JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 33, 1323-1341 (2017)
DOI: 10.6688/JISE.2017.33.5.13

Community Number Estimation for Community Detection in Complex Networks*

ZHIXIAO WANG[1], JINGKE XI[1,+], YAN XING[1] AND ZHIGUO HU[2]
[1]*School of Computer Science and Technology*
China University of Mining and Technology
Xuzhou Jiangsu, 221116 P.R. China
[2]*School of Computer and Information Technology*
Shanxi University
Taiyuan, 030006 P.R. China
E-mail: xjk@cumt.edu.cn

Most current community detection methods for complex networks focus on partition. Community number estimation does not have the due attention it deserves, and the community number is only a by-product of community partition. In fact, knowing the community number in advance can speed up the partition process, especially for large scale and dynamic complex networks. This paper proposes a community number estimation method based on topology potential. In the topology potential field, the potential distribution of nodes shows a natural peak-valley structure, and each community corresponds to a local high potential area. The number of local maximum potential nodes is the estimated community number. Experiments on real world networks and artificial networks show that the proposed method gives very good performance in community number estimation. The more noticeable the peak-valley structure of the corresponding topology potential field is, the closer the estimated community number will be to the ground truth. Furthermore, compared with state-of-the-arts methods, our proposed method is not sensitive to the tuned parameter, and shows good efficiency.

Keywords: complex network, community number estimation, topology potential, peak-valley structure, community detection

1. INTRODUCTION

Many real networks present a "community structure", *i.e.* groups of vertices that have a higher density of edges within them and a lower density of edges between them [19]. For example, in social networks, a person usually gets involved in different social groups such as family, friends, and colleagues; in protein networks, proteins with a similar function may participate in the same module. Community detection is beneficial for understanding the structure of networks and forecasting the behavior of networks [20]. For example, it can be used to forecast the information propagation in social networks, and to recognize functions of proteins in bioengineering networks [23].

Community detection approaches can be categorized into three types: (1) disjoint community detection; (2) overlapping community detection; and (3) dynamic community

Received August 7, 2016; revised November 23, 2016 & January 20, 2017; accepted February 18, 2017.
Communicated by Tzung-Pei Hong.
+ Corresponding author.
* This work was supported in part by National Key Research and Development Program (No. 2016YFC-060908), the National Natural Science Foundation of China (No. 61402482, No. 51674255), China Postdoctoral Science Foundation (No. 2015T80555), Jiangsu Planned Projects for Postdoctoral Research Funds (No. 1501012A), and Qing Lan Project of Jiangsu Province.

detection. For disjoint community detection, one node belongs to only one community. A comparative evaluation of disjoint community detection algorithms is shown in [17]. Since overlapping is indeed a significant feature of real world networks, a considerable amount of work has focused on overlapping community detection. The state-of-the-art overlapping community detection algorithms are reviewed in [21]. Besides overlapping, dynamic is another important feature of complex networks. Much research has also been conducted in dynamic complex networks [6].

Estimating the community number is able to benefit these community partition methods. On the one hand, the community number in some sense can be an alternative terminal condition for time-consuming partition iteration [3] which is considered as the hardest nut to crack. On the other hand, as an alternative parameter, the community number can make the partition process more explicit. With the booming of large-scale networks like Facebook, Twitter, blogs and personal mobile communication, more attention has been attached to the detection speed than accuracy.

However, community number estimation does not have the due attention it deserves. To the best of our knowledge, it is only a by-product of most current community detection methods, and only limited research has been done to solve this. Some traditional methods obtains the optimal community number via modularity Q. High Q value means good partition, and the partition corresponding to a maximum should be the best [15]. However, the limitations of the modularity function have been reported in [2]. In addition, Ulrik Brandes *et al.* [1] proved that the modularity optimization is an NP-complete problem. Therefore, a more simple and effective method for community number estimation is needed. FU *et al.* [3] put forward a fast community number estimation algorithm based on maximizing separability measure, which uses the separability function to select large degree nodes as core node candidates, and then determines the final core nodes via an improved shortest path seeking algorithm. The core node number is the estimated community number. He *et al.* [8] took the discrete nodal domain theory as a criterion to determine the community number. They introduced the weak-nodal-domain to spectral partition, and identified the cluster number by exploiting the community structure information contained in the Laplacian eigenvectors. However, when the community structure is not clear, it is difficult to obtain the proper community number from the eigenvector elements [14]. Li *et al.* [11] detected the optimal community number based on the theoretic information. They proposed a statistic index $\Omega(c)$ based on the similarity of communities, and found that when the $\Omega(c)$ reaches its local maximum, the corresponding c is likely to be the optimal community number. This method can only reveal the community number, and cannot identify core nodes of each community. However, these core nodes are very crucial in the follow-up community partition, especially for local-expansion-based methods. Furthermore, when a complex network does not have a definite community structure, this method cannot get any maximum. M.E.J Newman and Gesine Reinert [16] described a mathematically principled approach for finding the number of communities in a network using a maximum-likelihood method. Liu *et al.* [13] adopted the cluster analysis method to estimate the optimal number of communities. Their method bases on the idea that community centers are characterized by higher density than their neighbors.

Despite the efforts mentioned above, community number estimation is still not totally solved [18]. In this paper, we apply the topology potential field to community number estimation. The contributions of this paper are summarized as follows:

- The proposed method utilizes the inherent peak-valley structure of the topology potential field to estimate the community number. In the topology potential field, the potential distribution of nodes shows a natural peak-valley structure, and each community corresponds to a local high potential area. The number of local maximum potential nodes, located at the center of each local high potential area, is the estimated community number. Compared with other methods, our proposed method is not sensitive to the tuned parameter (*i.e.* impact factor), and shows good performance in efficiency.
- The proposed method also defines a concavity-convexity measure to analyze the relationship between the accuracy of the estimated number and the characteristics of the peak-valley structure. Experimental results show that the more noticeable the peak-valley structure is, the more precise the estimated community number will be.
- The proposed method can accurately identify the core nodes of each community before community partition. These core nodes are very crucial in the follow-up community partition, especially for local-expansion-based community detection methods.

The rest of this paper is organized as follows. Section 2 introduces the topology potential field. Section 3 analyses the peak-valley structure of the topology potential field. Section 4 presents the community number estimation algorithm based on topology potential. Section 5 discusses the experimental results and Section 6 provides the conclusion of this paper.

2. TOPOLOGY POTENTIAL FIELD

In physics, the "field" concept was proposed to describe a non-contact interaction between material particles, which can be divided into two types: short-range fields and long-range fields. With the development of the classical field theory, it has become a mathematical model describing the non-contact interactions between objects [4]. The nodes in a complex network are not isolated but linked by edges. Therefore, the field model can be used to describe the interaction and the association among network nodes. Each node is regarded as a field source, and these nodes interact with each other, forming a field called the topology potential field. Interactions among nodes are always within the local neighborhood and do not influence all the nodes in the whole network. Each node's influence will quickly drop with the distance increasing, in accordance with the properties of short-range fields. Hence, we define the topology potential in the form of Gaussian function, which belongs to the nuclear force field, a typical short-range field.

Definition 1: Given a complex network $G = (V, E)$, where $V = \{\upsilon_i | i = 1, \ldots, n\}$ indicates a set of nodes, n denotes the total number of nodes, and $E = \{(\upsilon_i, \upsilon_j) | \upsilon_i, \upsilon_j \in V\}$ represents a set of edges, the topology potential of node υ_i can be defined as follows [15]:

$$\varphi(\upsilon_i) = \sum_{j=1}^{n} m(\upsilon_j) \times e^{-\left(\frac{d_{ij}}{\sigma}\right)^2} \tag{1}$$

where, $\upsilon_i, \upsilon_j \in V$, $\varphi(\upsilon_i)$ represents the topology potential of node υ_i; $m(\upsilon_i)$ represents the mass of node υ_j; d_{ij} refers to the hops between node υ_i and υ_j; σ indicates an impact fac-

tor used to control the influence scope of the node. According to the properties of Gaussian function, for a given σ, its influence region approximates to $\lfloor 3\sigma / \sqrt{2} \rfloor$ hops [7].

There are three parameters in Formula (1): $m(\upsilon_i)$, d_{ij} and σ. Almost all research on the topology potential field ignores the mass difference between nodes, such as [4, 7, 22]. Similar to these references, this paper sets $m(\upsilon_i) = 1$. d_{ij} is the distance between node υ_i and υ_j. If $d_{ij} > \lfloor 3\sigma / \sqrt{2} \rfloor$, then the topology potential component produced by node υ_j on node υ_i is ignored. Thus, only one parameter needs to be tuned in Formula (1), *i.e.* the impact factor σ. Potential entropy is used to select the optimal impact factor σ [18].

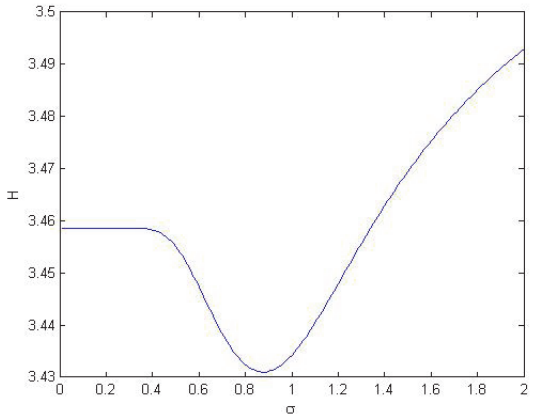

Fig. 1. The optimal σ selection of the Zachary network.

Definition 2: Given a complex network $G = (V, E)$, where $V = \{\upsilon_i | i = 1, \ldots, n\}$ indicates a set of nodes, $P = \{\varphi(\upsilon_i) | i = 1, \ldots, n\}$ represents the corresponding topology potential value of nodes, and n refers to the total number of nodes, the potential entropy H can be defined as follows:

$$H = -\sum_{i=1}^{n} \left[\frac{\varphi(\upsilon_i)}{Z} \times log\left(\frac{\varphi(\upsilon_i)}{Z} \right) \right] \tag{2}$$

where $Z = \sum_{i=1}^{n} \varphi(\upsilon_i)$ denotes a normalization factor.

This paper takes a typical social network – the Zachary network as an example to show the optimal σ selection. Fig. 1 shows the changes of the potential entropy H with the value of the impact factor σ. According to the potential entropy theory, when H reaches its minimum value, the topology potential field has minimum uncertainty and the corresponding σ is the best. Zhang *et al.* have proved the existence of the optimal σ from a mathematical perspective in [22]. In this paper, we adopt the IFO (Influence Factor Optimization) algorithm proposed by Li *et al.* in [12] to obtain the optimal impact factor. With IFO algorithm, the complexity of the optimal impact factor selection is no more than $O(n^2)$. Li *et al.* [12] found that the optimal impact factor varies within a narrow range, and the influence scope of nodes in topology potential fields of most middle scale realworld networks is 2 or 3 hops. Zachary is a small scale network, and the optimal σ

for this network is 0.84 (as shown in Fig. 1). Thus, the influence scope of nodes in the topology potential filed of this network is $\lfloor 3\sigma / \sqrt{2} \rfloor$=1 hop.

Actually, our proposed method is not sensitive to this impact factor, which can still estimate the correct community number even when the optimal value of the impact factor is not given. The experimental results in Section 5.1 will confirm this point.

3. PEAK-VALLEY STRUCTURE OF TOPOLOGY POTENTIAL FIELD

As can be seen from Formula (1), the topology potential of a node is a composition of the topology potential components produced by its neighbors. It is defined as the differential position of each node in the topology, *i.e.*, the potential of each node in its position [7]. The topology potential value of each node reflects its degree to be influenced by other nodes in the network, and the potential distribution characterizes the structure of nodes in the topology space. In the topology potential field, nodes with higher influence on others are grouped together, and each group corresponds to a local high potential area [7]. So we can utilize this property to estimate the community number of networks. There are two main types of nodes in the topology potential field: peak nodes and valley nodes. Peak nodes are the local maximum potential nodes, located at the center of local high potential areas. They are the representative nodes of each community. The number of local maximum nodes is the estimated community number. The valley nodes are located at relatively low positions in the topology potential field, jointing local high potential areas. They are the overlapping nodes among communities. Fig. 2 shows the topology potential distribution of the Zachary network nodes. On the whole, the topology potential field presents a natural peak-valley structure. Some nodes, with considerably large topology potential value, are located at relatively high positions, such as Node 1 and Node 34. These two nodes correspond to two communities, respectively. Some nodes, with relatively small topology potential value, are located at relatively low positions, such as Node 10 and Node 20.

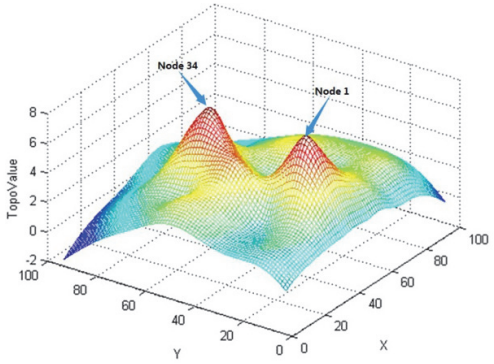

Fig. 2. Topology potential distribution of nodes in the Zachary network.

We can directly identify the community number based on the natural peak-valley structure of the topology potential field. In practice, the community number of some networks are easy to estimate, while some others' may be very difficult. In this paper, we

define a measure named *cc* for mining the relationship between the accuracy of community number estimation and characteristics of the peak-valley structure in the topology potential field. The *cc* measure quantifies the concavity-convexity of the peak-valley structure in the topology potential field.

Definition 3: Given a complex network $G = (V, E)$, where $V = \{\upsilon_i | i = 1, \ldots, n\}$ indicates a set of nodes, $P = \{\varphi(\upsilon_i) | i = 1, \ldots, n\}$ denotes the corresponding topology potential value of nodes, the concavity-convexity of the topology potential field can be defined as follows:

$$cc = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(k(\upsilon_i)\cdot\varphi(\upsilon_i)^* - \overline{k(\upsilon_i)\cdot\varphi(\upsilon_i)^*})^2}, \tag{3}$$

$$\varphi(\upsilon_i)^* = \frac{\varphi(\upsilon_i) - min(P)}{max(P) - min(P)}. \tag{4}$$

In Formulas (3) and (4), *cc* denotes the concavity-convexity of the topology potential field; $min(P)$, $max(P)$ represent the maximum and minimum of P, respectively; $\varphi(\upsilon_i)$ refers to the topology potential of node υ_i, $1 \le i \le n$; $\varphi(\upsilon_i)^*$ indicates the normalization of $\varphi(\upsilon_i)$; $k(\upsilon_i)$ denotes the *k*-shell value [9] (*i.e. ks*) of node υ_i, $1 \le i \le n$.

The *k*-shell value reflects the location and importance of a node in a complex network [9]. It can be obtained with the following steps. Firstly, we delete the 1-degree nodes, and then scan the whole network to delete the new emerging 1-degree nodes. Repeat this step until there are not any 1-degree nodes. These deleted nodes are assigned *ks*=1. Secondly, we delete all 2-degree nodes in the same way, and assign *ks*=2 to these deleted nodes. Finally, all nodes will be assigned a *ks* value by repeating these steps. The bigger the *k*-shell value is, the more significant the node will be. More details about *k*-shell can be found in [9].

4. COMMUNITY NUMBER ESTIMATION BASED ON TOPOLOGY POTENTIAL

The topology potential of a node is a composition of the topology potential components produced by its neighbors. Some nodes have more neighbors than others, and naturally the topology potential components produced by their neighbors are more than others, thus these nodes will become local maximum potential nodes in the topology potential field. Also some nodes may only have a few neighbors, and the topology potential components produced by the neighbors of these nodes are relatively less, leading to low potential positions in the topology potential field. That is, the distribution of the topology potential value in the topology potential field is in accordance with the clustering of nodes in the complex networks. So we can utilize the concavity-convexity of the peak-valley structure in the topology potential field to estimate the community number of networks.

In this section, we propose a community number estimation algorithm based on Topology Potential. The topology potential field shows a natural peak-valley structure. Each community corresponds to a local high potential area. The corresponding local

maximum potential node can be regarded as the representative node of each community. The number of local maximum potential nodes is the estimated community number. Note that if the distance between two local maximum potential nodes is shorter than the maximum influence scope of nodes, *i.e.* $\lfloor 3\sigma / \sqrt{2} \rfloor$ hops, the node with the smaller topology potential value will be regarded as the combined representative node.

Local maximum potential nodes search is crucial for community number estimation. Hill-climbing [7] is a popular method for handling this task, which may leave out some local maximum potential nodes. Furthermore, initial node selection will affect the search performance. Different from the traditional Hill-climbing method, the proposed method directly determines whether a node is a local maximum potential node or not.

Algorithm 1: Community number estimation algorithm based on topology potential.
Input: network $G = (V, E)$, $|V| = n$, $|E| = m$;
Output: local maximum potential nodes $C_{maximum}$, and estimated community number k.
1. $\{C_{maximum} = \phi$;
2. For int $i = 1$ to n
3. {Calculate $\varphi(\upsilon_i)$ with Formula (1);
4. Select the optimal impact factor σ with Formula (2);}
5. For int $i = 1$ to n
6. If $\varphi(\upsilon_i) > \varphi(neighbor(\upsilon_i))$
7. $\{\upsilon_i$ is a local maximum potential node;
8. $C_{maximum} = C_{maximum} \cup \{\upsilon_i\};\}$
9. int $q = |C_{maximum}|$;
10. For int i=1 to q
11. For int j=i+1 to q
12. If $distance(\upsilon_i, \upsilon_j) < \lfloor 3\sigma / \sqrt{2} \rfloor$
13. If $\varphi(\upsilon_i) \geq \varphi(\upsilon_j)$
14. $\{C_{maximum} = C_{maximum} - \{\upsilon_j\}$;
15. Mark υ_j as a combined representative node;}
16. Else
17. $\{C_{maximum} = C_{maximum} - \{\upsilon_i\}$;
18. Mark υ_i as a combined representative node;}
19. $k = |C_{maximum}|;\}$

Now, we analyze the complexity of the proposed algorithm. Step 1 is the initialization; Steps 2-4 are for the topology potential calculation. According to [4, 7], the complexity of topology potential calculation is $O(n^2)$. Steps 5-8 search for all local maximum potential nodes. The process is completed after a one-time node traversal, and the complexity is $O(n)$. Steps 9-19 identify local maximum potential nodes within $\lfloor 3\sigma / \sqrt{2} \rfloor$ hops, and the complexity is $O(q(q-1)/2)$, $q = |C_{maximum}|$, $1 < q \ll n$. The above analysis shows that the total complexity of the proposed algorithm is $O(n^2)+O(n)+O(q(q-1)/2) = O(n^2)$.

5. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method, we compare it with other four typical methods, including MSM (Maximizing Separability Measure in [3]), WNDP

(Weighted Nodal Domain Partition in [8]), $\Omega(c)$ measure in [11], and the maximum-likelihood method in [16] (named MIM in this section), through exhaustive experiments on artificial networks and real world networks. The artificial networks are generated using the LFR Benchmark generator [10], which can produce the required networks with implanted communities. It is true that, even with the same parameters, the LFR generator cannot produce the exactly same artificial networks in each run. In order to obtain relatively accurate results, we use the average performance of 50 runs as the final results. Real world networks include the Zachary network, Dolphin network, Political book network, Football network, Jazz musicians network, C. elegans metabolic network, PGP network, Collaboration network, DBLP collaboration network, and Youtube online social network. Some of them are taken from http://www-personal.umich.edu/~mejn/netdata/, and others are from http://snap.stanford.edu/data/index.html.

5.1 The Sensitivity of Our Proposed Method to the Impact Fact σ

In the topology potential calculation, only one parameter needs to be tuned, *i.e.* the impact factor σ. In this subsection, the Zachary and Dolphin networks are selected to evaluate the sensitivity of our proposed method to the impact factor σ. In order to evaluate the sensitivity of our proposed method to the impact factor σ, the value of the impact factor ranges from 0 to 5, and the corresponding local maximum potential nodes are identified by our proposed method. Fig. 3 shows the results of our proposed method on the Zachary network with the varying impact factor. Fig. 4 shows that of the Dolphin network.

The Zachary network consists of two known communities. The optimal impact factor σ of the Zachary network can be obtained by the IFO method described in Section 2, and the corresponding result is 0.840. However, Fig. 3 reveals that as long as the impact factor $0 < \sigma < 0.942$, our proposed method can always identify two local maximum potential nodes, and get the correct community number.

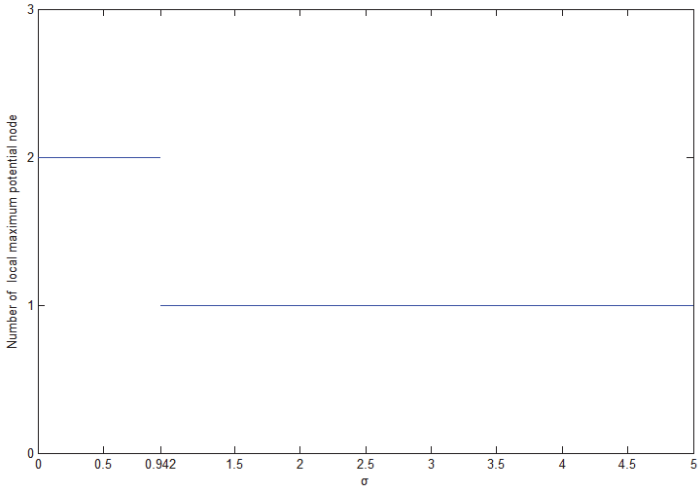

Fig. 3. Local maximum potential nodes of the Zachary network with the variation of σ.

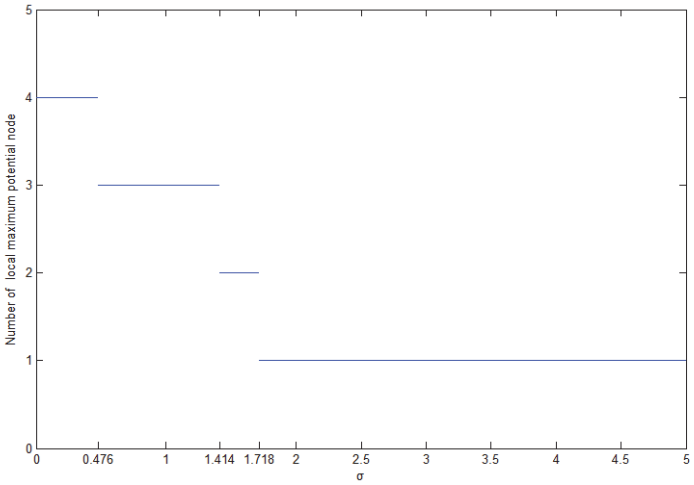

Fig. 4. Local maximum potential nodes of the Dolphin network with the variation of σ.

The Dolphin network is composed by two families. The IFO method of Section 2 reveals that the optimal impact factor σ of the Dolphin network is 1.3. Fig. 4 shows that when the impact factor $0.476 \leq \sigma < 1.414$, 3 local maximum potential nodes are identified: Nodes 15, 18 and 21. The distance between nodes 15 and 18 is 2 hops. If $0.943 \leq \sigma < 1.414$, then $\lfloor 3\sigma / \sqrt{2} \rfloor = 2$, and nodes 15 and 18 are in the same community according to Algorithm 1. Thus, the final estimated community number is 2 even 3 local maximum potential nodes are identified. When $1.414 \leq \sigma < 1.718$, 2 local maximum potential nodes are detected, which is consistent with the ground truth. To sum up, as long as the impact factor $0.943 \leq \sigma < 1.718$, we can get the correct community number.

Figs. 3 and 4 show that the performance of our proposed method is not sensitive to the impact factor σ. Our proposed method can still estimate the correct community number even if the impact factor varies within a certain range rather than being given the optimal value.

5.2 Community Number Estimation on Real World Networks

5.2.1 Zachary network

The Zachary karate network has 34 nodes and 78 edges, which describes the social interactions between members of the karate club at an American university. The Zachary network consists of two known communities.

The proposed method identifies 2 local maximum potential nodes (*i.e.* Node 1 and Node 34, with their respective topology potential of 7.2 and 7.4), indicating there are 2 communities, which is equal to the well-proven community number in the Zachary network. Fig. 2 shows the topology potential distribution of nodes in the Zachary network, where there are 2 obvious local high potential areas. The community number estimated by MSM, WNDP, and $\Omega(c)$ are all 2. For the MIM method, the number 2 has the highest probability.

5.2.2 Dolphin network

The Dolphin network describes the frequent associations between 62 dolphins living off Doubtful Sound, New Zealand, which is composed by 2 dolphin families.

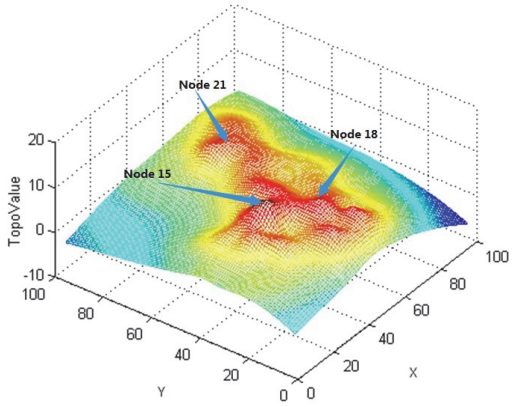

Fig. 5. Topology potential distribution of nodes in the Dolphin network.

The proposed method gets 3 local maximum potential nodes: Node 15, Node 18 and Node 21, with their respective topology potential of 9.61, 7.10 and 8.23. For the Dolphin network, the corresponding optimal σ is 1.3, then $\lfloor 3\sigma / \sqrt{2} \rfloor$=2. The distance (*i.e.* hops) between Node 15 and Node 18 is 2 hops. According to the proposed method, they belong to the same community, and Node 18, with a smaller topology potential value than Node 15, is deleted from the set of local maximum potential nodes. Consequently, the final local maximum potential nodes are Node 15 and Node 21, which tells a 2-community network. This result is consistent with the known community number in the Dolphin network. Fig. 5 shows the topology potential distribution of nodes in the Dolphin network. Apparently, Node 15 and Node 18 are located at the identical local high potential area. However, the community number estimated by MSM method is 3. The reason is the threshold of the shortest path length is set at 1 hop. Therefore, Node 15 and Node 18 are not regarded as in the same community. The community number identified by WNDP and $\Omega(c)$ are all 2. For the MIM method, the number 2 has the highest probability.

5.2.3 Political book network

The Political book network contains 105 books taken from the online bookseller Amazon.com. These books describe different political views in 2004 around the US election. Edges between books represent frequent co-purchasing by a same buyer. There are 3 communities in the Political book network corresponding to liberal, neutral and conservative fractions.

The proposed method derives 2 local maximum potential nodes: Node 13 and Node 85 with corresponding topology potential 10.29 and 9.51 respectively. This indicates that the proposed method finds 2 communities (*i.e.* liberal and conservative communities),

leaving the neutral community out. The reason is that the neutral community contains only 13 nodes, virtually scattering at the borders between liberal and conservative communities. With their smaller topology potential, it is hard to detect the third local maximum potential node. Fig. 6 shows the topology potential distribution of nodes in the Political book network. As a matter of fact, considerable state-of-the-art methods split the Political book network into two communities rather than three communities, and they regard the neutral nodes as the overlapping nodes between liberal and conservative communities [7]. The community number identified by MSM, WNDP and $\Omega(c)$ are all 2. For the MIM method, the number 3 have the highest probability.

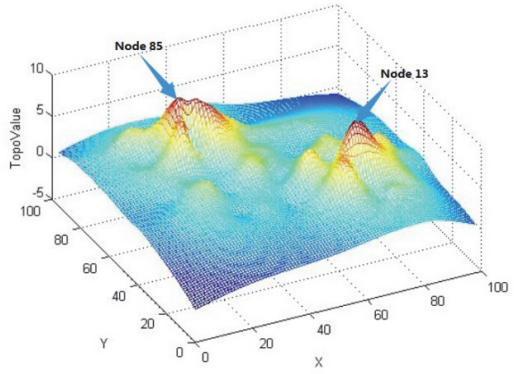

Fig. 6. Topology potential distribution of nodes in the political book network.

5.2.4 Football network

The American College Football network [5] contains 115 teams, among which 616 games are carried out. In the network, nodes represent teams and edges games. All teams are organized into 12 conferences, each of which contains about 8-12 teams. Since most matches are carried out within conferences, the conference number 12 is regarded as the community number.

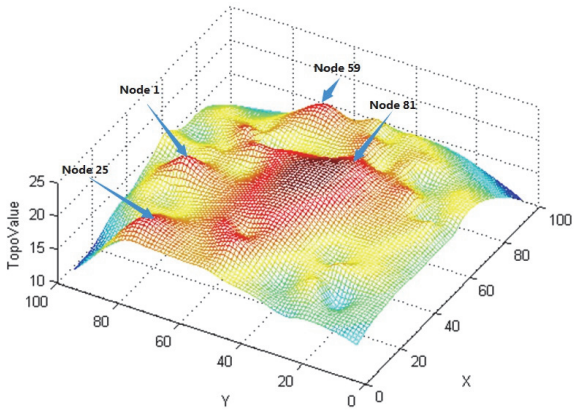

Fig. 7. Topology potential distribution of nodes in the football network.

The application of the proposed method locates 4 local maximum potential nodes: Node 1, Node 25, Node 59 and Node 81. Their corresponding potential values are 20.43, 20.18, 21.10, 21.27. This reveals the identified community number is 4. Failure to detect other 8 communities is caused by the fact that edges between nodes are comparatively evenly distributed, with no clear core nodes available. Their topology potential ranges from 14.12 to 21.27, difference being insignificant. Therefore, these comparatively loose communities hide themselves in the above 4 bigger communities. Fig. 7 shows the topology potential distribution of nodes in the Football network. There are some minor peaks surrounding the 4 found local maximum potential nodes (*i.e.* Node 1, 25, 59, and 81). The community number estimated by the MSM method is 18. The reason is that the node degree in the Football network is comparatively evenly distributed. It is very difficult to distinguish one core node from others. Furthermore, the threshold of the shortest path length is set at 1 hop, and few of these core nodes can be merged in the same community. The community number estimated by WNDP and $\Omega(c)$ are 7, 9, respectively. For the MIM method, the number 11 has the highest probability.

5.2.5 A wider range of real world networks

We examine algorithm performance on a wider range of real world networks, including Jazz musicians network, C. elegans metabolic network, PGP network, Collaboration network, DBLP collaboration network, and Youtube network. Experiment results on these real world networks are shown in Table 1, including network name, network size, ground truth, estimated number by our method, MSM (the detailed parameter setting can be found in [3]), WNDP [8], $\Omega(c)$ [11] and MIM [16]. Compared with MSM, WNDP and $\Omega(c)$ methods, the estimated number of MIM and our method is close to the ground truth one. Our proposed method shows good performance, especially for networks with high *cc* value.

Table 1. Results of different methods on real word networks.

Network Name	Network Size	Ground Truth	Our Method	MSM	WNDP	$\Omega(c)$	MIM
Zachary network	34	2	2	2	2	2	2
Dolphin network	62	2	2	3	2	2	2
Political book network	105	3	2	2	2	2	3
Football network	115	12	4	18	7	9	11
Jazz musicians network	198	4	4	5	3	9	4
C. elegans metabolic network	453	10	10	9	13	7	11
PGP network	10680	80	76	79	57	105	80
Collaboration network	65276	600	328	1483	828	1356	452
DBLP collaboration network	317080	13477	11983	21729	10249	18724	11142
Youtube online social network	1134890	8385	8267	15149	5738	1246	9358

5.3 Community Number Estimation on Synthetic Networks

5.3.1 Ad hoc network

The generated Ad-Hoc network, with 128 nodes, is split into 4 communities containing 32 nodes each. In the experiment, the parameter *p_in* is set 0.46. The proposed method finds 4 local maximum potential nodes – Node 71, Node 96, Node 98 and Node 126, whose topology potential values are 28.91, 29.11, 28.71 and 28.13, respectively. This indicates there are 4 communities, which is consistent with the true community number of the Ad hoc network. Because the Ad hoc network has a clear community structure, the community number estimated by other four methods is all 4. Table 2 shows the corresponding results.

Table 2. Results of different methods on the Ad hoc network.

Network Name	Network Size	Known Number	Our Method	MSM	WNDP	$\Omega(c)$	MIM
Ad hoc network	128	4	4	4	4	4	4

5.3.2 LFR network

In this subsection, we first produce 7 small-scale synthetic networks. The size of generated networks is 100 nodes. The average degree k is set as 3 or 4. The parameter μ ranges from 0.1 to 0.4, which determines the overlapping probability among communities. Other important parameters are: maxk=15, maxc=50, minc=1. In generally, as μ increases, the community structure of the LFR network becomes ambiguous gradually, making it difficult to carry out community detection.

The proposed method is used to identify the community number of these networks. Results are shown in Table 3. The experiment results confirm the above points. The bigger the μ is, the bigger gap between generated community number and estimated one will be.

Table 3. Results of our method on LFR networks with different μ.

Generating Parameters	Generated Number	Estimated Number by our method
k=4, μ=0.1	3	3
k=4, μ=0.2	5	5
k=4, μ=0.3	4	3
k=4, μ=0.4	5	2
k=3, μ=0.1	6	6
k=3, μ=0.2	6	4
k=3, μ=0.3	6	2

Then, we produce another 8 synthetic networks of various parameters, and compare the proposed method with other four methods on these LFR networks. The generated networks are middle scale with 1000 nodes. The average degree k is set as 10 or 15.

Table 4. Results of different methods on LFR networks.

Generating Parameters	Generated Number	Our Method	MSM	WNDP	$\Omega(c)$	MIM
n=1000, k=10, γ=2, μ=0.1	36	28	12	20	15	27
n=1000, k=15, γ=2, μ=0.1	9	8	9	5	7	8
n=1000, k=10, γ=3, μ=0.1	34	27	23	19	41	29
n=1000, k=15, γ=3, μ=0.1	11	9	15	18	21	14
n=1000, k=10, γ=2, μ=0.2	12	10	15	18	20	11
n=1000, k=15, γ=2, μ=0.2	10	8	13	21	17	8
n=1000, k=10, γ=3, μ=0.2	37	16	22	18	45	28
n=1000, k=15, γ=3, μ=0.2	11	8	14	13	19	8

The exponent of degree distribution γ is set as 2 or 3. In order to generate a clear community structure, a small mixing μ is selected, *i.e.* 0.1 or 0.2. Other important parameters are: maxk=150, minc=100, minc=1, the exponent of community size distribution β=1. Results are shown in Table 4. The proposed method shows well performance on synthetic networks, and the estimated community number is close to the generated one.

5.4 The efficiency of community number estimation

The above community number estimation results on real world and synthetic networks show that the proposed method gives good performance, and MIM is another method with comparative performance. In this subsection, our proposed method is compared with the MIM method from the perspective of the execution time. The ten real world networks in Table 1 are adopted. For each method, 50 times of community number estimation are conducted and the average time is recorded. Table 5 shows the execution time of the two methods on the selected ten networks. Our proposed method outperforms the MIM method in efficiency.

Table 5. The execution time of the two methods on different networks.

Network Name	Network Size	Our Method (s)	MIM (s)
Zachary network	34	0.0023	0.0075
Dolphin network	62	0.0029	0.0093
Political book network	105	0.0074	0.0139
Football network	115	0.0099	0.0317
Jazz musicians network	198	0.0128	0.0326
C. elegans metabolic network	453	0.0404	0.0612
PGP network	10680	3.348	5.161
Collaboration network	65276	8.739	12.824
DBLP collaboration network	317080	57.385	86.479
Youtube online social network	1134890	308.372	489.628

5.5 The Relationship Between Community Number Estimation and *cc* Value

The above experiments imply that the more noticeable the peak-valley structure is, the more accurate the estimated community number will be. According to Definition 3,

cc evaluates the concavity-convexity of the peak-valley structure in the topology potential field. In this subsection, we analyze how the accuracy of community number estimation is related to the *cc* value. For this purpose, we define a community number estimation accuracy measure named EstimationError:

$$EstimationError = \frac{\left|N_{generated} - N_{estimated}\right|}{N_{generated}}. \tag{5}$$

In Formula (5), $N_{generated}$ represents the generated community number, and $N_{estimated}$ denotes the estimated community number.

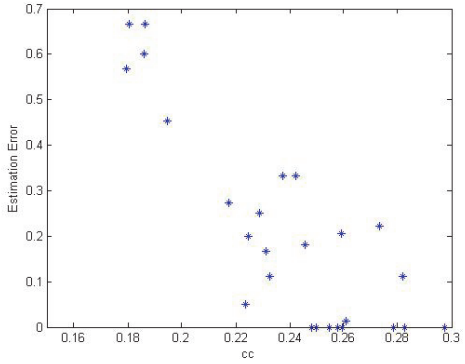

Fig. 8. Relationship between EstimationError and *cc*.

Fig. 8 shows the relationship between *cc* values of the above 26 networks (10 real world networks and 16 synthetic networks) and the EstimationErrors. A big *cc* means that the topology potential value gap between most peak and valley positions is very big, thus the concavity-convexity of the whole peak-valley structure is obvious and clear. The clearer the peak-valley structure of the corresponding networks' topology potential field is, the more precise the estimated community number will be. Contrarily, a small *cc* reveals that the topology potential value gap between most peak and valley positions is very small, thus the concavity-convexity of the whole peak-valley structure is relatively flat and smooth, leading to likelihood of error in local maximum potential node search, and consequentially likelihood of error in community number estimation.

Experimental results also show that, when the peak-valley structure is not very clear, the estimated community number is always smaller than the actual number. The reason is explained as follows. In this case, the local maximum potential nodes of some loose and small-scale communities are in the influence scope of other local maximum potential nodes from large-scale communities. Therefore, these comparatively loose communities will be absorbed by large-scale communities around them, decreasing the estimated community number.

5.6 Discussions

This subsection analyzes the characteristics of our method based on the above experiments results.

- Our proposed method is based on global measure since the topology potential of a node is a composition of topology potential components produced by all the other nodes in the field. Therefore, our proposed method can effectively identify the community number from a global view. In contrast, many other methods are based on node degree in essence. Degree is a typical local measure, which may be less effective sometimes in community number estimation since it neglects the global structure of the network. The flaw will become more notable in large-scale networks. The above experiments results confirm this point.
- Our proposed method only needs one parameter σ, which is unique and determined by the minimum of potential entropy, while other methods are different. For example, MSM method needs three input parameters: λ (the threshold of degree to distinguish core nodes from noncore ones), s (the maximum of available degree range), and th (the threshold of shortest path length). λ is determined by maximizing separability measure. The other two parameters are given based on experience, and different values may result in quite different estimated community numbers.

6. CONCLUSION

Estimating the community number is an important problem of community structure analysis in complex networks. In this paper, we propose a community number estimation method based on topology potential, in which the potential distribution of nodes is utilized. The topology potential field shows a natural peak-valley structure, and each community corresponds to a local high potential area. The number of local maximum potential nodes, located at the center of each local high potential area, is the estimated community number. Experimental results have shown that the estimated community number by our method on many real world networks and artificial networks is closer to the ground truth. Furthermore, our proposed method is not sensitive to the tuned impact factor parameter, and shows good performance in efficiency.

Our method may benefit many applications both theoretically and practically. For example, it can be used in local-expansion-based community detection methods where the seed identification is the most basic and critical step. Just like other topology measurements in complex networks, such as degree, the topology potential can also be used to reflect the differential position and influential ability of each node in the topology [7]. In future, we will continue to exploit the relationship between the estimated community number and the characteristics of the peak-valley structure in the topology potential field. We will also try other methods to determine the proper value of the impact factor more efficiently for handling large-scale networks, such as Maximum Likelihood Estimation.

REFERENCES

1. U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 20, 2008, pp. 172-188.
2. S. Fortunato and M. Barthelemy, "Resolution limit in community detection," in *Proceedings of National Academy of Sciences of USA*, Vol. 104, 2007, pp. 36-41.

3. P. H. Fu, S. N. Zhu, A. Zhu, and X. Dong, "A fast estimation algorithm of community number in large scale-free complex networks," *International Journal of Modern Physics B*, Vol. 28, 2014, ID 1450039.
4. W. Y. Gan, N. He, D. Y. Li, and J. M. Wang, "Community discovery method in networks based on topology potential," *Journal of Software*, Vol. 20, 2009, pp. 2241-2254.
5. X. F. Gong, K. Li, M. H. Li, and C. H. Lai, "A spectral algorithm of community identification," *Europhysics Letters*, Vol. 101, 2013, pp. 73-75.
6. C. H. Guo, J. J. Wang, and Z. Zhang, "Evolutionary community structure discovery in dynamic weighted networks," *Physica A*, Vol. 413, 2014, pp. 565-576.
7. Y. N. Han, D. Y. Li, and T. Wang, "Identifying different community members in complex networks based on topology potential," *Frontiers of Computer Science in China*, Vol. 5, 2011, pp. 87-99.
8. B. He, L. Gu, and X. D. Zhang, "Nodal domain partition and the number of communities in networks," *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2012, 2012, p. 02012.
9. Z. Wang, Y. Zhao, J. Xi, and C. Du. "Fast ranking influential nodes in complex networks using a *k*-shell iteration factor," *Physica A: Statistical Mechanics and its Applications*, Vol. 461, 2016, pp. 171-181.
10. A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, Vol. 80, 2009, ID 016118.
11. Z. F. Li, Y. Q. Hu, Be. S. Xu, Z. R. Di, and Y. Fan, "Detecting the optimal number of communities in complex networks," *Physica A: Statistical Mechanics and its Applications*, Vol. 391, 2012, pp. 1770-1776.
12. H. B. Li, J. P. Zhang, J. Yang, J. B. Bai, and Y. Chu, "Identification of overlapping communities and structural holes between communities based on topological potential," *ACTA Electronica Sinica*, Vol. 42, 2014, pp. 62-69.
13. D. Liu, C. Wang, and Y. Jing, "Estimating the optimal number of communities by cluster analysis," *Internal Journal of Modern Physics B*, Vol. 30, 2016, p. 1650037.
14. U. V. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, Vol. 17, 2007, pp. 395-416.
15. M. E. J. Newman, "Modularity and community structure in networks," in *Proceedings of National Academy of Sciences of USA*, Vol. 103, 2006, pp. 8577-8582.
16. M. E. J. Newman and G. Reinertz, "Estimating the number of communities in a network," *Physical Review Letters*, Vol. 117, 2016, p. 078301.
17. G. Ke, V. L. Orman, and H. Cherifi, "Comparative evaluation of community detection algorithms: a topology approach," *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2012, 2012, p. 08001.
18. E. Weinan, T. Li, and E. Vandeneijnden, "Optimal partition and effective dynamics of complex networks," in *Proceedings of the National Academy of Sciences of USA*, Vol. 105, 2008, pp. 7907-7912.
19. Z. Wang, Y. Zhao, Z. Chen, and Q. Niu, "An improved topology-potential-based community detection algorithm for complex network," *The Scientific World Journal*, Vol. 2014, 2014, pp. 121609.

20. Z. Wang, Z. Li, X. Ding, and J. Tang, "Overlapping community detection algorithm based on node location analysis," *Knowledge-based Systems*, Vol. 105, 2016, pp. 225-235.
21. J. R. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: the state-of-the-art and comparative study," *ACM Computing Surveys*, Vol. 45, 2013, p. 43.
22. J. Zhang, H. Li, J. Yang, J. Bai, L. Zhang, and Y. Chu, "Variable scale network overlapping community identification based on identity uncertainty," *Acta Electronica Sinica*, Vol. 40, 2012, pp. 2512-2518.
23. W. D. Zhou and L. Nakhleh, "Convergent evolution of modularity in metabolic networks through different community structures," *BMC Evolutionary Biology*, Vol. 12, 2012, p. 181.

Zhixiao Wang (王志晓) was born in 1979 and received the Ph.D. degree in the Department of Computer Science and Engineering at Tongji University in 2011. He is an Associate Professor in the School of Computer Science and Technology, China University of Mining Technology. He has published more than 20 papers in international conferences and journals. His research interests include field theory application, and social network analysis.

Jingke Xi (席景科) is an Associate Professor of College of Computer Science and Technology, China University of Mining and Technology. He received his Ph.D. degree at China University of Mining Technology in 2012. His research interests include community detection and data mining.

Yan Xing (邢艳) is a Ph.D. student in School of Computer Science and Technology at China University of Mining and Technology, China. She received her bachelor degree in Computer Science from China University of Mining and Technology in 2010. Her research focuses on data mining, complex network, and community detection.

Zhiguo Hu (胡治国) received B.S. from Artillery College, ShenYang, China, in 2001, M.S. degree from Artillery College in HeFei, China in 2006, and the Ph.D degree in Computer Science from TongJi University, China in 2012. In 2015, he joined the Shanxi University at Taiyuan, where he is currently a Lecturer in the School of Computer and Information Technology. His research interests include voice and video quality measurement, IP network measurement and characterization, network performance monitoring and prediction.