ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20190712024843
Original scientific papers

Hierarchical Semantic Community Detection in Information Networks: A Complete Information Graph Approach

Guilan SHEN, Jie SUN, Yaohui HAO

Abstract: In order to detect the hierarchical semantic community which is helpful to discover the true organization of information network,we propose a complete information graph approach. In this method, we first use complete information graphs including semantic edges and link edges to represent information networks. Then we define semantic modularity as an objective function, a measure that can express not only the tightness of links, but also the consistency of content. Next, we improve Lovain's algorithm and propose simLV algorithm to detect communities on the complete information graph. This recursive algorithm itself can discover semantic communities of different sizes in the process of execution. Experiment results show the hierarchical community detected by the simLV algorithm performs better than the Louvain in measuring the consistency of semantic content for our approach takes into account the content attributes of nodes, which are neglected by many other methods. It can detect more meaningful community structures with consistent content and tight structure in information networks such as social networks, citation networks, web networks, etc., which is helpful to the application of information dissemination analysis, topic detection, public opinion detection, etc.

Keywords: complete information graph; content attributes; information network; semantic hierarchical community

1 INTRODUCTION

Community is one of the important features of complex networks, and the nodes in the real network often belong to different hierarchical structures. That is to say, a large community may contain a small community, and a small community may contain some smaller community structures. A node can belong to multiple communities at the same time. Fig. 1 shows us a network with a hierarchical community structure. It is very meaningful to analyze the network hierarchical community, which can be helpful to detect the central organization of the network, better understand the phenomena in the network [1], provide the representation forms of different granularity for the system represented by the network, and comprehensively reveal the hidden rules of the network.

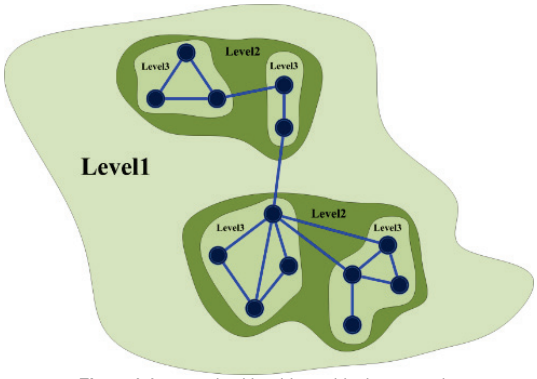

Figure1 A network with a hierarchical community

In fact, the study of hierarchical community reinforces the concept of community, for it performs a hierarchical analysis of the original detected communities. We usually abstract many real networks into complex network represented by simple graph, which only focus on the network structure. As a result, most existing researches naturally represent the internal strength of the community as the tightness of links, and the purpose of hierarchical community detection is to find the community structure with different tightness of links. However, we know that for different types of networks, the purpose of community detection may be different, and the measurement of the internal strength of the community should be different.

Information network is a complex network which nodes have content attributes, such as social network, web network, science citation network and so on. It is important to consider the content when analyzing these networks. For example, it is more practical to identify tightly linked nodes with the same interests and hobbies in the social network, which can be used for precision marketing. Therefore, in addition to focusing on structure attributes, content attributes should also be considered in hierarchical community detection of information community networks. The internal strength of community should have the dual characteristics of link tightness and content consistency. Although many researchers have paid attention to the significance of combining network structure and node content attribute to detect community, some detection methods of semantic community [2-4] have also been developed, there are still few researches on hierarchical semantic community.

In this paper, we propose a complete information graph method to detect semantic hierarchical communities in information networks. The following contents in this paper include: the second section is literature review and related work introduction; the third section introduces the hierarchical community detection method based on complete information graph; the fourth section is the experiment part and the fifth section is the conclusion.

2 RELATED WORKS

In the past 10 years, lots of methods have been developed to detect the hierarchical structure of the networks. These methods can be summarized as follows.

2.1 Methods Based on Generating Tree Graph

Tree graph is a classical method to describe hierarchical structure. To reveal multiple levels of network, a tree graph can be generated by some approach, and then different methods are used in the tree graph to obtain multiple cut values or resolution thresholds. SalesPardo [5]

adopts the top-down method to detect the community. After measuring the similarity between nodes according to their closeness, this method uses the block box to infer the network hierarchy according to the similarity between nodes. The method itself can directly reveal the network hierarchy. Vieira [6] defines the distance between communities in the network by taking modularity as the community quality metric, and then generated tree graph with spectral method to reveal the hierarchical community. There are some other network clustering methods that can generate hierarchical tree graph, but there is no good way to determine the cut threshold to divide tree graph.

2.2 Methods Based on Multi-Resolution Parameters

The most obvious feature of hierarchical communities is multi-resolution structure. Using different resolutions to describe the community has been the main method for a period of time [7-9]. In general, such methods are based on multi-scale quality functions according to the real organizational structure. By adding a resolution parameter to the quality function, the module size of the community can be adjusted on the basis of optimizing the partition. Based on the assumption that in the network, network flows will stay in the tightly linked community for a long time, Rosvall [10]，Renaud [11] and Delvenne [12] use the length of time that the flow stays in the network as the standard of partition quality and the time consumed by markov random walk in the network as the measure. Such methods take the time as the resolution adjustment parameter, and as time increases, they can reveal various organizations of different sizes in the system.

The biggest limitation of these methods is how to choose the appropriate resolution parameters. Moreover, even when the resolution parameters are fixed, the multi-resolution quality function is as limited as the modularity function.

2.3 Methods Based on Local Optimization of Community Quality

This kind of methods usually adopts greedy search strategy to optimize the local maximum of community quality, in which hierarchical communities of different scales can be found in the network. For example, the Louvain method [13] takes the modularity as the community quality function and recursively performs the optimization in a multi-scale form. Mutilevel Infomap method [14] is based on the network flow and information theory, and converts the problem of how to detect the communities into the problem of how to compress maximally the information coding of nodes in the network, that is, how to minimize the total length of information encoding of nodes in the network. In order to solve this optimization problem, this method defines the hierarchical Map Equation of multilevel information compression as the objective function and adopts the algorithm idea similar to Louvain method to find the hierarchical community.

It has been shown that it is meaningful to divide intermediate communities in the process of local community quality maximization. The advantages of these methods are that they are fast and do not need to adjust resolution parameters, but they lack theoretical basis. Moreover, even if the system is not multi-scale, or even random network graph, they can also generate hierarchical structure.

2.4 Methods Based on Probabilistic Model

Such methods treat the network structure as a probabilistic process of building edges among groups of nodes and then identify the most likely clustered groups. Clauset [15] directly uses the tree random graph to represent the hierarchical structure of the network, and then infer a group of tree random graphs that could better represent the hierarchical structure of the network by using the maximum likelihood estimation. According to these random tree graphs, the hierarchical structure of the network was obtained. Peter Ronhovde [16] uses the porter model to accurately quantify the hierarchical or multi-resolution structure in the graph. Tiago P [17] constructs a nested generation model, which can completely describe the whole network hierarchy on multiple scales, and this method can also avoid resolution problems caused by the detection method based on modularity. Based on the principle of simplification, these methods can avoid noise even if the resolution is increased, and there will be no miscalculation in the sparse network. But the method of probability model usually has high complexity and is not suitable for large scale network.

In general, researchers have put forward many corresponding methods to detect the hierarchical communities. These methods can find the community structure consistent with the actual system level organization, that is, the sub-communities of different sizes and scales nested in the large community. Although the target functions of detection are different, none of these methods consider the node content attributes, which is not suitable for detecting meaningful communities for information network. That is to say, the detection of these hierarchical communities does not focus on the inherent requirements of semantic hierarchical communities with tight structure and consistent semantics.

3 RESEARCH METHOD

3.1 Complete Information Graph

Simple graph is the common representation of the real networks, here, nodes represent individuals, and edges represent the links between individuals. However, this classical representation method has some limitations in dealing with information networks which nodes have content attribute. For the information network, the connection between nodes is reflected in two aspects, one is the direct link relationship, and the other is the semantic relationship caused by the similarity of node content. However, the simple graph cannot show the semantic relation. For example, there is no citation relationship between literature A and literature B in the citation network, although they are similar in content, the semantic similarity relationship cannot be directly represented by the edge in the simple graph. To reflect the content relationship between nodes, we propose the concept of complete information graph.

Definition Complete Information Graph Let $G = \{V, E\}$ be the simple graph of information network, then $CG =$

$\{V, E'\}$ is called the complete information graph of information network, where $\forall e = (u, v) \in E'$, if $e \in E$, then e is called the linked edge of the complete information graph CG, otherwise, if $e \in E' - E$, then e is called the semantic edge of the complete information graph CG.

Obviously, a complete information graph CG can represent the two relationships between all nodes in the information network in terms of structure and content. Let's look at an example, as shown in figure 2, node 6 has no linked edges with node 1, node 2 and node 3, but they have semantic consistency, so the semantic edges in the complete information graph shown in figure 3 reflect this relationship.

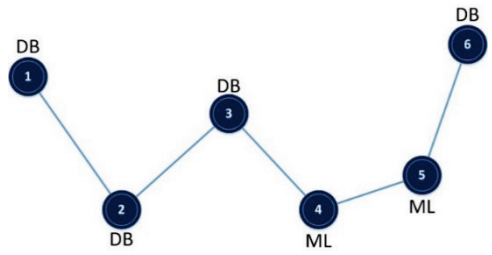

Figure 2 simple graph of an information network

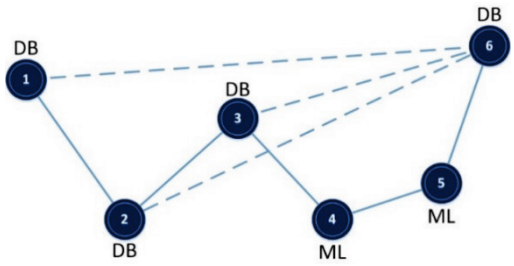

Figure 3 complete information graph of an information network

We can use complete information graph to represent all type information network. In different types of information network, different methods can be adopted to build semantic edge. For the purpose of generality, function $fun(sim(u, v))$ is defined in our paper. When the following conditions are met, the two nodes u and v without linked edges can build semantic edges.

$$fun(sim(u, v)) = \begin{cases} 1 & \exists e = (u, v) \in E' - E \\ 0 & otherwise \end{cases} \quad (1)$$

where $sim(u, v)$, is the content similarity of node u and v. The content similarity can be calculated by using cosine similarity, KL divergence or Pearson correlation coefficient, etc. which decide by the detailed representation of node content.

Depending on the different type, size and application scenarios of the information network, the function fun can adopt three different strategies:

Strategy 1: threshold method. This strategy sets a similarity threshold γ which is used to build semantic edges for two nodes with no linked edges whose content similarity is greater than the threshold.

Strategy 2: top N method. This strategy presets the number of semantic edge N, and then selects two node pairs with content similarity within the top N and without linked edges to build semantic edges.

Strategy 3: KNN method. This strategy selects the average degree value K of the simple graph of information network, and for each node, selects K most similar nodes without link edges to build semantic edges.

3.2 The Structure and Content Fusion Approach

Most existing methods of integrating structure attribute and content attribute are based on the premise that the nodes in the network have linked edges. According to the content attribute of nodes, the corresponding content similarity calculation method is adopted, and the content similarity of node pairs is taken as the edge weight, for example, Ester proposed a fusion model for CkC problem [18]. However, such methods do not solve the key problem of how to take advantage of content similarity between the unlinked edge nodes in the simple graph.

In a complete information graph, link edge directly reflects the structural relationship between nodes, while semantic edge reflects the potential semantic relationship between nodes. If the two are separated which simply considers the content attribute of nodes or the link information of nodes, it is inevitable to miss some core information to measure the close relationship between nodes. Based on the idea of transforming nodes' similarity into edge weights, we convert nodes' structure similarity and content similarity into edge weights in complete information graphs. In general, let the content similarity of two nodes u and v be $sim_c(u, v)$, and the structural similarity be $sim_s(u, v)$, then the similarity of nodes converted to edge weights is expressed as shown in formula 2

$$w_{uv} = sim(u, v) = \alpha \cdot sim_c(u, v) + (1 - \alpha) \cdot sim_s(u, v) \quad (2)$$

Here, α is the parameter for adjusting the proportion of content similarity and structure similarity. It is between 0 and 1.

As mentioned above, content similarity can be measured in different forms according to the modeling method of node content attributes. In our paper, the text vector space model is adopted to represent the node contents, which are represented as weight vectors. Let the document set composed of all node contents be D, and $V = \{t_1, t_2, \ldots, t_{|v|}\}$ is a group of different words, that is the glossary of document data set, then the content attribute of each node u can be expressed as a word vector $content_u = (w_{1u}, w_{2u}, \ldots, w_{|V|u})$, and each weight w_{iu} can be calculated by word reverse document frequency $tf\text{-}idf$, which is shown as formula 3

$$w_{iu} = tf_{iu} \times idf_i = \frac{f_{iu}}{\max\{f_{1u}, f_{2u}, \ldots, f_{|V|u}\}} \times \log \frac{N}{df_i} \quad (3)$$

where N is the number of nodes in the information network, df_i is the number of documents containing at least one word t_i, and f_{iu} is the number of t_i times that the word appears in the content of the node.

Here, the content similarity of two nodes u and v is calculated by using the Angle cosine similarity between vectors.

$$sim_c(u,v) = \cos(content_u, content_v) = \frac{\sum_{i=1}^{|V|} w_{iu} \times w_{iv}}{\sqrt{\sum_{i=1}^{|V|} w_{iu}^2} \times \sqrt{\sum_{i=1}^{|V|} w_{iv}^2}} \quad (4)$$

In order to calculate the structural similarity of nodes, we extend the classic ternary closure principle in social network analysis to the general information network. That is to say, we believe that the more common neighbors two nodes have, the more similar the two nodes are. For example, two scientists with a common collaborator in the cooperative network of scientists are more likely to cooperate in the future [19]. Based on this principle, Jaccard index of common neighbor is adopted to measure the structural similarity of two nodes. This method is only based on local information and can avoid excessive computational complexity.

For node u in the network, its neighbor set is defined as $\Gamma(u)$, then the Jaccard structural similarity of two nodes u and v is defined as

$$sim_s(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \quad (5)$$

3.3 Hierarchical Semantic Community Detection Method

Louvain algorithm proposed by Blondel et al. is an aggregation algorithm for hierarchical community structure analysis, which can be applied to networks with millions of nodes and has the characteristics of fast speed and high accuracy. Because the optimization goal of the algorithm is modularity proposed by Newman, the algorithm can find the hierarchical communities with high link tightness. In order to find the hierarchical communities with tight links and consistent semantics in the information network, this paper gives the definition of semantic modularity on the basis of the modularity. Semantic modularity is essentially a multiplicative model integrating modularity and content similarity. Given the complete graph CG of the information network with n nodes and m edges, the definition of modularity is shown in formula 6, then the definition of semantic modularity is shown in formula 7

$$Q = \frac{1}{2m} \sum_{uv} \left(G_{uv} - \frac{k_u k_v}{2m} \right) \delta(C_u, C_v) \quad (6)$$

$$Q_{sim} = \frac{1}{2\sum_m sim(u,v)} \times \times \sum_{uv} \left(G_{uv} - \frac{k_u k_v}{2\sum_m sim(u,v)} \right) sim(u,v) \delta(C_u, C_v) \quad (7)$$

where $\sum_m w_{uv}$ is the sum of weights of edges in the complete graph CG, $sim(u, v)$ is the similarity of node u and node v in the network, $\sum_{i \in \Gamma(u)} sim(u, i)$ is the sum of the similarity of node u and all its neighbors, and can also be regarded as the link density of node u, is the expected weight corresponding to the similarity between node u and node v in the zero model.

This semantic modularity not only reflects the closeness of community nodes, but also considers the semantic similarity between nodes. This multiplicative model avoids adjusting parameters when measuring the structural and semantic characteristics of communities. In this paper, the semantic modularity is taken as the optimization objective, and our proposed simLV algorithm similar to the Louvain is applied in the complete graph of the information network to explore the hierarchical structure of the information network. The algorithm is divided into two stages.

The first phase is community initialization, also known as coarsening phase. At first, we assign a community number to each node in the network, that is, each node is considered a separate community. Then, for any node u and v, ΔQ_{sim} is the increment of modularity of the corresponding semantic community. When the node u joins the community c where the neighbor node v is located. When ΔQ_{sim} is positive, the neighbor node with the corresponding ΔQ_{sim} maximum value is selected and the node u is added to the community where the neighbor node is located. If all of ΔQ_{sim} are negative, node u remains in the original community. Repeat the above consolidation process until the entire network is stable and no more consolidation occurs, then the smallest level of communities is divided.

In the second stage, using the results of the first stage to construct a new network, the network nodes are the first stage of communities, the weight of connecting edges between nodes is the total weight of all connecting edges between two communities. Then, the community division of the new network is carried out with the algorithm of the first stage, and the community structure of the second smallest level is obtained.

Repeat the process until a higher level of community structure is no longer possible, Thus, a hierarchical semantic community structure is detected.

4 EXPERIMENTS

4.1 Experimental Evaluation Metrics

In experiments, data sets usually lack prior knowledge, so it is not possible to effectively determine whether hierarchical communities are valid or not. In order to reasonably measure the effect of the hierarchical semantic community detection, we evaluate the semantic community quality from two perspectives including the overall level and each sublevel.

We selected three metrics to evaluate the quality of hierarchical communities, includes semantic modularity Q_{sim}, normalized mutual information NMI [20] and $Purity$ of community. When the prior knowledge of community classification is unknown, the quality of hierarchical communities can be evaluated with semantic modularity. When the prior knowledge of community classification is available, it can be evaluated with NMI and $Purity$.

Given the standard communities $G = \{G_1, G_2, \ldots, G_S\}$ the communities detected by the algorithms is represented by $C = \{C_1, C_2, \ldots, C_S\}$. To evaluate the consistency in topics, the $Purity$ proposed by Strehl A etc. [23] is employed. The $Purity$ of C_i is defined as:

$$Purity(C_i) = \frac{1}{|C_i|}\max_j \{C_i \cap G_j\} \quad (8)$$

Usually, the detected community C includes nodes that belong to other G in the ground-truth. For C, we compute the intersection set with each standard community G_j, then take the maximum as the result for it. So the *Purity* of C is defined as:

$$Purity(C) = \frac{1}{K}\sum_{i=1}^{K} Purity(C_i) \quad (9)$$

The average *Purity* of the detected communities is measured by the average *Purity* of each community. The higher *Purity* means that results are closer to the ground-truths.

Normalized mutual information *NMI* is defined as

$$I_{norm}(X:Y) = \frac{H(X)+H(Y)-H(X,Y)}{(H(X)+H(Y))/2} \quad (10)$$

where, $H(X)$ is the information entropy of the random variable X associated with the generated partition C, and $H(Y)$ is the information entropy of the random variable Y associated with the real partition G. $H(X, Y)$ is joint entropy. The value of mutual information $I_$norm $(X:Y)$ is normalized to [0, 1], where 1 indicates that the generated community is completely consistent with the standard community, and 0 indicates that the generated community is completely unrelated to the standard community.

4.2 Datasets

We select three real datasets, including web information network Wisconsin, and two science citation networks CiteSeer and Cora. For simplicity, we handle all networks formed by these datasets as undirected network. The statistical information of specific datasets is shown in Tab. 1.

Table 1 The statistical information of datasets

Dataset	|*Class*|	|*V*|	|*Edges*|	*Average Degree*
Wisconsin	5	262	459	3.50
CiteSeer	6	3312	4536	2.38
Cora	7	2708	5429	5.70

4.3 Effect Analysis about Links and Content Fusion in Complete Information Graph

To verify the effect of the complete information graph on merging content attributes and links attributes, we first test what kind of node pairs need to build semantic edges nto the original network structure to form a complete information graph. Then we design and test the following four strategies to verify which method of setting edge weight the complete information graph is the best.

(1) Based on node content similarity: This strategy calculates the cosine vaule of each pair nodes content as the edge weight of the graph, which is represented by the symbol *S* in the experiment.

(2) Based on node link structure: This strategy sets the edge weight of all connected edges, including link edges and semantic edges, as 1, which is represented by the symbol *T* in the experiment.

(3) Based on node structural similarity：This strategy calculates the Jaccard value of each pair nodes structure as the edge weight, which is represented by the symbol *J*.

(4) Based on the linear combination of node content similarity and structure similarity: This strategy calculates the content similarity and structure similarity of all nodes in the complete information graph, and converts the calculated values into the edge weights of the complete information graph by means of weighted linear combination, which is represented by the symbol *H*. The weight value was set to 0.5 in our experiment for convenience, and content similarity and structure similarity were regarded as equally important.

In the experiment, we select randomly the Citeseer data set to construct a complete information graph using the similarity threshold method. Here, we calculate the similarity of a pair of nodes by using the cosines of vectors. Firstly, we take different γ content similarity value between [0.3, 0.8] to build content edges to form different complete information graphs and then apply different strategies into them to verify the quality of the detected communities.

The evaluation results of the first layer of the hierarchical community detected are shown in Fig. 4.

Experimental results show that the community quality detected in the complete information graph increases gradually with the increase of γ. When the threshold value reaches 0.7, the effect is the best; however, when the threshold value exceeds 0.7, the community quality detected decreases slightly. This phenomenon is caused by the fact that the vaule similarity threshold is higher, the quality of content edge is better, which plays the role of meaningful edge connection to the original information network, in addition, the fewer semantic edges can be supplemented. However, the number of semantic edges is too small, although it can improve the quality of semantic community detection, but it cannot achieve the best detection results. Therefore, the threshold γ can be set to be slightly less than the maximum similarity value in the data set.

Experimental results also show that adopting the fourth strategy, namely, the content similarity and structural similarity of the nodes of linear fusion, has a better overall community detection effect than the other three strategies, and the number of detected communities was comparable to those based on node content similarity and node link structure. Therefore, the method we proposed is effective in detecting semantic community

It is worth mentioning that we observed a special phenomenon. When we use the method of structural similarity to transform the edge weight of complete information graph, the two performance indexes of semantic community detected in semantic modularity and purity are very prominent. But it does not mean this method for detecting the semantics community quality is good. Due to the sparsity of the Citeseer data set itself, a large number of scattered and fragmented small communities were detected by this method, and the number of communities with the smallest number detected in different similarity thresholds reached 2112. It can be seen that this method has poor effects in sparse networks.

4.4 Effect Analysis on Hierarchical Community Detection

To evaluate the effectiveness of the proposed method, we compare our method with the baseline method Louvain in the complete information graph. At first, we build complete information graphs for three data sets respectively. According to the experimental conclusions in the previous section, three datasets, including Wisconsin, Citeseer and Cora set the similarity thresholds of semantic edges as 0.5,0.7 and 0.5, respectively. Next, the edge weights in complete information graphs are calculated by the method based on the linear combination of node content similarity and structure similarity. Then we use the proposed algorithm simLV and the classic Louvain algorithm in the complete information graph of each dataset to detect the hierarchical community, and use the *NMI*, *SimQ* and *Purity* metrics to quantify the performance of each algorithm. The results for three datasets are shown in Tab. 2 to Tab. 4.

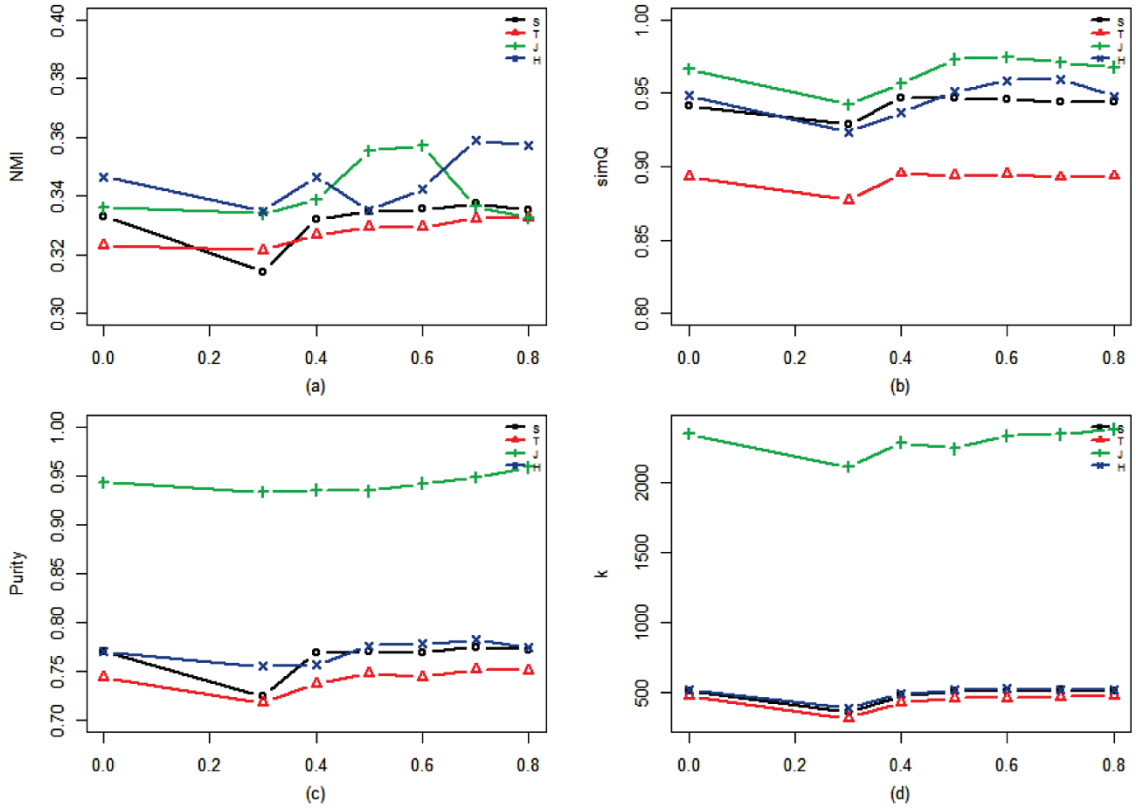

Figure 4 Selection analysis of content similarity threshold

Table 2 Comparison results of hierarchical communities of Wisconsin dataset

Level	Louvain				SimLV			
	k	*NMI*	*Purity*	*simQ*	*k*	*NMI*	*Puriy*	*simQ*
3	45	0.232	0.647	0.277	67	**0.286**	**0.705**	**0.302**
2	15	0.225	0.598	0.351	23	**0.247**	**0.673**	**0.358**
1	10	0.202	0.581	0.376	10	**0.214**	**0.597**	**0.389**

Table 3 Comparison results of hierarchical communities of Citeseer dataset

Level	Louvain				SimLV			
	k	*NMI*	*Purity*	*simQ*	*k*	*NMI*	*Puriy*	*simQ*
4	1045	0.337	0.825	0.633	1158	**0.371**	**0.877**	**0.786**
3	531	0.306	0.737	0.796	615	**0.365**	**0.816**	**0.847**
2	435	0.289	0.693	0.832	546	**0.356**	**0.789**	**0.936**
1	420	0.291	0.687	0.833	523	**0.359**	**0.783**	**0.960**

Table 4 Comparison results of hierarchical communities of Cora dataset

Level	Louvain				SimLV			
	k	*NMI*	*Purity*	*simQ*	*k*	*NMI*	*Puriy*	*simQ*
4	278	0.441	0.833	0.756	305	**0.439**	**0.894**	**0.789**
3	202	0.430	0.802	0.771	226	**0.436**	**0.836**	**0.865**
2	123	0.427	0.781	0.795	145	**0.425**	**0.808**	**0.893**
1	95	0.428	0.768	0.816	108	**0.420**	**0.783**	**0.901**

The proposed algorithm simLV and Louvain algorithm both can find hierarchical community, and for the principle of the two algorithms is the same, the number of levels detected in the community for the same data set is the same. In different levels of community quality detection, firstly, in terms of the number of communities detected, because simlv algorithm considers semantic consistency and easily destroys the original tight connection structure, it detects more than Louvain algorithm at all levels. Secondly, in the aspect of purity measurement, it is superior to Louvain algorithm in each data set, which indicates that the algorithm has good effect in the semantic consistency of detected communities at all levels. In addition, in terms of semantic modularity measurement, the algorithm is also superior to the Louvain algorithm in all data sets, which indicates that the algorithm has a good effect in detecting link tightness and semantic consistency.

In brief, the experimental results show that the proposed method can detect hierarchical semantic communities better.

5 CONCLUSION

In community detection of information networks, nodes with similar content but no link edges are difficult to be classified as the same community. In view of this phenomenon, we propose the concept of a complete information graph which merges the linked edges and the

semantic edges. Specifically, on the basis of the original network graph, it adds semantic edges to the nodes without linked edges but with similar semantic content through relevant strategies, and converts the linear combination of content similarity and structural similarity of nodes into the edge weight of complete information graph for hierarchical community detection.

With the proposed semantic modularity as the objective function, we adopt simLV algorithm, which is similar to Louvain algorithm, to carry out the recursive optimization of local semantic modularity. In the process of optimization, hierarchical communities are found by using the method itself, and there is no need to adjust the resolution parameters. The feasibility and effectiveness of the proposed algorithm are verified by real datasets.

Due to the limitations of the experiment, there is no effective verification of the consistency level between the detected hierarchical community and the real community, which needs further research in the future.

Acknowledgments

This paper is supported by Scientific Research Project of Beijing Educational Committee (KM201711417004), Premium Funding Project for Academic Human Resources Development in Beijing Union University (BPHR2019CZ03), Project of philosophy and Social Sciences in Henan Province (2018BXW007).

6 REFERENCES

[1] Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *NATURE, 453*(7191), 98-101. https://doi.org/10.1038/nature06830

[2] Cao, J., Jin, D., Yang, L., & Dang, J. (2018). Incorporating network structure with node contents for community detection on large networks using deep learning. *Neurocomputing*, S0925231218300985. https://doi.org/10.1016/j.neucom.2018.01.065

[3] Fan, X., Xu, R. Y. D., Cao, L., & Song, Y. (2017). Learning nonparametric relational models by conjugately incorporating node information in a network. *IEEE Transactions on Cybernetics, 47*(3), 589-599. https://doi.org/10.1109/TCYB.2016.2521376

[4] Combe, D., Largeron, C., Egyed-Zsigmond, E., & Géry, M. (2012). Combining Relations and Text in Scientific Network Clustering. *International Conference on Advances in Social Networks Analysis & Mining. IEEE Computer Society*. https://doi.org/10.1109/ASONAM.2012.215

[5] Salespardo, M., Guimerà, R., Moreira, A. A., & Amaral, L. A. N. (2007). Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences of the United States of America, 104*(39), 15224-15229. https://doi.org/10.1073/pnas.0703740104

[6] da Vieira, F. V., Xavier, C. R., Ebecken, N. F. F., & Evsukoff, A. G. (2014). Modularity based hierarchical community detection in networks. *Computational Science and Its Applications – ICCSA 2014*. Springer International Publishing. https://doi.org/10.1007/978-3-319-09153-2_11

[7] Reichardt, J. & Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters, 93*(21), 218701. https://doi.org/10.1103/PhysRevLett.93.218701

[8] Reichardt, J. & Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E, 74*(1), 016110. https://doi.org/10.1103/PhysRevE.74.016110

[9] Arenas, A., Fernández, A., & Gómez, S. (2008). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics, 10*(5), 053039. https://doi.org/10.1088/1367-2630/10/5/053039

[10] Rosvall, M. & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences, 105*(4), 1118-1123. https://doi.org/10.1073/pnas.0706851105

[11] Lambiotte, R. (2010). Multi-scale Modularity in Complex Networks. *International Symposium on Modeling & Optimization in Mobile. IEEE*.

[12] Delvenne, J. C., Yaliraki, S. N., & Newman, B. M. (2010). Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences of the United States of America, 107*(29), 12755-12760. https://doi.org/10.1073/pnas.0903215107

[13] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), 0-0. https://doi.org/10.1088/1742-5468/2008/10/P10008

[14] Rosvall, M. & Bergstrom, C. T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PLOS ONE*, 6. https://doi.org/10.1371/journal.pone.0018209

[15] Clauset, A., Moore, C., & Newman, M. E. J. (2008). Hierarchical structure and the prediction of missing links in networks. *NATURE, 453*(7191), 98-101. https://doi.org/10.1038/nature06830

[16] Ronhovde, P. & Nussinov, Z. (2009). Multiresolution community detection for megascale networks by information-based replica correlations. *Physical Review E, 80*(1), 016109. https://doi.org/10.1103/PhysRevE.80.016109

[17] Peixoto, T. P. (2013). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X, 4*(1), 346-354. https://doi.org/10.1103/PhysRevX.4.011047

[18] Ge, R., Ester, M., Gao, B. J., Hu, Z., Bhattacharya, B., & Ben-Moshe, B. (2008). Joint cluster analysis of attribute data and relationship data. *ACM Transactions on Knowledge Discovery from Data, 2*(2), 1-35. https://doi.org/10.1145/1376815.1376816

[19] Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Working Papers, 64*(2), 025102. https://doi.org/10.1103/PhysRevE.64.025102

Contact information:

Guilan SHEN, PhD, Assoicate Professor
Corresponding author
Beijing Union University,
A3, Yanjingdongli, Chaoyang district, Beijing,100025, China
E-mail: guilan.shen@buu.edu.cn

Jie SUN, PhD, Assoicate Professor
Beijing Union University,
A3, Yanjingdongli, Chaoyang district, Beijing,100025, China
E-mail: jie.sun@buu.edu.cn

Yaohui HAO, PhD, Lecturer
State Key Laboratory of Mathematical Engineering and Advanced Computing,
62 Science Avenue, Zhengzhou City, Henan Province 450001, China
E-mail: hao_yaohui@126.com