Abstract. Recently, various heterogeneous complex networks featured as “small-world” and “scale-free” have become a common research area of different disciplines. Especially, network topology mining and community detection have become focal topics. Through the investigations of typical features in complex networks, we propose a network nodes evaluation model based on a multivariate hierarchy method. With this model, network core nodes are extracted and a new algorithm about network topology reconstruction is put forward to implementing network backbone topology mining, which provides a new way for data mining and information retrieval. Furthermore, we propose two approaches for network community detection: broken edge clustering and center point diffusing. Experiments show that the methods presented in this paper are of high accuracy with good performance.

Keywords. Complex network, topology mining, community detection, node evaluation, topology reconstruction

AMS (MOS) subject classification: ***

1 Introduction

Complex network exists almost everywhere in the real world, from Internet to WWW, from aerial routes to large electric grids, from very large to inter-personal relationship networks, from cell neuron networks to epidemic processes[1]. Even the parasynonym relationship in linguistics can be analyzed as a complex network problem[7].

However, because the research objects of complex networks are often complicated topology networks consisting of thousands of nodes, analysis is time consuming, if the whole complicated topology network is treated as a process object. Therefore, how to reduce the complex network topology, evaluate the core node, and extract backbone topology efficiently and accurately becomes critical in the research of complex networks. Network topology mining has become a key problem in many disciplines, such as relationship network mining in social complex networks, virus propagation networks[9], etc. Graphic network and social network mining is now employed in the discovery of communities in the hyperlink Internet[6]. In recent years, network community detection becomes a topic along with continuous research on physical and
mathematical network features. Meanwhile, facing the increasing data, people are not satisfied with database query, and they put forward a new requirement: how to extract information and knowledge from data for decision making, that is, how to do data mining. On the other hand, the analysis of complex networks plays an important role in web data mining[2]. At present, node evaluation and topology mining of complex networks are in the ascendant[3], as the work done by Newman[8], where many methods need huge computation resources. Therefore, there remains much work to do in network topology mining and community detection to improve the efficiency and accuracy of network analysis.

This paper is organized as follows: section 2 gives some insights into topology features and makes comparisons of these features; network nodes evaluation, and network backbone topology mining are detailed in section 3; section 4 describes network community detection for which new methods are proposed; section 5 is the main result with discussion of our methods; the main ideas and discoveries are summarized up in section 6.

2 Complex network features

Presently, many models are employed to describe real networks, such as regular graphs, random graph, and small-world and scale-free complex network.

In a regular graph, each node has about the same degree, and the degrees obey $\delta$ distribution. In 1960s, Erdős and Rényi introduced a probability method into graph theory and proposed the ER random graph model. Random networks have small mean distances and convergence coefficients, and their degree distributions obey Poisson distribution: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, $k = 0, 1, 2, ...$ However, random networks theory can not explain the existence of pivotal nodes in real networks. More and more findings show that real networks do not match the supposed random graph model[5]. So networks featured as small-world networks with big convergence coefficient and small mean distance appear[10], as illustrated by Figure 1.

Figure 1 Regular graph to random graph

Moreover, when analyzing real complex network, scientists found that the degree distributions of many complex systems are widely different from Poisson distribution, but obey power law distributions[1]. That is, few core nodes dominate the system, with few nodes having most connections, but most
nodes having few connections. Because the node degree distribution of such a network is \( P(k) = ck^{-\gamma} \), the distribution being unrelated to the network scale, such types of networks are called scale-free network. In 1999, Barabási and Albert proposed a preferential attachment model\[4\], which describes the formation mechanism of scale-free network: growth and preferential connection. Experiments show that with same scale and mean degree, a scale-free network has a smaller mean distance and a bigger convergence coefficient. Judged from mean distance (D) and convergence coefficient (C), random graphs and real complex networks are quite different.

Table 1 Comparison of features between typical networks

<table>
<thead>
<tr>
<th></th>
<th>Random graphs</th>
<th>Real complex networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean distance</td>
<td>Small / Large</td>
<td>Small</td>
</tr>
<tr>
<td>Convergence coefficient</td>
<td>Small</td>
<td>Large (Small-world feature)</td>
</tr>
<tr>
<td>Degree distribution</td>
<td>Binomial / Poisson</td>
<td>Power-law</td>
</tr>
</tbody>
</table>

From regular graphs, random graphs, small-world networks to scale-free networks, we can see that the change tendency of degree distributions in typical networks are from even to uneven, and from uniform to centralization in Figure 2.

Figure 2 Trend of network features with posson to power-law

Based on such features of complex networks, we propose an approach to reduce complexity of large scale complex networks. Most real complex networks are scale-free, that is, their essential features are unrelated to the scales of the networks. Therefore, reduction can be conducted to real complex networks using rational methods in order to facilitate topology mining.

3 Network topology mining

Based on the analysis of network features, the extraction of network core nodes and formation of network backbone topology are implemented. Scale-free effects reveal that the importance of nodes in a real complex network is unequal. Therefore, the way to evaluate the importance of nodes is critical. We propose a new method for the extraction of network core nodes—multivariate hierarchy, and then put forward two network topology mining methods: topology reconstruction and data field virtual core.
3.1 Core nodes extraction

Based on the major features discussed above—small mean distance, big convergence coefficient and power law distribution, multivariate hierarchy is proposed on complex network core nodes extraction.

3.1.1 First Layer: multivariate weight method

The large amount of nodes in a complex network makes the dilemma of efficiency and accuracy in the research of network topology. Taking the overall features of complex networks into consideration, a multivariate weight method is of high efficiency and accuracy. The algorithm is described as follows:

**Input:** for network $G(V, E)$, $V$ stands for the node set, $E$ stands for the edge set; $N$ is the number of nodes, $N_1$ is the number of core nodes, and $R$ stands for the values of nodes, for instance, the related degree on mining demand, the corresponding quality, etc.

**Output:** primary network core topology

**Procedure:**

Step 1: Compute the degree of $G(V, E)$, $De$

Step 2: Compute the inner-node weight, $W_I = R/MAX(R) \times De/MAX(De)$ (MAX() is a function for the maximal value)

Initialize the outer-node eight, $W_O = 0$; $\forall u, v \in V, (u, v) \in E, W_O(u) = W_O(u) + W_I(v)$; $W_O(v) = W_O(v) + W_I(u)$; $W_O = W_O/MAX(W_O)$ (SUM() is a function for the sum, the time complexity is $O(N)$)

Step 3: Compute the convergence coefficient of $G(V, E)$, $C$

Step 4: Set the proportion of inner weight $p$, outer weight proportion $q$, the combined weight of nodes, $W = p \times W_I + q \times W_O + (1 - p - q) \times (1 - C)$; select $N_1$ nodes with highest weight as primary core nodes $V_1$, $V_1 = CORE(W, V, N_1)$, $CORE()$ is an extraction function

Step 5: Produce primary from the $N_1$ nodes and their topology relationship: $G_1(V_1, E_1) : \{V_1 = \{\forall v \in V_1\}, E_1 = \{\forall u, v \in V_1, (u, v) \in E\}$

3.1.2 Second Layer: topology quality extraction

Network $G(V, E)$ is reduced to primary core topology $G_1(V_1, E_1)$ by using the multivariate weight method with time complexity $O(N)$. And it makes good preparation for further backbone topology mining.

This method combines betweenness, degree, minimum mean distance and two-dimensional cloud extract core nodes from primary core topology. The algorithm is described as follows:

**Input:** primary core topology $G_1(V_1, E_1)$

**Output:** second level core nodes $V_2$

**Procedure:**

Step 1: Retain nodes and the relationship that connects with $V_1$, extend primary core topology $G'_1(V'_1, E'_1) : \{V'_1 = V_1 + \{v | \forall v \in V, u \in V_1, (v, u) \in E\}, E'_1 = E_1 + \{(u, v) | u \in V_1, v \in V'_1, (u, v) \in E\}$
Step 2: Initialize betweenness $B=0$; compute degree $D_e$ according to $G'_1(V'_1,E'_1)$; compute the shortest distance $D(u,v)$ and route $P(u,v)$ between any two nodes: $\forall u, v \in V'_1, (u,v) \in E'_1$ if $w \in V_1, w \in P(u,v)$, $B(w) = B(w) + 1$; end if; define mean distance: $\forall i \in V_1, D_i = \sum_{j=1}^{N_1} D(i,j) / N_1$

Step 3: $\forall v \in V_1, E_B = \max(B), E_D = \max(D)$, compute the topology quality of node $v$: $M(v) = W(v) \times \exp(-B(v) / E_B - D(v) / E_D) / \sqrt{\frac{E_B}{3} \times \frac{E_D}{3}}$

Step 4: Determine the number of nodes, $N_2$, select $N_2$ nodes according to $M$ values as the secondary nodes $V_2$. $V_2 = \text{CORE}(M, V_1, N_2)$.

3.1.3 Third Layer: Data field virtual core method

This method employs the shortest distance $D$, betweenness $B$ computed in the first level, and topology quality $M$, secondary core topology, to produce network virtual core. The algorithm is described as follows:

Input: $D, B, M$ and $V_2$

Output: network virtual core nodes $V_3$

Procedure:

Step 1: Update $D, B, M$: $D = \{D(v) | \forall v \in V_2\}, B = \{B(v) | \forall v \in V_2\}, M = \{M(v) | \forall v \in V_2\}$

Step 2: Define a topology potential function:

$S(v) = \sum_{i=1}^{N_2} M(u) \times \exp(-\frac{D(u,v)}{E_B}) | u = V_2(i)$

Step 3: Determine the number of core nodes $N_3$, select $N_3$ nodes according to $S$ values as core nodes in virtual core: $V_3 = \text{Core}(S, V_2, N_3)$.

Potential function $S(V)$ defined above represents interrelationship among nodes. The impact of nodes is in inverse proportion to their distance. Each node in the topology space will affect other nodes, and is affected by other nodes, which form a data filed. The potential of any point in the data field is regarded as the sum of potentials in this point. Taking efficiency into consideration, we adopt the potential function to form data filed in the final core nodes extraction. This method, which represents the overall impact among nodes, performs well in accuracy but is bad in efficiency (time complexity $T = O(n^2)$). However, through the former two reductions, the scale of the network decreases sharply, which makes the computation feasible.

3.2 Network backbone topology mining

Based on the network core nodes extraction method–multivariate hierarchy, we propose a new method–topology reconstruction to build original network backbone topology using core nodes. The algorithm is described as follows:

Input: first level topology $G_1(V_1, E_1)$, core nodes $V_1, V_2, V_3$

Output: reduced network topology

Procedure: Set level of topology mining $R$: $R=1, 2$ or $3$

(1) if $R=1$, return $G_1(V_1, E_1)$. 

Complex network topology mining and community detection
(2) if R=2, extend \( G_1(V_1, E_1) \) to \( G'_1(V'_1, E'_1) \): \( V'_1 = V_1 + \{ v | \forall v \in V, u \in V_1, (v, u) \in E \}; \\
E'_1 = E_1 + \{ (u, v) | u \in V_1, v \in V'_1, (u, v) \in E \}; \)
\( \forall u, c \in V_2, \) if \( (u, c) \in E \), link \( (u, c) \); else calculate \( \text{Path}(u, c) \) in \( G'_1(V'_1, E'_1) \), if \( \forall k \in V_2, k \neq u, k \neq c, k \notin P(u, c) \), link \( (u, c) \) → \( E_2 \), return \( G_2(V_2, E_2) \).
(3) if R=3, \( \forall u, c \in V_3 \), if \( (u, c) \in E \), link \( (u, c) \); else calculate path of \( (u, c) \) in \( G_2(V_2, E_2) \), get \( P(u, c) \); if \( k \in V_3, k \neq u, k \neq c, k \notin P(u, c) \), link \( (u, c) \) → \( E_3 \), return \( G_3(V_3, E_3) \).

This method outputs a reduction of complex network topology, which keeps most of the major characteristics of the original network. Therefore, it facilitates further research of complex networks.

4 Network community detection

Section 3 proposes three solutions for complex network backbone topology mining. From the point of detail and microcosmic, we should decompose complex network in order to conduct mining on nodes community, which can discover a detailed relationship between nodes. In this paper, we propose two methods for complex network community detection: broken edge clustering and center point diffusing.

4.1 Broken edge clustering

Input: network \( G(V, E) \), where \( V \) is the node set, \( N \) is the number of node, \( E \) stands for the edge set and \( u, v, w, x \) refer to nodes in the network.
Output: communities in network \( G \).

Procedure:
Step 1: \( \forall w, x \in V, B(w, x) = 0 \), set the number of community \( m \), community counter \( t=0 \);
Step 2: Compute the shortest distance \( D(u, v) \) and shortest route \( P(u, v) \) of any nodes pair \( (u, v) \in G(V, E) \); \( \forall x \in V, \) if \( E(w, x) \in P(u, v) \), \( B(w, x) = B(w, x) + 1 \);
Step 3: Select edge \( E(i, j) = \text{MAX}(B) \), break \( E(i, j) \) from \( G(V, E) \); \( E(i, j) \notin E, B(i, j) = \text{REALMIN}; \) recompute the shortest distance \( D(i, j) \); if \( D(i, j) \geq \text{REALMAX} \), then \( t = t + 1 \); \( C(t, 1) = i; C(t, 2) = j \); Here, \( ct \) refers to community counter, \( C \) counts marginal nodes among communities; \( \text{REALMIN} \) and \( \text{REALMAX} \) stand for the minimum and maximum distances of the network;
Step 4: Repeat Step 3 until community counter \( t = M \)

4.2 Center point diffusing

The algorithm is described as follows:
Input: network \( G(V, E) \), where \( V \) is the node set, \( N \) is the number of nodes and \( E \) stands for the edge set; select core nodes set \( C \), number of core nodes \( P, C \subseteq V, u, v \in V, c, w \in C \)
Output: communities in network \( G \)
Procedure:

Step 1: Set the number of community $M$, generally $M \leq P$; if $M > P$, then increase the number of nodes to make $P > M$;

Step 2: $\forall v \in V$, compute the shortest distance of $v$ to each core node $c \in C$, make the core node with the shortest distance as $v$’s parent node, that is, $F(v) = \{c|D(v, c) = \max(D(v, C))\}$ produce a community with core nodes being parent nodes $Q(c) = \{v|F(v) = c\}$, and set community counter $m = P$

Step 3: Select the minimum subcommunity in current community $Q$: $Q(c) = \min(size(Q))$, update the current core nodes set $C' = \{w|w \in C, w \neq c\}$, compute the distance $D(c, C')$ between parent node $c$ and all other core nodes; make the node with shortest distance as its attachment neighbor $Nc = \{w|D(c, w) = \min(D(c, C'))\}$, then select the maximum scale sub-community as attachment community of $Q(c)$; update the new community $Q(w) = \{Q(w), Q(c)\}$; community counter $m = m - 1$;

Step 4: Repeat Step 3 until $m = M$

5 Results and discussion

In order to show the advantages of the algorithms proposed in this paper, we choose a typical complex network $G(V, E)$ with node number $N=53$, and finally an UML class diagram with more nodes as examples. In $G(V, E)$, nodes number $N=53$, edges number $E=56$, core nodes $M=10$, the topology and degree distribution diagram are shown in Figure 3.

![Figure 3: Topology of $G(V,E)$ and degree distribution](image)

The results of $G(V, E)$ on multivariate hierarchy method is shown in Figure 4.

![Figure 4 Core nodes extraction and backbone mining($R$ from 1 to 3)](image)
In order to further verify the significance of the algorithms proposed in this paper, we employ a real complex network UML class diagram topology $U(V, E)$ with $N=224, E=580$ as follows.

Figure 5 Topology of $U(V, E)$ and degree distribution

Figure 6 Core nodes extraction $M=50$

Figure 7 Backbone of $U(V, E)$

The results above show that the first-level nodes extraction performs well with high efficiency and accuracy. If efficiency is not the primary concern, the second-level topology extraction will get higher accuracy. The results show that topology reconstruction can represent the original network topology with good performance.

Meanwhile, the results of network community detection of $G(V, E)$ using broken-edge clustering and center point diffusion are shown in Figures 8 and 9. We adopt center point diffusion in the community detection of $U(V, E)$ with community number $M=10$, and the result is significant as Figure 10.

Figure 8 Broken edge clustering

Figure 9 Center point diffusing
We found that although broken edge clustering performs better than center point diffusion in accuracy, the latter is a better choice for efficiency and representing preferential attachment in complex networks.

6 Conclusions

Through the analysis of complex network features—small-world, scale-free, etc, his paper has proposed a new method for evaluation and extraction of complex network core nodes—multivariate hierarchy method. Then, a novel method based on the extracted core nodes—network topology reconstruction has been proposed to conduct topology mining. We have also proposed two methods for network community detection—broken edge clustering and center point diffusing. Experiments have shown that our methods perform well on complex network topology mining and community detection in both efficiency and accuracy.

The innovative aspects of this paper are as follows: first, we hold a macroscopic view of complex network features from even to uneven, from uniform to centralization; second, the multivariate hierarchy method takes all critical factors—degree, betweenness, convergence coefficient and mean distance into consideration; third, community detection methods proposed in this paper deal with efficiency and accuracy quite elegantly.

7 Acknowledgments

The authors thank Peng Liang for his professional English Teaching and wonderful suggestions and Ph.D Bin Liu, Jin Liu and Shu-Liang Wang for providing data and information for the network simulation. This work was funded in part by the National Natural Science Foundation of China under grant 60496323. This work was also supported by the Provincial Natural Science Foundation of Hubei under Grant 2005ABA123 and 2005ABA240.
8 References