = \sum_{j=1}^{N} q_s^{t-1}(j) Q_{ij}, \tag{10}$$

where Q equals to transition matrix except for $Q_{si} = 0, i = 1 \ldots N$. Then we define the measurement of belonging coefficients as the probability walking from node i to node s within T steps:

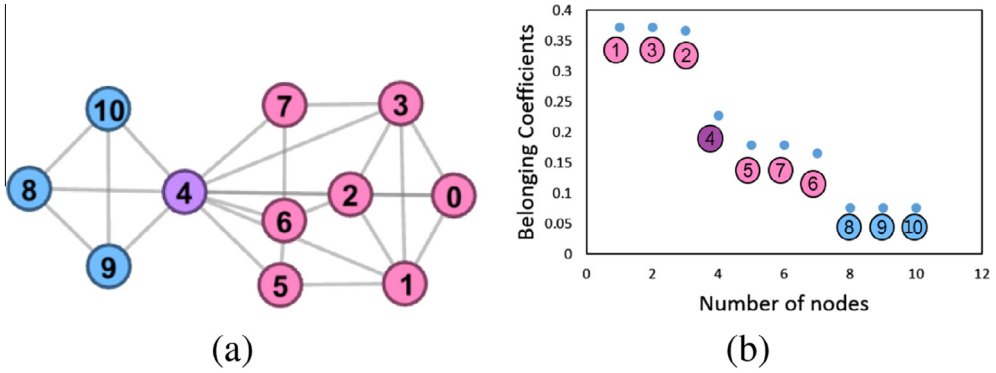


Fig. 2. An example network and its sorted order of belonging coefficients. (a) Shows the HSM. (b) Shows the belonging coefficients to the community with seed node 0.

$$BC_s^T(i) = \sum_{t=1}^{T} q_s^t(i), \tag{11}$$

which means how close the node i is connected to the community with source s. s is also called the seed of the community.

This measurement can satisfy the requirements above. Because more links are inside the community than outside, the probability of the random walker walking inside is higher than walking outside. Thus the gap in the sorted order of belonging coefficients can indicate the boundary of community structure. Since the nodes with more paths and less steps to the seed have higher belonging coefficients, nodes with belonging coefficients above any positive threshold can construct a connected component. For the toy network in Fig. 2a, belonging coefficients to the community with seed node 0 are shown in Fig. 2b. There are two gaps that indicate the core community (node 0,1,2,3) and the pink community respectively. Furthermore, there's no constrain about the sum of node's belonging coefficients and we can give fine belonging coefficients for overlapping nodes (node 4).

For each node i in Eq. (10), we just need to enumerate its neighbors to update $q_s^t(i)$. Thus the time to update all nodes with one step is $O(M)$ where M is the number of links in the network. The total time complexity of measurement is $O(TM)$ where T is the number of steps. In the next part, we will use the order of nodes sorted by their belonging coefficients to generate HSM. As analyzed in [12], the order will keep unchanged through a number of steps. We also discuss the selection of step T in the experiment.

3.4. Detection of levels

After having the belonging coefficients of each node, we divide the nodes into levels by their belonging coefficients to generate HSM. The division should keep the *Consistency* of members.

We sort the belonging coefficients in descending order and denote the result sequence as L. Observing the sequence of a toy network in Fig. 2a, nodes in the same level tend to form a line and the whole sequence can be regarded as a combination of multiple lines (Fig. 2b). Nodes in the same level have small differences relative to the nodes in other levels, so they tend to be on a line. Similar phenomenas can be observed in real social networks (Fig. 3a) too. That's why we use line segments to map the sequence and define the consistency as Eq. (5). The consistency can also catch the gap, because the breakpoint on the gap has higher consistency than the breakpoint beneath the gap.

For a given K, we should select the best $K-1$ breakpoints to minimize the inconsistency which is the reciprocal of consistency. To solve this optimization problem, we design an algorithm based on the dynamic programming. We define the state f_{ij} as the minimum inconsistency of the first j nodes within i levels. Assuming

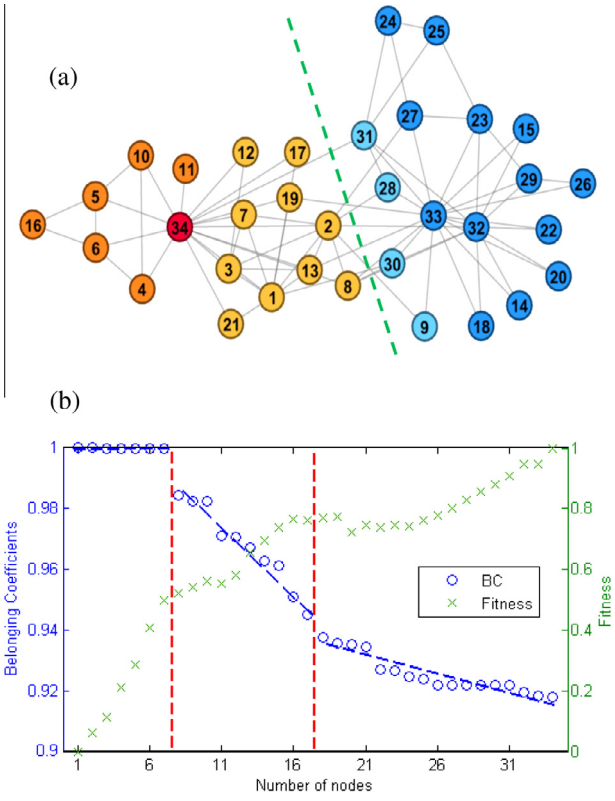


Fig. 3. The HSM in the karate club network. (a) Shows the structure of network. Nodes at different levels are drawn by different colors, i.e. orange, yellow, light blue and deep blue. The seed node 34 is colored in red. The green dash line in (a) indicates the top 2 levels. (b) Shows the values of BC and Fitness corresponding to the nodes ranked by BC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

that $f_{i-1,j'}(j' = 1,2,\ldots,j)$ have been calculated, we can enumerate the last breakpoints to calculate f_{ij}. Then the transition function is defined as

$$f_{ij} = min\{f_{i-1,k} + cost(k+1,j)\}, \quad k = i-1 \ldots j-1, \tag{12}$$

$$g_{ij} = argmin_k\{f_{i-1,k} + cost(k+1,j)\}, \quad k = i-1 \ldots j-1, \tag{13}$$

where g_{ij} records the breakpoint selected by f_{ij} and the $cost(p,q)$ is the minimum residual error to fit the part of order from the pth node to the qth node, which is defined as

$$cost(p,q) = \sum_{i=p}^{q} (y(i) - L_i)^2 = \sum_{i=p}^{q} (wi + b - L_i)^2. \tag{14}$$

The minimal *cost* is computed by the least square method and the parameter w and b are given by Eqs. (15) and (16) respectively.

$$w = \frac{n\sum_i iL(i) - \sum_i i\sum_i L(i)}{n\sum_i i^2 - \sum_i i\sum_i i}, \tag{15}$$

$$b = \frac{\sum_i i^2\sum_i L(i) - \sum_i i\sum_i iL(i)}{n\sum_i i^2 - \sum_i i\sum_i i}, \tag{16}$$

where $i = p, \ldots, q$ and $n = q - p + 1$. Calculating f iteratively, we can obtain the division for the HSM by the breakpoints stored in g.

We can further improve the complexity of the algorithm. After state f_{ij} is calculated, the state chooses $k = g_{ij}$ as the better decision than previous decisions p which satisfies $p < k$. It means that breaking the last level at k is better than at p, since $\{node_{k+1}, \ldots, node_j\}$ and $\{node_{p+1}, \ldots, node_k\}$ are very inconsistent. For latter states $f_{i,j'>j}$, if $\{node_{j+1}, \ldots, node_{j'}\}$ and $\{node_{p+1}, \ldots, node_j\}$ are inconsistent, $f_{ij'}$ prefers to break at k than p; otherwise, $f_{ij'}$ prefers to break at p. However, it's hard to satisfy the latter condition because we have the conclusion that $\{node_{k+1}, \ldots, node_j\}$ and $\{node_{p+1}, \ldots, node_k\}$ are very inconsistent. Hence, for later states $f_{i,j'>j}$, it's more probable to prefer decision k to p, which means the decisions are non-descending. Observing the decisions based on real networks (Table 2), we count the number of decisions g_{ij} satisfying $g_{ij} >= g_{i,j-1}$ and show the result in Table 1. There are over 99% decisions which are non-descending with j. Then we assume that if $f_{i-1,k} + cost(k+1,j) <= f_{i-1,p} + cost(p+1,j)$, where $p < k, f_{i-1,k} + cost(k+1,j+1) <= f_{i-1,p} + cost(p+1,j+1)$ is as well for $f_{i,j+1}$. With this assumption, we can reduce the time complexity of division from $O(KN^2)$ to $O(KN)$ where N is the number of nodes in the network. Although there are near 1% conditions that cannot satisfy this assumption, RWLR with this improvement still has competitive performance in the experiments.

Proposition 1 (*Time complexity*). *When calculating f_{ij}, if $f_{i-1,k} + cost(k+1,j) <= f_{i-1,p} + cost(p+1,j)$, where $p < k$, we will remove p from the decisions of $f_{i,j'>j}$, and the calculation of $f_{K,N}$ can be reduced to $O(KN)$.*

Proof. Calculating f_{ij} from $j = 1$ to $j = N$, we can construct a queue to hold the decisions. For f_{ij}, we add decision $k = j - 1$ into the queue and remove the decisions p that $\exists k, p < k$ and $f_{i-1,k} + cost <= f_{i-1,p} + cost$. Then the head of the queue is the best decision for f_{ij}. Each decision k is added in the queue and removed from the queue only once. The time complexity for $f_{i,\cdot}$ is $O(N)$ and the total time complexity is $O(KN)$. □

As for the number of levels K, we select it according to the *Stability* of the community structure. The community with local maximal quality is regarded as stable one. With increasing K, there are more line segments that can better map the order, then the inconsistency keeps decreasing. So we want to use as less lines as possible to obtain good community structure. Starting from $K = 1$, we test whether $K + 1$ can obtain better community structure. If the best community quality among all the levels cannot be increased with $K + 1$ levels, we regard this local best community structure as the stable one and take K as a result. Otherwise, we keep increasing K by one and repeat the test. To measure the quality of the community, we choose the Fitness function proposed by [24] which provided good results in a wide range of synthetic and empirical networks [15] (Eq. (7)).

In the experiments, the selected K is often small which ranges in $[2, 8]$. It deserves noticing that the selection does not influence the time complexity which is still $O(KN)$. The pseudo-code of the RWLR method is shown in Algorithm 1.

Table 1
The non-descending decisions of dynamic programming.

Networks	Non-descending (%)
Karate	0.968944
Dolphins	0.972244
University	0.978979
Amazon	0.999925
DBLP	0.9998
Youtbe	0.99965
Average	0.997158

Algorithm 1. RWLR.

Input: $G(V, E)$, seed S, step $T = 50$
Output: $level_i$, number of levels K
$n = |V|$
for $i = 1$ to n **do**
  calculate $B_S^T(i)$
end for
Set initial number of levels $K = 1$
Set previous best quality of community $preQua = 0$
Set current best quality of community $curQua = 0$
for $i = 1$ to n **do**
  $f_{1,i} = cost(1, i)$
  $g_{1,i} = -1$
end for
repeat
  $K = K + 1$
  for $i = 1$ to n **do**
    calculate $f_{K,i}$ and $g_{K,i}$
  end for
  $k = K$
  $i = n$
  repeat
    clear $level_k$
    for $j = i$ to $g_{k,i} - 1$ **do**
      add node j to $level_k$
    end for
    $i = g_{k,i}$
    $k = k - 1$
    until $k > 0$
    $preQua = curQua$
    $curQua = max(FL_i, i = 1 \ldots K)$
  until $curQua > preQua$
  $K = K - 1$
  $k = K$
  $i = n$
  repeat
    clear $level_k$
for $j = i$ to $g_{k,i} - 1$ **do**
      add node j to $level_k$
    end for
    $i = g_{k,i}$
    $k = k - 1$
  until $k > 0$

4. Experiments

There are three parts with five experiments in this section. First, we give qualitative analysis of HSM in small networks. Second, we

test the efficiency of RWLR method qualitatively. The accuracy is evaluated by detecting community structures on LFR benchmarks. The speed is tested to show the scalability. Besides, we analyze the selection of the only parameter, step T, in our method. Finally, we apply HSM to do social network analysis, including visualization of community structures in large social networks and interactive recommendations in the Amazon network.

4.1. Datasets

We test the method in both real networks and synthetic networks. The real networks are composed of three small networks and three large networks whose profiles are shown in Table 2. The synthetic networks are of two kinds, networks with overlapping communities and networks with hierarchical structure of communities (HSC) [13].

4.2. Qualitative analysis

To show the properties of HSM, we conduct experiments in three real networks (i.e. karate, dolphins and university networks), and a synthetic network, with ground truths.

4.2.1. Karate club network

The first one is the classical karate club network consisting of friendships between 34 members of a karate club [32]. This network is of particular interest because the club is split in two parts during the course of observations as a result of an internal dispute between the president and the instructor. We detect the HSM from the view of instructor (node 34) in Fig. 3.

The selected number of levels K is 2 (i.e. green dash line in Fig. 3a) and the division correctly maps the real partition except for node 8 which is at the end of the first level. Increasing K to three levels, the method splits the first level into two levels (orange and yellow). The orange level indicates a lower resolution which is the same as the result in [5]. Different from [10], which recognized the orange nodes as the last level, we can correctly recognize the multi-resolution of community. With K increased, the blue level is split into two levels (light blue, deep blue), and the light blue nodes are overlapping nodes.

We also show the order of nodes by their belonging coefficients (Eq. (11)) and the corresponding Fitness (Eq. (7)) in Fig. 3b, where levels are drawn by red dash lines and the best fit line segments are drawn as blue dash lines on data points. In the figure, the line segments map the consistency of belonging coefficients. The gap between the 7th node and the 8th node, and the bend between the 17th node and 18th node are also caught by RWLR. Comparing the belonging coefficient with Fitness, we can see that the local maximal Fitness points correspond to big gaps or bends in the line of belonging coefficients, which shows that the detected levels can form meaningful community structures.

If we just maximize the Fitness functions, the maximal result will be the whole network which is meaningless. Local expansion methods [15,14] try to detect the local maximal result (the stable community) which corresponds to the first 10 nodes which are not consistent. Combining the stability of community structure with the consistency of belonging coefficients can obtain better result.

4.2.2. Dolphins network

The second network is the network of bottle nose dolphins living in Doubtful Sound (New Zealand) analyzed by Lusseau [19]. Nodes represent dolphins and edges are set between animals that are seen together more often than expected by chance. The dolphins fall into two groups after a dolphin leaves the place for some time. We detect the HSM with the seed node 9.

In the Fig. 4a, the real parts are separated by the dash line, and the two levels (orange, blue) we detected correctly match it, except for the node 39 which is often regard as the overlap node. In order to show the influence of different seed nodes, we depict the sequences of belonging coefficients with each orange node as seed in Fig. 4b. In the figure, values in each sequence are distributed differently. It's not suitable to set a common threshold to filter out the communities, which is often used by fuzzy detection methods [7,25]. Although these sequences are different, the general shapes are similar, and so are the divisions. There are several dash lines to indicate the levels in each HSM. Most HSMs contain the breakpoint corresponding to Fig. 4a and we highlight it with a thick dash line. Consequently, from different seed nodes in the same community, we construct different sequences of belonging coefficients but the generated HSMs are still similar.

Table 2
Profiles of real networks.

Networks	Nodes	Edges	Diameter
Karate [32]	34	78	5
Dolphins [19]	62	159	4
University [22]	81	577	4
Amazon [33]	334,863	925,872	44
DBLP [33]	317,080	1,049,866	21
Youtbe [33]	1,134,890	2,987,624	20

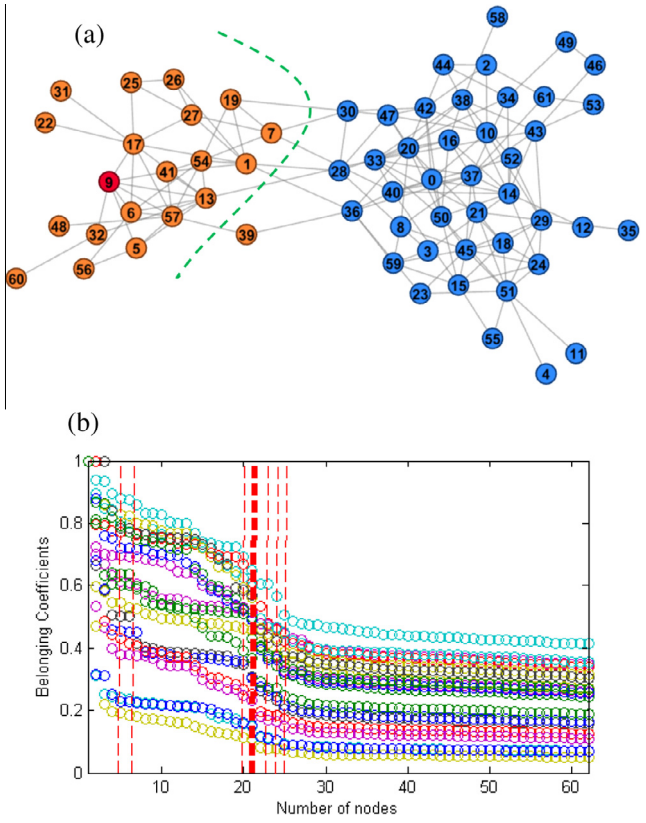


Fig. 4. The HSM in the dolphins network. (a) Shows the HSM and the green dash line indicates the top 1 level. Regarding each orange node as seed, (b) shows all the orders of BC and the red dash lines indicate divisions for each HSM whose numbers of levels are mostly 2 or 3. Most HSMs have the same division drawn by thick dash line which is also the real boundary of community. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2.3. University network

The third network is the university network of the academic staff of a given Faculty of a UK university consisting of three separate schools [22]. In the network, nodes represent person and links represent friendship. The friendship is measured with questionnaires, and all academic staff participate in the survey.

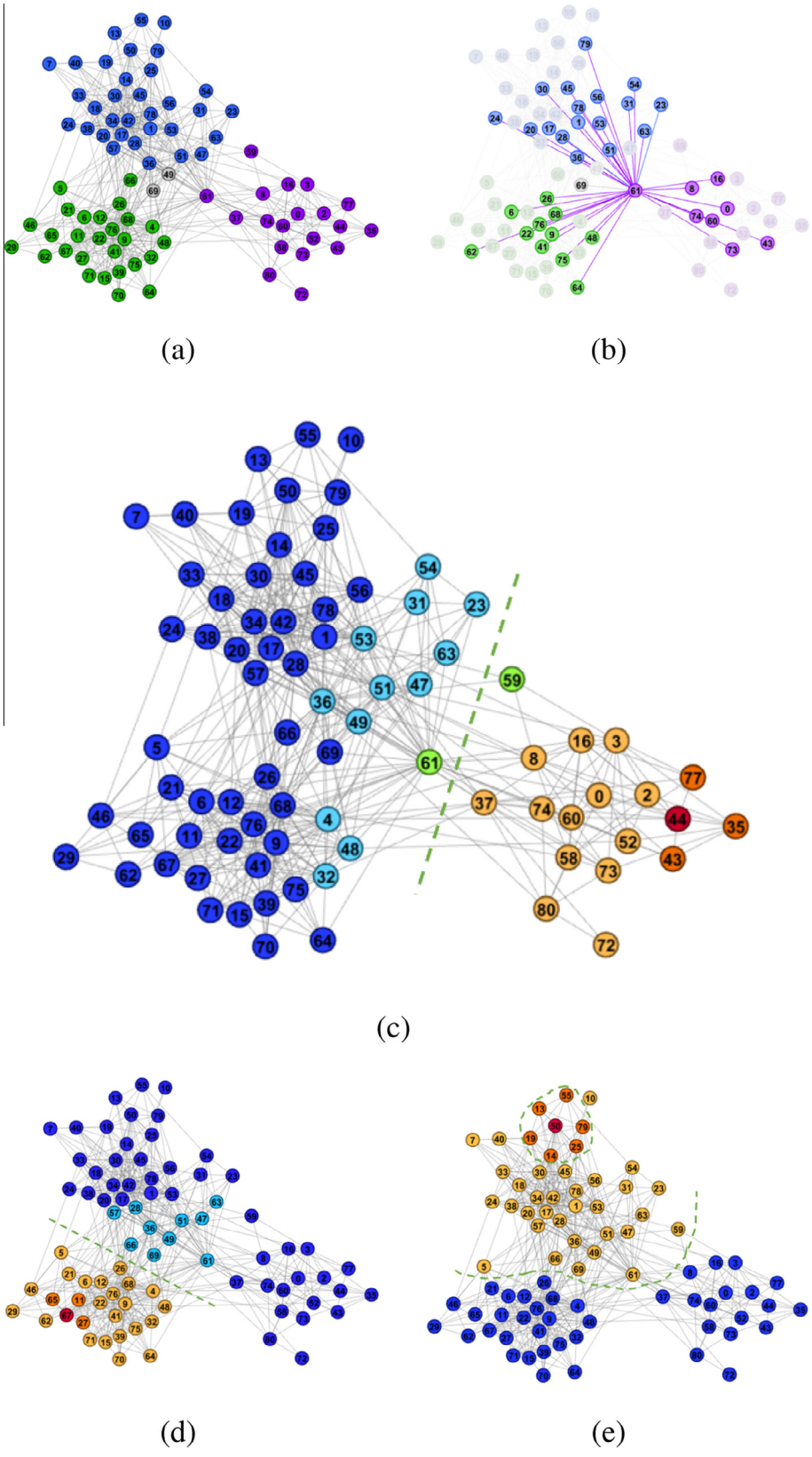


Fig. 5. The HSM in the university network. (a) Shows three real communities with two undetermined nodes (gray). (b) Shows the neighborhood of overlapping node 61. Three HSMs are shown in (c–e) separately, where levels are marked by different colors and the dash lines indicate the divisions automatically determined by RWLR. (c) Is zoomed for further analysis. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The real partition of the network is shown in Fig. 5a where communities indicate schools. There are two gray nodes that have not been determined. We detect HSMs with the seed node 50, 67, 44 respectively which correspond to three schools. In Fig. 5c,d,e the dash lines correctly separate the real communities from others except for few overlapping nodes.

In order to analyze the overlapping nodes, we increase the number levels for each HSM and obtain finer resolutions of communities. In Fig. 5c, the first level (orange) has the core members which densely connect to seed, and the second level (yellow) is almost the real community. Node 61 and node 59 are at the third level (green). Node 59 has two links connecting to two communities respectively, and it can belong to both communities. Node 61 is more interesting, because it has lots of links connecting to three communities in Fig. 5b, and three communities we detected all include node 61 at larger resolutions. In [22], node 61 is also called bridge node which plays an important role in the network. In the Fig. 5c, we can regard these special nodes as a single level in HSM where levels can indicate their roles in the network. Nodes at the fourth level are also fuzzy but not strong than the third level. Similar phenomena is observed in Fig. 5d.

4.2.4. Hierarchical synthetic network

In order to show the hierarchy in HSM, we generate a hierarchical synthetic network in Fig. 6a. There are 128 nodes and 1228 links in the network where more links are inside the communities than outside. 8 small communities have 16 nodes respectively and 2 large communities which consist of 4 small communities have 64 nodes respectively. All the communities are circled by dash lines in the figure. We detect HSM with the seed node 1 and the result has four levels separated by colors.

In the figure, we can correctly reveal the small community which is the first level in HSM. The large community can also be revealed successfully by the third level except for few nodes with many connections to this community. Except for these two levels, we can also reveal a level with overlapping nodes (the second level), which are very important for information spreading in social networks [28].

In the Fig. 6b, we show that the divisions of HSM can catch the local maximal Fitness. There is a local maximal Fitness at the 34th node, and it is the combination of two small communities A and B in Fig. 6c. However, we did not set this property when generating the network. Observing the neighbors of community A in Fig. 6c, A has dense connections with B, C and D but a little more with B. So A and B cannot combine to an independent community to C or D. Thanks to the HSM, we prevent from the unsuitable local maximization and obtain multi-resolution of the community.

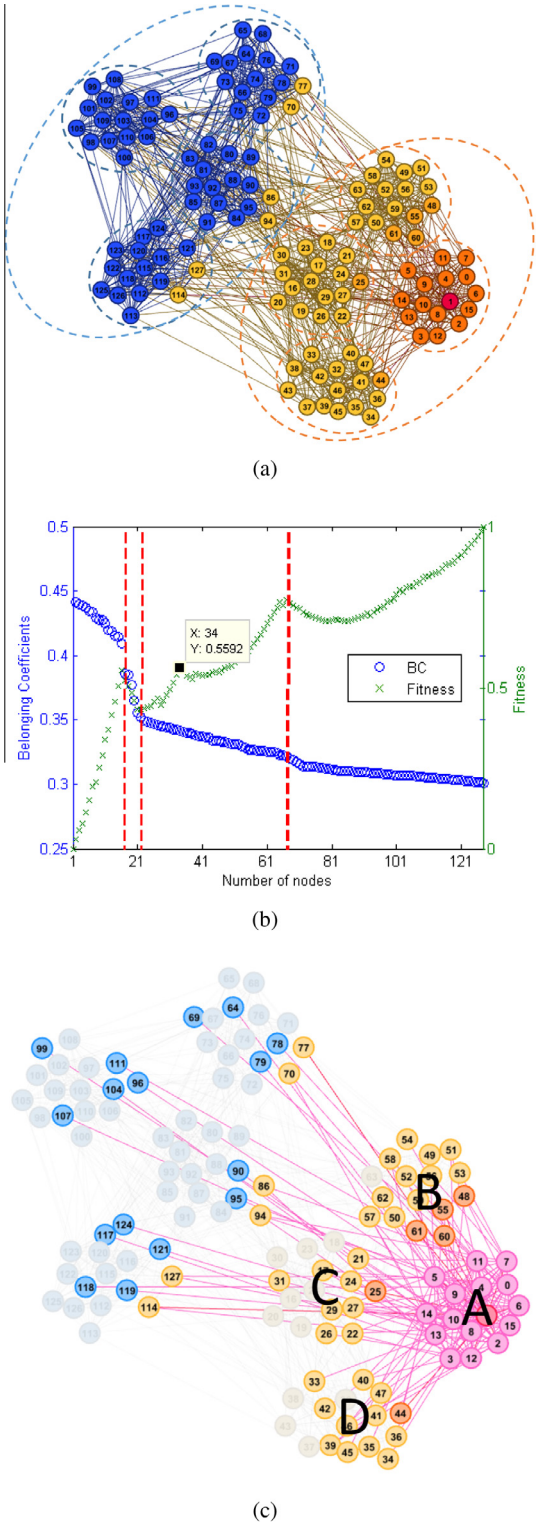


Fig. 6. The HSM in hierarchical synthetic network. The real partitions are circled in (a). There are two large communities and further four small communities in each large community. The seed node is colored in red. The corresponding order of belonging coefficients and Fitness are shown in (b). In order to analyze the relations among small communities, we highlight the community A and its neighborhood in (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. Quantitative analysis

4.3.1. Accuracy analysis

The comparison of community detection is conducted with three competitive methods, UEOC [12], LFM [14] and Copra [7]. LFM constructed the community from a seed independently by a greedy algorithm that maximized the local Fitness which is similar to our method, excepting that we detect the HSM at the same time. The Fitness function is widely used to evaluate the quality of communities [10,15], which is also the quality function used in our method. UEOC combined random walk and annealed network to unfold communities effectively. With the same measurement, random walk, we show similar properties in community detection. Copra is an overlapping community detection method based on dynamic process and has good performance in the recent study [28], which is the baseline in the comparison to show the performance among current methods.

These methods are tested in $4 * 1100$ benchmark networks. The benchmark network is generated by LFR [13] and the accuracy is measured by the extended NMI [14]. The LFR parameters are

similar to [8]: the network size n is 1000 or 5000; the minimum community size $cmin$ is 10 or 20; the maximal community size $cmax$ is $5 * cmin$; the average degree is 16 and the maximal degree is 40; the exponents parameters $\tau 1$ is -2 and $\tau 2$ is -1; the proportion of links between communities is 0.1 and the maximal number of communities that each node can belong to is 2. The fraction of overlapping nodes (on/n) varies from 0 to 0.5 with interval 0.05. For each set of parameters, we generate 100 networks and use the average result that is detected in the 100 networks to eliminate the random factor. The parameters in the compared method are set default values. The Fitness alpha in LFM is 1 which is the same as RWLR. To obtain good performance of Copra, we set the maximal number of communities that a node can belong to as the real value 2 and repeat the method 10 times to select the best result.

Although, there are several methods can detect the HSC or generate the benchmarks of HSC [13], the HSM has finer resolutions than HSC, so they cannot directly be compared. Hence, in this section, we mainly test the quality of a single level in HSM and left the test with hierarchy to future work. Similar with [12], we select the level with the best Fitness in HSM.

The comparison results are shown in Fig. 7. In the figure, the performance of RWLR is not influenced by the size of network, while UECO has lower performance in larger network. Copra has the similar performances no matter how the network size or community size changes. In all these conditions, RWLR performs better than UECO and Copra. When communities are large, RWLR has the best performance than all the other methods. UEOC also has better performance for large communities. It's the advantage of random walk.

Local expansion method LFM is good at detecting small communities. However, in real social networks, the size of community ranges largely and large communities are more important. So RWLR has better performance than LFM in real social networks. Although RWLR has lower NMI than LFM in small communities, its NMI can keep upon 0.8 which is much larger than other methods.

With the increase of the overlapping nodes, the community structures become mixed and the NMI should go down. The downtrend of RWLR is similar to UECO in large communities and slower than Copra and UECO under other conditions. It's interesting that LFM has lower NMI for small overlaps compared with the larger one. When there's a community structure with small overlaps, RWLR can accurately detect it; when there's one with large overlaps, RWLR also can yield an accuracy over 0.8 on small communities and over 0.9 on large communities. In conclusion, RWLR can detect community structures accurately, especially in the networks with large range of community sizes and small overlaps.

4.3.2. Scalability analysis

Since our method detects not only the community structure but also the hierarchy, and the time complexity depends on the selection of seeds which is not part of our method, it's unfair to compare time complexity with normal community detection method. So in this section, we analyze the relation between the theoretical time complexity and the practical time complexity.

The experiment is conducted on a PC with Intel(R) Core(TM)2 Quad CPU and 4 GB RAM. We generate networks with nodes range from 100 to 100,000 by the same method in Section 4.3.1. For each number of nodes, we generate 100 networks and run the RWLR method 100 times in each network. Finally, the average time is used to eliminate noises. The result is shown in Fig. 8. Since the range of X axis is large, we use logarithmic axis to show all the numbers. The time cost of RWLR is linear to the number of nodes, which indicates that the time complexity is $O(N)$, where N is the number of nodes in the network. Although the theoretical time complexity is $O(KN)$, the number of levels K is small (ranges in $[2, 8]$) that does not influence much on the time.

4.3.3. Parameter analysis

Furthermore, we analyze the only parameter in our method, the number of step T in Eq. (11). After each step, the belonging coefficients of nodes will be updated and the order of the sorted nodes

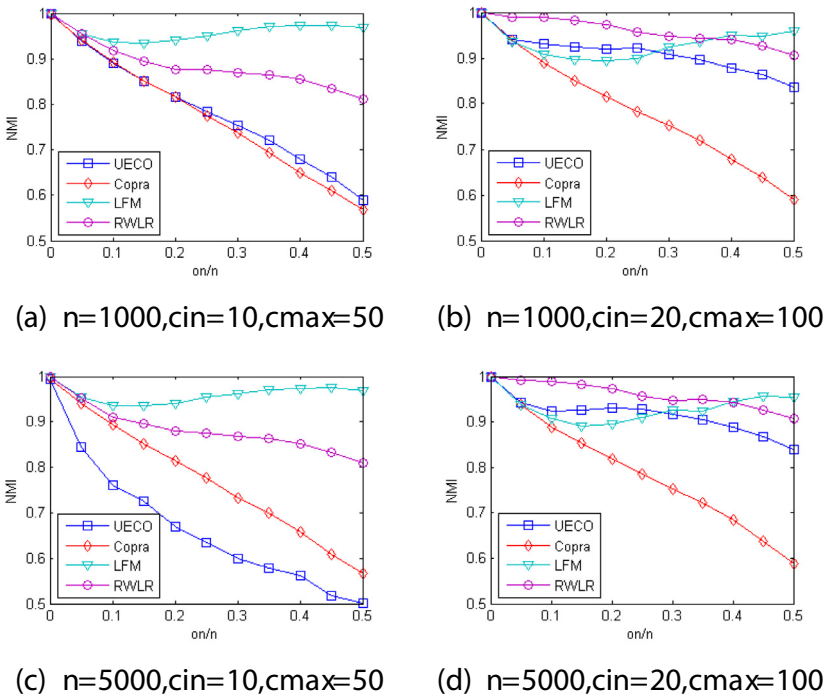


Fig. 7. Comparison results in benchmark networks.

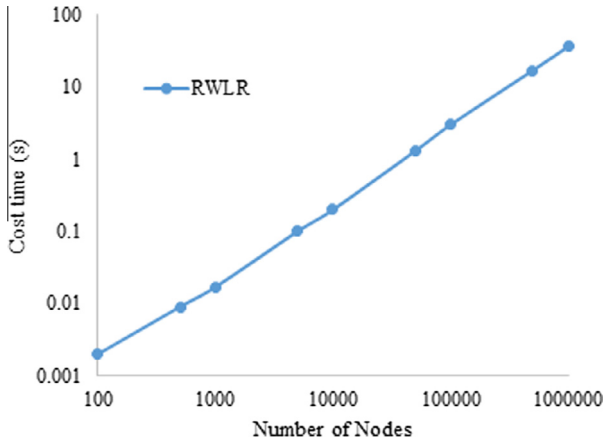


Fig. 8. The cost time that runs by RWLR in networks with nodes ranging from 100 to 1,000,000. *X* axis and *Y* axis are logarithmic that can show all the numbers clearly.

will be updated as well. Until the order is stationary, the nodes with close connections will be in the top of the order, which is beneficial to the division of HSM. To show the relationship between the number of step and the belonging coefficients, we depict Fig. 9. The difference between the two consecutive sorted node order, $O1$ and $O2$, is defined as $nnz(O1 - O2)$. We test it in six networks and the details are listed in Table 2.

For small networks in Fig. 9a, the orders convergence quickly and become stationary after 20 steps. For large networks, it's hard to walk enough among all the nodes within 20 steps, because most nodes have less connections to the seed. Then, we turn to focus on the first 100 nodes in the order which are often in the top levels of HSM. In Fig. 9, the orders can converge after 50 steps which is enough to catch the important top levels. With less connections to the seed, most nodes will be divided into bottom levels in HSM where we do not need to consider their order in the same levels. Consequently, we set the step T to a constant 50.

Apart from the convergence of the order, the number of step also influences the average value of belonging coefficients. In Fig. 9b,d, with the steps increasing, the belonging coefficients will increase as well. Small networks will have larger values than large networks since the random walker has more probability to arrive the seed. That's why the values in Fig. 3a are upon 0.9 while the values in Fig. 6b are much less.

4.4. Applications

4.4.1. Community visualization

In this section, we use HSM to analyze large real networks. It's hard to observe the whole graph in large networks. With the help of HSM, we map similar nodes into each level and obtain the overall picture of the whole network for the observed seed. In the picture, white circles indicate levels. The radius of each circle indicates the number of members at that level. Ranging from warm to cold, colors indicate the belonging coefficients from high to low. Since the number of nodes is large, we log them before normalizing them. The original number of nodes at each level are also labeled in the picture. We test in three large networks, Amazon, DBLP and Youtube networks, whose statistics are shown in Table 2. In Amazon network, nodes represent products. If two products are frequently co-purchased, a link between them will be contained in the network. In DBLP network, nodes represent authors. If two authors published at least one paper together, the network contains a link between them. In the Youtube network, nodes represent users and links represent friendship.

In Fig. 10, the pictures of HSMs with random seeds show the properties of each community. Fig. 10c has smooth color and uniform radii, which means the community boundaries are fuzzy. On the contrary, Fig. 10b has sharp change of colors between the first two levels, which means that there is a large gap of belonging

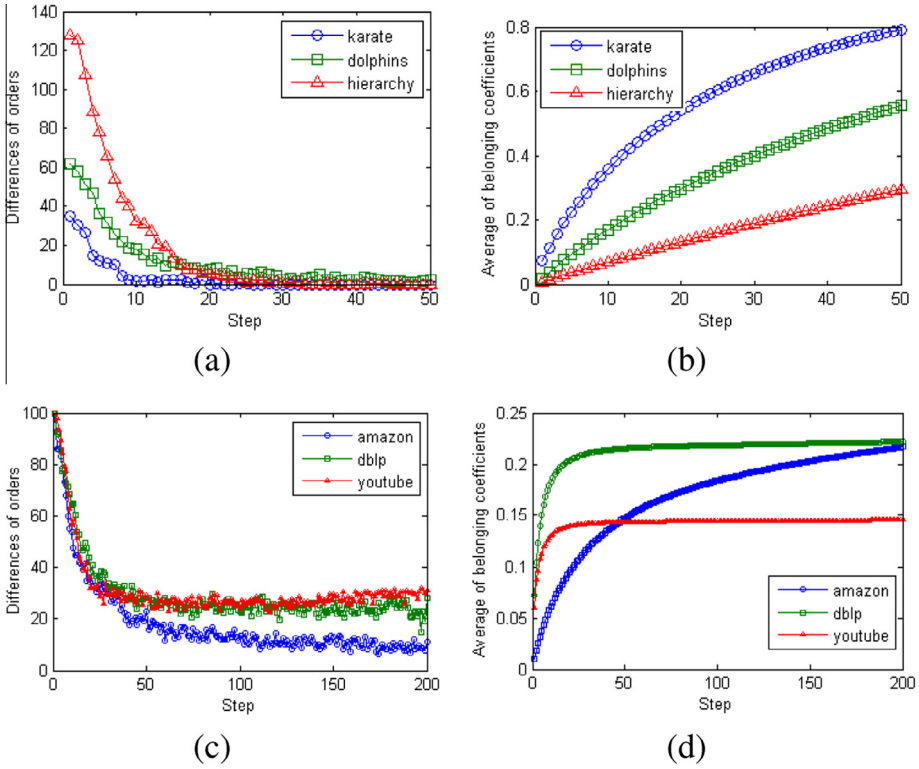


Fig. 9. Analysis of parameter step T. (a) and (c) Show the relationship between parameter step T and the rank of nodes. (b) and (d) Show the relationship between step T and belonging coefficients. The number of nodes are small in (a) and (b), while large in (c) and (d).

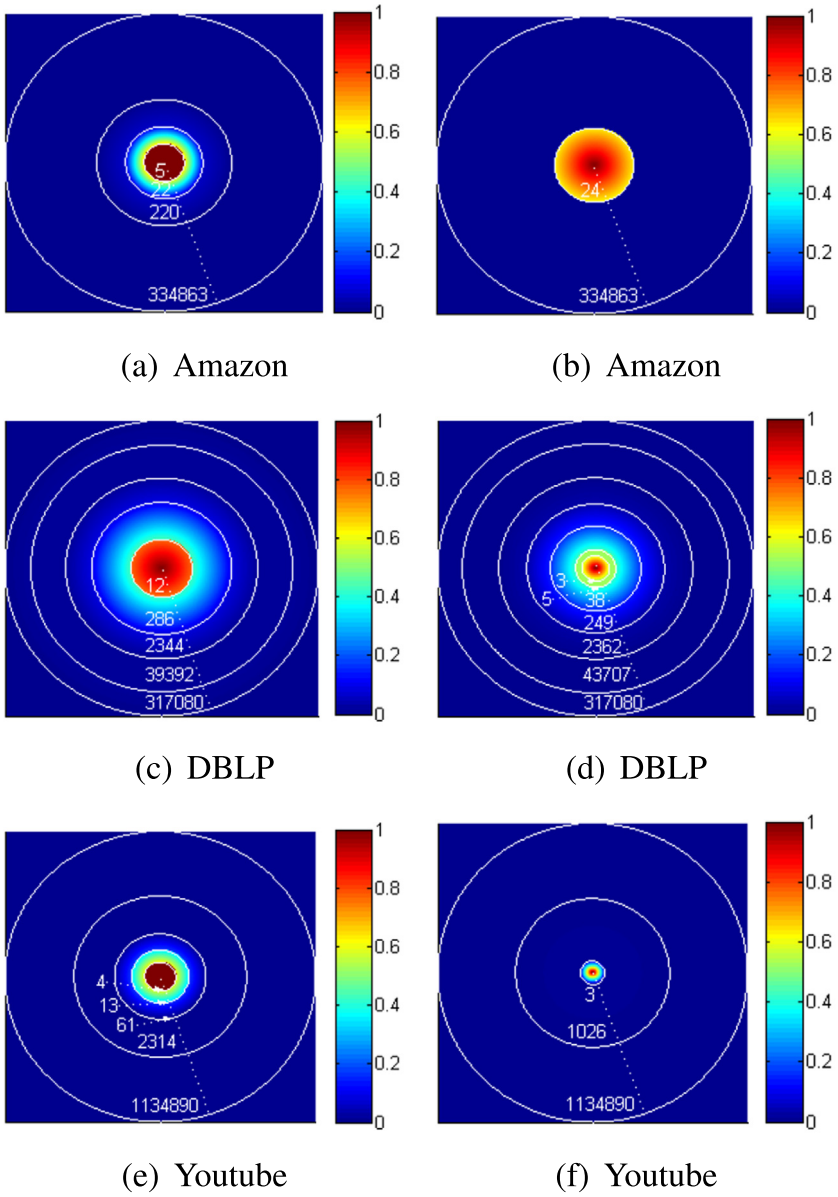
coefficients and the first level is more important to the community than others. Fig. 10d also has clear change of colors but not as sharp as Fig. 10b. Different from the figures mentioned above, Fig. 10a and e have few varied colors at their first levels. That means their core members are equally important and have similar connections in the network.

For all the figures, most belonging coefficients are gathered in the first several levels. In Fig. 10f and b, we should pay more attention to the first levels because the sum of their belonging coefficients almost equals to 1. With further division of levels, we can obtain finer resolutions in Fig. 11. From the figure, we can learn that the first 10 nodes are more closely connected, though they are not as strong as the first level in Fig. 10a and e. The colors in finer levels are as smooth as the original level, which shows their nodes are consistent in belonging coefficients.

The structure of HSMs are different for seeds in these different communities, while seeds in the same community have similar HSMs shown in Fig. 12 shows. Consequently, HSM can give an overall observation of the communities in large networks.

4.4.2. Interactive recommendation

Moreover, for those users who have purchased some specific products, we can use HSM to provide an interactive recommendation according to the product co-purchase relations. Traditional methods usually recommend top K products according to their similarity scores. However, small K has high precision but low recall, while large K has high recall but low precision. An interactive recommendation can be a better solution that provides multiple levels of products with dynamic K. User can choose the next level or zoom in the current level, where products have consistent probability to be purchased in the same level. We take product ID7438 as an example and its HSM is shown in Fig. 13a. In the figure, the products at the first level have the closest connections and we can take them as the first recommendation. If the user does not find the interest product, he can require the next level of recommendations, the second level. With the categories labeled in [33], we can see the two levels are two categories of the product ID7438. Since there are too many products at the second level, as shown in Fig. 13b, user can zoom the level, i.e. divide it further into



(a) Amazon (b) Amazon

(c) DBLP (d) DBLP

(e) Youtube (f) Youtube

Fig. 10. Visualizing different communities in large networks.

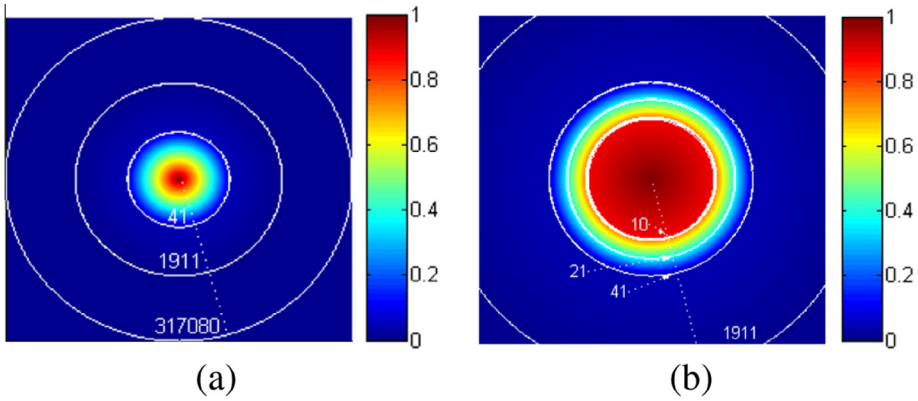


Fig. 11. The HSM with different number of levels. We divide the first level in (a) to obtain finer levels in (b).

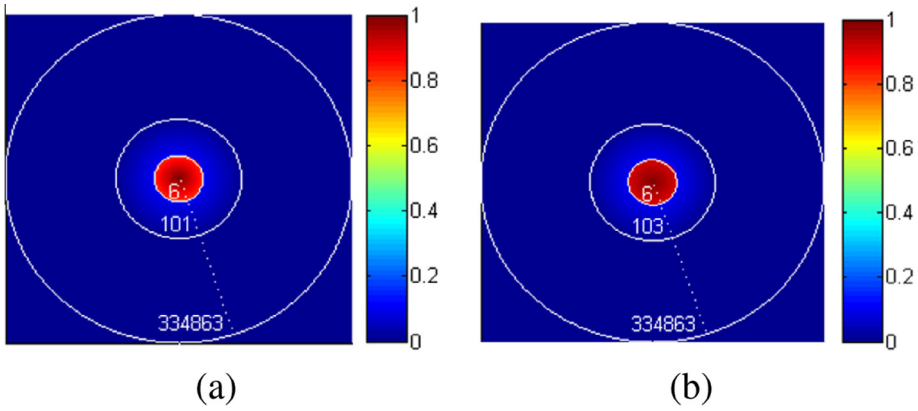


Fig. 12. The seeds of two HSMs in the same community.

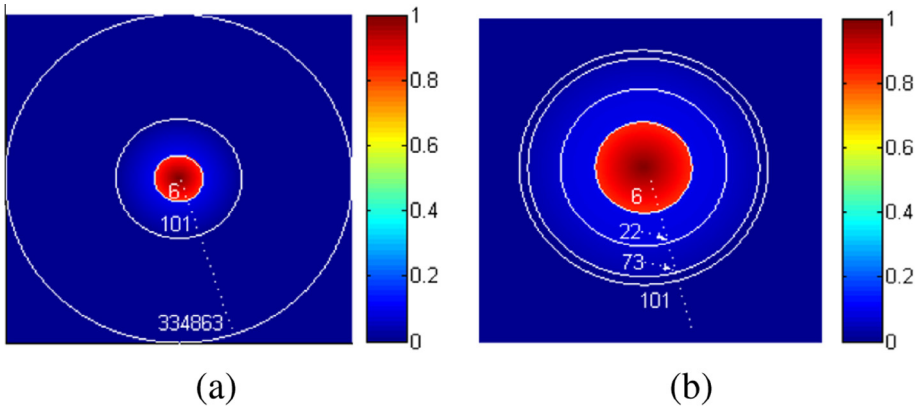


Fig. 13. The HSM of product ID7438 in Amazon network is shown in (a). We divide the second level to obtain finer levels in (b).

finer levels, and take the first 22 products as the next recommendation.

5. Discussion

The RWLR method has several advantages. First, RWLR satisfies the properties of HSM. The continuity and comparability is satisfied by the measurement based on random walk; the consistency is satisfied by the linear regression, and the stability is satisfied by selecting proper number of levels. Second, RWLR is also scalable. We use improved dynamic program to maximize Eq. (5) in linear time complexity $O(KN)$, where K is the number of levels and N is the number of nodes. Third, RWLR has few parameters that the only parameter, step T, can be set as a constant, such as 50 in our experiments. Finally, the experimental results show that the detected HSMs are meaningful. In Fig. 5, nodes on each level of HSM represent core member, overlapping member, bridge member, and so on.

As a two phase method, we can use any other measurements satisfying continuity and comparability, which are more suitable for specific networks. We use random walk based measurement as an example in this paper, which performances better in networks with large communities. Although RWLR focuses on detecting the levels of a community but not the partition of a graph, it can achieve common performance in the community detection.

In addition, we note that the efficiency depends on the selection of seeds, which is also important. There are methods [6] that focus on the selection of seeds which may improve the performance. The discussion of seeds is out of our scope in this paper under the community detection framework (see Fig. 1).

6. Conclusions and further study

In this paper, we introduce a novel concept, hierarchical structure of members (HSM), to describe the community structure in a 'level' way. Each level of HSM can be regarded as a stable community boundary, which forms the multi-resolution of community. Besides, nodes in the same level have consistent belonging coefficients, which keeps them sharing similar properties. HSM can reveal more information from the relations among levels and the relations among members in the same level. We also propose a RWLR method to detect the HSM. Experiments show that the detected HSMs can reveal multi-resolution of communities and the relations among members. The only parameter, step T, can be set to a constant. The method is scalable since it's time complexity is linear to the number of nodes in the network. We also test the community structures on benchmarks against competitive methods, and the result shows the good performance of RWLR in networks with large communities and small overlaps. Taking the consistency of belonging coefficients into account, we can prevent unreasonable community generated by the local maximal Fitness. Furthermore, we apply the HSM to draw an overall picture of the large network and provide interactive recommendations in Amazon network.

In the future, we will apply the HSM to more social network analysis and mine its potential power. Besides, we will improve the detection method by trying other measurement and taking seed selection into consideration.

Acknowledgements

The research was supported in part by Natural Science Foundation of China (No. 60903071), National Basic Research Program of China (973 Program, No. 2013CB329605), Specialized Research Fund for the Doctoral Program of Higher Education of China, and Training Program of the Major Project of BIT.

Appendix A. Proof of the existence of the path from any node to source

In a connected component, we need to prove that for any node i, if there exists a neighbor j that $BC(j) > BC(i)$, we move it to node j. Otherwise, i is the source node.

Proof. (By Contradiction) We assume that there exists a node i that has no neighbor with higher BC and is not source node. The neighbors of i that connect to i and source node are noted as S. For any node j in S, j is in a path from i to source node, and j is more closer to source node than i, so $BC(j) > BC(i)$. It is contrary to the assumption. Therefore, for node i, moving along the ascending order of BC, it will arrive at source node. □

References

[1] B. Amiri, L. Hossain, J.W. Crawford, R.T. Wigand, Community detection in complex networks: multiobjective enhanced firefly algorithm, Knowl.-Based Syst. 46 (0) (2013) 1–11.
[2] Y. Bo, J. Liu, J. Feng, On the spectral characterization and scalable mining of network communities, IEEE Trans. Knowl. Data Eng. 24 (2012) 326–337.
[3] D. Chen, M. Shang, Z. Lv, Y. Fu, Detecting overlapping communities of weighted networks via a local algorithm, Physica A: Stat. Mech. Appl. 389 (19) (2010) 4177–4187.
[4] Y. Cui, X. Wang, J. Li, Detecting overlapping communities in networks using the maximal sub-graph and the clustering coefficient, Physica A: Stat. Mech. Appl. 405 (2014) 85–91.
[5] S. Fortunato, Community detection in graphs, Phys. Reports 486 (3) (2010) 75–174.
[6] D.F. Gleich, C. Seshadhri, Vertex neighborhoods, low conductance cuts, and good seeds for local community methods, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 597–605.
[7] S. Gregory, Finding overlapping communities in networks by label propagation, New J. Phys. 12 (2010).
[8] S. Gregory, Fuzzy overlapping communities in networks, J. Stat. Mech.: Theory Exp. 2011 (02) (2011) P02017.
[9] R. Guimera, L.A.N. Amaral, Functional cartography of complex metabolic networks, Nature (2005).
[10] F. Havemann, M. Heinz, A. Struck, J. Gläser, Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels, J. Stat. Mech.: Theory Exp. 2011 (01) (2011) P01023.
[11] J. Huang, H. Sun, J. Han, H. Deng, Y. Sun, Y. Liu, SHRINK: a structural clustering algorithm for detecting hierarchical communities in networks, in: International Conference on Information and Knowledge Management, 2010, pp. 219–228.
[12] D. Jin, B. Yang, C. Baquero, D. Liu, D. He, J. Liu, A markov random walk under constraint for discovering overlapping communities in complex networks, J. Stat. Mech.: Theory Exp. 2011 (05) (2011) P05031.
[13] A. Lancichinetti, S. Fortunato, Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities, Phys. Rev. E 80 (1) (2009) 016118.
[14] A. Lancichinetti, S. Fortunato, J. Kertész, Detecting the overlapping and hierarchical community structure in complex networks, New J. Phys. 11 (3) (2009) 033015.
[15] C. Lee, F. Reid, A. McDaid, N. Hurley, Detecting highly overlapping community structure by greedy clique expansion, arXiv preprint arXiv:1002.1827, 2010.
[16] J. Leskovec, K.J. Lang, A. Dasgupta, M.W. Mahoney, Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters, Computing Research Repository abs/0810.1, 2008.
[17] Y. Li, H. Wang, J. Li, H. Gao, Efficient community detection with additive constrains on large networks, Knowl.-Based Syst. 52 (0) (2013) 268–278.
[18] J. Liu, Fuzzy modularity and fuzzy community structure in networks, Eur. Phys. J. B 77 (2010) 547–557.
[19] D. Lusseau, The emergent properties of a dolphin social network, Proc. Roy. Soc. Lond. Ser. B: Biol. Sci. 270 (Suppl. 2) (2003) S186–S188.
[20] A. McDaid, N. Hurley, Detecting highly overlapping communities with model-based overlapping seed expansion, in: Advances in Social Network Analysis and Mining, 2010, pp. 112–119.
[21] A. McDaid, N. Hurley, Detecting highly overlapping communities with model-based overlapping seed expansion, in: 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2010, pp. 112–119.
[22] T. Nepusz, A. Petrczi, L. Ngyessy, F. Bazs, Fuzzy communities and the concept of bridgeness in complex networks, Phys. Rev. E 77 (2008).
[23] I. Psorakis, S. Roberts, M. Ebden, B. Sheldon, Overlapping community detection using Bayesian non-negative matrix factorization, Phys. Rev. E 83 (2011).
[24] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. 101 (2004) 2658–2663.
[25] W. Ren, G. Yan, X. Liao, L. Xiao, Simple probabilistic algorithm for detecting community structure, Phys. Rev. E 79 (2009).
[26] A. Stanoev, D. Smilkov, L. Kocarev, Identifying communities by influence dynamics in social networks, Phys. Rev. E 84 (4) (2011) 046102.
[27] J.J. Whang, D.F. Gleich, I.S. Dhillon, Overlapping community detection using seed set expansion, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, ACM, 2013, pp. 2099–2108.
[28] J. Xie, S. Kelley, B.K. Szymanski, Overlapping Community Detection in Networks: the State of the Art and Comparative Study, 2011.
[29] J. Xie, B.K. Szymanski, Towards linear time overlapping community detection in social networks, in: Advances in Knowledge Discovery and Data Mining, Springer, 2012, pp. 25–36.
[30] J. Yang, J. Leskovec, Community-affiliation graph model for overlapping network community detection, in: 2012 IEEE 12th International Conference on Data Mining (ICDM), IEEE, 2012, pp. 1170–1175.
[31] J. Yang, J. Leskovec, Overlapping community detection at scale: a nonnegative matrix factorization approach, in: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, ACM, 2013, pp. 587–596.
[32] W. Zachary, An information flow model for conflict and fission in small groups1, J. Anthropol. Res. 33 (4) (1977) 452–473.
[33] M.J. Zaki, A. Siebes, J.X. Yu, B. Goethals, G.I. Webb, X. Wu (Eds.), 12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10–13, 2012, IEEE Computer Society, 2012.
[34] S. Zhang, R.-S. Wang, X.-S. Zhang, Identification of overlapping community structure in complex networks using fuzzy c-means clustering, Physica A: Stat. Mech. Appl. 374 (1) (2007) 483–490.