Journal of Theoretical and Applied Information Technology
20th April 2013. Vol. 50 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

LOSSLESS NETWORK COMPRESSION BASED ON TOPOLOGY POTENTIAL COMMUNITY DISCOVERY

WANG XUHUI
Handan College, Handan 056005 China

ABSTRACT

A research of lossless network compression is carried out. To meet the different needs, two approaches of lossless network compression are proposed in this research. One approach, judging importance of the nodes according to their roles playing in the community composition, quantifies the importance of every node in communities, and achieves lossless network compression through layers; another approach, judging importance of the nodes according to the distances from the community representative nodes to them, differentiates the nodes with different distances, and achieves lossless network compression through compression ratio. Comparative experiments show that the two approaches not only can achieve perfect compression ratio, and retain the relationship between the communities, but also can reserve the important nodes or basic community structures during the compression process according to the needs.

Keywords: *Topology Potential; Topology Potential Entropy; Uncertainty Measure; Overlapping Community Discovery; Variable Scale Community; Structural Holes Between Communities; Lossless Network Compression*

1. INTRODUCTION

With the development of network society era, in-depth study is the inevitable demand of the times put on social networks. Visualization of social networks, such as knowledge discovery research should relate to the social network compression. With the increasing size of social network, community detection has become an indispensable step [1] social network application process. The community as a social network structure characteristics, in the process of compression retains its important nodes or basic structure and maintains the relationship between them has important meaning and value. However, the existing compression methods from view, community research object are still very rare compression.

In view of the above questions, in the discovery method based on topological potential network community based on, this chapter launches the compression method of network, puts forward two kinds of lossless compression method of network. The first method called social network compression algorithm SNC (Social Network Compression), is essentially a graph compression method. But in order to distinguish from other Internet community for the consideration of the premise, the method is called compression methods for social network. The method will be the first to use topological potential theory of the social network community discovery and the importance of distinguishing between community node, then the node importance level compression based on community. Second methods of lossless compression method of social network will still topology potential theory community discovery and quantified the node importance based on community, then according to the compression rate of the network compression. Hereinafter referred to as the method for NSNC (New SNC) method.

Although the above two methods that based on the social network are proposed, because of the social network also belong to the complex network (although the two are indeed has some differences, such as nodes in social networks are generally known as actors, more emphasis on initiative node), and complex network also exists in a community structure [2], therefore these two methods are fully applicable to the complicated network compression.

2. LOSSLESS COMPRESSION METHOD FOR SNC SOCIAL NETWORK

Because the SNC method will be found in the topological potential community, network compression according to the importance of community nodes, so this section will first to topological potential community discovery nodes found in the analysis of the importance of community.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.1 Analysis of community node importance

From the community structure level, the application of topology discovery based on potential theory of community importance nodes found in the community is the existence of difference. In order to illustrate the importance of different nodes and the conclusion that community, here are the following theorem and its corollary.

Theorem 1 let u、v nodes in a network of community representatives, v^* a attract chain, a and u in the v jump, v v^* at the $a+1$ jump, $a=0,1,2,...,h-1$, u、v, v^* topology potential contribution ratio

$$R_{u\leftarrow v}(a,a+1) =. e^{\frac{2a+1}{\sigma_{opt}^2}} \quad (1)$$

Shown by the formula (1) shows an arbitrary attract node p chain on the community representative topology potential contribution of v^* in

$$A_{v^*\leftarrow p}(\sigma_{opt},l)=\frac{1}{n}e^{-(\frac{l}{\sigma_{opt}})^2} \quad (2)$$

The l p left the minimum hop number v^*.

By type, namely u, v topology potential contribution to the amount of v^* respectively.

$$A_{v^*\leftarrow u}(\sigma_{opt},a)=\frac{1}{n}e^{-(\frac{a}{\sigma_{opt}})^2}$$

and

$$A_{v^*\leftarrow v}(\sigma_{opt},a+1)=\frac{1}{n}e^{-(\frac{a+1}{\sigma_{opt}})^2}$$

The contribution of the two ratio

$$R_{u\leftarrow v}(a,a+1)=\frac{A_{v^*\leftarrow u}(\sigma_{opt},a)}{A_{v^*\leftarrow v}(\sigma_{opt},a+1)}= e^{\frac{2a+1}{\sigma_{opt}^2}}$$

Table 1 lists the number of network distance between node hops contribution ratio of representative points, which can be used to verify the correctness of the above theorem and its corollary. From Table 1, we found in the HCD community, neighbor nodes than the neighbor node contribution represent topological potential larger; with the representative point distance increasing, the contribution of exponential decline node. Therefore, with the local extremism point to its nearest neighbor node representing the community number more, the relationship between them is also more closely and form the core structure of the community; topological potential neighbor nodes that represent the contribution is relatively small, the number is relatively less, interaction among them is more sparse. To sum up, from the community composition level, neighbor nodes represent points compared to the neighbor node is more important.

Table 1 $R_{u\leftarrow v}(a,a+1)$ values of several networks

Node u leave the representative point v* hop a	Node u leave the representative point v* hop a+1	Karate club (σopt=1.0204)	Dolphin Society (σopt=1.1782)	Word adjacencies (σopt=1.0043)	Les miserables (σopt=1.0435)	Books about US politics (σopt=0.9803)
1	2	17.8365	8.6810	19.5772	15.7225	22.6869
2	3	121.7630	36.6679	142.2059	98.6741	181.8129
3	4	831.2309	154.8820	1.0330e+003	619.2763	1.4571e+003

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2The basic idea of SNC method

We can know from the previous analysis, the importance of the neighbor node points found in the HCD community is a hop-by-hop reduced. Accordingly, SNC will use the first method based on topological potential theory of community found, then the network compression steps to effectively reduce the size of the network.

The SNC method uses relatively representative point compression from outside to inside the way, most can be compressed to the network only representative points. One of the advantages of this method is: in the compression process can not only compress off some nodes are less important to reduce the scale of the network effectively, but also retains the basic structure of the community important node or community where necessary.

Different compression methods in general graphs, SNC method without user specified parameters in the compression process, and the optimal range of influence is determined according to the method of automatically only under the guidance of H specified to be compressed to hop. Figure 1 shows a optimal range of influence of 2 jump community compression diagram, one-way arrows represent one from a jump can be compressed to a jump.

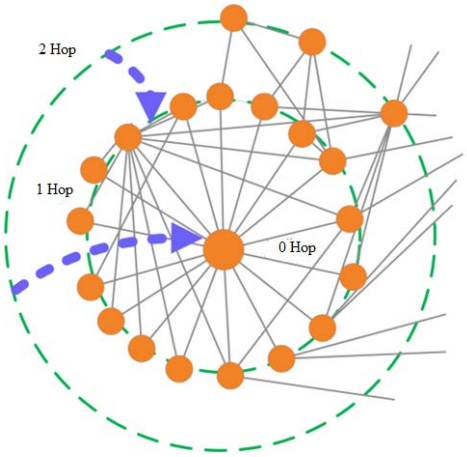


Fig. 1 *Community compression schematic diagram*

The basic idea of SNC method is first to topological potential community discovery method of community discovery, and then compress. According to this idea, the proposed SNC method consists of two parts. The first part of the paper and third chapters of the overlapping community discovery algorithm based on greedy strategy are similar to the GS algorithm for community detection. The algorithm considers not only overcome community edge nodes and other nodes linked by the small number of community nodes overlap artificially separated problems and overcome the shortcomings of HCD method, but also consider the offer ready and support for the following network compression. To avoid ambiguity, hereinafter referred to as the algorithm for the NGS (New Greedy Strategy). The second part puts forward the importance of a compression algorithm according to the node level network, referred to as BL (Based on Layers).

2.3 The SNC methods of basic data structure

In order to realize lossless compression, first of all to the data structure design of SNC method. In addition to some basic data structure, data structure design must also take into account the following factors: SNC discovery process is required for all node hops distance representative points in the community, at the same time in the compression process needs to keep the relationship between community. Based on the above considerations, designed the data structure of SNC method.

2.4 Describe SNC method

NGS algorithm in determining the community on behalf of all the nodes along with the greedy strategy to attract chain periodicity is representative point of attraction. It is described as follows:

Method name: algorithm of NGS discovery based on greedy strategy network overlapping communities

Algorithm input: networkG=(V,E)(|V|=n , |E|=m)

The output of the algorithm: Ci(I community representative point number)

The BL algorithm is obtained by interactive way to hop to the compression, and then compressed operation. The detailed description of BL algorithm is as follows:

Method name: according to the node importance of network level compression algorithm based on BL

Algorithm input: network G=(V,E)(|V|=n , |E|=m)optimization of the impact factor of OptSigma

Algorithm output: compression community Ci(I as the representative point number, each Ci displays only the user specified number of nodes within the)

Journal of Theoretical and Applied Information Technology
20th April 2013. Vol. 50 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.5Analysis of SNC methods for complex

The SNC method is mainly used for NGS algorithm and BL algorithm for community detection network compression. In comparison, the time complexity of the NGS algorithm to the high number of. Because the NGS algorithm in the worst case of no more than $O(n^2)$。$(n$ is the number of nodes in the network), so the SNC method of the time complexity does not exceed $O(n^2)$。

3 . SNC EXPERIMENTS AND ANALYSIS

The proposed method is feasible and effective, through the experiment in the karate club network [3] and dolphin social network [4] two widely used data sets to test method.

3.1 Compression test of dolphins of the social network

Application of NGS algorithm for community detection in dolphin social network, square icon and the star icon were used to mark the three different communities, large icon is used to mark the representative community, and triangle icon is used to mark the overlapping nodes community. Icon in figure 2~4

Found in the community on NGS algorithm, BL algorithm is applied to 2, 1 and 0 hops to compress the discovered communities, the compression results are shown in figure2~4. Bidirectional arrows in figure 2~4 is used to mark the two communities of the relationship.

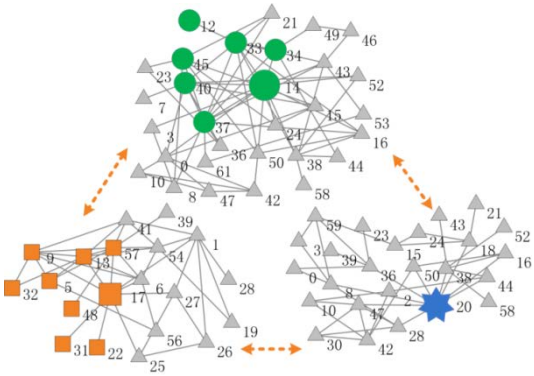


Fig. 2 *Two hops compression on dolphin social network*

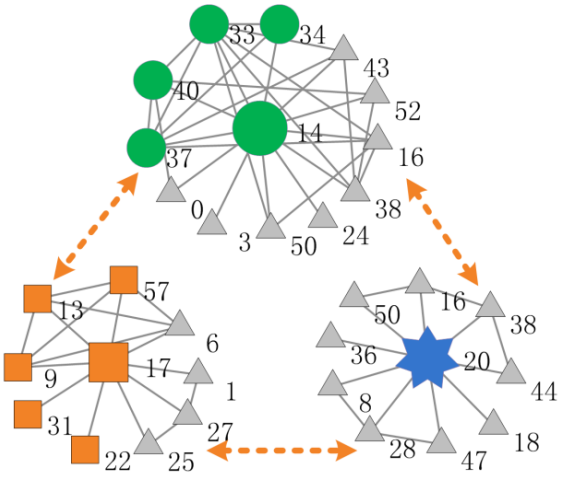


Fig. 3 *One hop compression on dolphin social network*

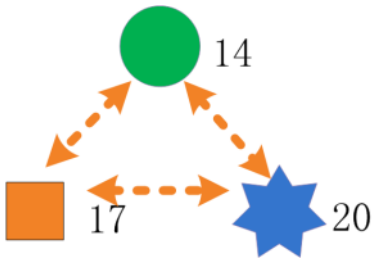


Fig. 4 *Zero hops compression on dolphin social network*

3.3Experimental analysis

Show the classic experiments on the data sets, using overlapping community discovery algorithm greedy strategy based on NGS and the importance of network node level compression algorithm BL, the number of nodes in the community can be efficiently compressed. Table 2 lists the various community karate club network and the social network in 2, 1 and 0 hops compression rate data (numbers and community representatives in the second column is the number of the community to maintain consistent). Compression ratio defined here as follows:

The definition of 1 if the $G^{'}=(V^{'},E^{'})$ network is a network of $G=(V,E)$ compression, $R=\frac{|G|-|G^{'}|}{|G|}$ compression rate is called network G.

Due to the optimization of two network range is h = 2, so it may lead to some nodes in the community was unable to attract the nodes in other communities, and the compression ratio is 0, such as the karate club network community C1 in

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

compression to the 2 jump compression ratio that belongs to this kind of situation. In general, in the optimization of some nodes in the community will still have the ability to attract other community nodes, so the compression rate of karate club network of community C34 and dolphin social networks in the community C14, C17 and C20 in the compressed to 2 jump is not 0, the highest compression ratio can reach 0.4314. Compared with the methods of literature [5-6] identified in the community, in the compression to the 1 jump after the community or still maintained the basic structure or preserves important node, while the maximum compression rate of 0.75, the lowest was 0.2917. In the compression to 0 jump, each community compression rate reached the highest, in more than 0.95.

Table 2 Community compression rate list

Network name	Community name	The community of node number	Compression to hop		
			2	1	0
Karate club network	C1	24	0	0.2917	0.9583
	C34	27	0.2222	0.3333	0.9630
Dolphin social network	C14	51	0.4314	0.7451	0.9811
	C17	23	0.1304	0.5652	0.9565
	C20	40	0.3750	0.7500	0.9750

4 . THERE ARE SNC APPROACH TO THE PROBLEM

The current map compression method has higher time complexity, dependence on a priori knowledge of parameter setting, parameter is adjusted too much, compression loss, [4] ignore the network community structure. Aiming at these problems, puts forward the community the importance of nodes social network compression method based on SNC. The method to realize network community for the compression, compression according to the importance of the node level, in the process of compression can choose whether to retain important node community or basic structure according to the need, and can maintain the relationship between community. However, the SNC method also exists the problem cannot control the compression ratio. To solve this problem, the following will present a new lossless compression method of NSNC social network.

5 . LOSSLESS COMPRESSION METHOD FOR NSNC SOCIAL NETWORK

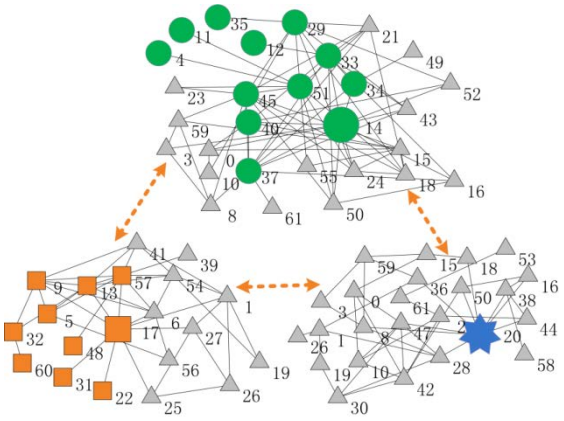
NSNC method has two algorithms. The first application based on the IS community node important degree sorting algorithms belongs to the uncertainty in the third chapter, quantitative characterization of the importance of community node, followed by going to proposed algorithm according to the compression ratio on the network compression. To illustrate the convenience, hereinafter referred to as the compression algorithm for the NSNC method for BIV (Based on Importance values). Because the NSNC method directly using IS algorithm, so this section of the IS algorithm no longer.

Differences between the SNC and NSNC node importance. Although the NSNC method and the SNC method use topology potential theory to judge the importance of community nodes, but the nodes importance of the two methods in the terms of the difference between them. The SNC method relates distance between nodes and community representative points as the basis for judging the importance of node importance, different distance between nodes. Importance of the node in the SNC method is a kind of distance on level of distinction, with hierarchical integrity. NSNC method to nodes in the community structure in the role as the basis to judge the importance of node, realize the importance of quantization on each node.

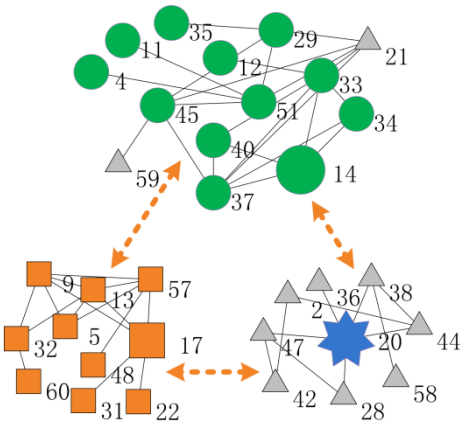
Analysis of NSNC algorithm time complexity. The NSNC method involves two main aspects of the operation, one is to sort the nodes in network community, and another is designated by the ratio of compression to the community. The NSNC method in the nodes are sorted in the community, the data structure and algorithm of the basic inheritance of the IS algorithm, the time complexity is less than knlnn (k is a constant, n is the number of nodes in the network); in a specified ratio of R compression, just according to the node importance ranking after a temporary

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

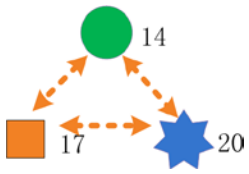
filter, time complexity is less than O (n). Therefore, the NSNC method of time complexity is still knlnn.



(a)R_{14}= 0.4314， R_{17}=0.1304， R_{20}=0.375



(b)R_{14}= 0.7451， R_{17}=0.5652， R_{20}=0.75



(c)R_{14}= R_{17}= R_{20}=0.99

Fig. 5 NSNC compression on dolphin social network

6 . CONCLUSIONS

With the development of image compression method and technology in the semantic label network, network retrieval application in many fields has been used more and more widely, the related research has been paid more and more attention. In view of the existing compression methods in higher time complexity, dependence on a priori knowledge of parameter setting, parameter is adjusted too much compression problem, loss and ignore network community structure, expand the research of lossless network compression. First of all, determine the lossless compression method of social network SNC community the importance of nodes based on topological potential community method. Node importance found in the proposed topology potential method and the community related theorems and inferences based on, firstly, the NGS algorithm based on greedy strategy for community discovery and mining community in the different levels of importance of node, then compress the community through social network compression algorithm SNC based on the node importance. The feasibility and effectiveness of the method is verified by the classic experiments on the data sets. The experimental results show that, this method can not only in the compression process keep the relationship between community, but also has the ideal community compression rate, up to 0.95 or more, and can retain the basic structure of the important nodes or the community in need. Secondly, in view of the existing network compression method in SNC cannot flexibly control the compression ratio, the addition of a lossless compression method for NSNC network. In the NSNC method, using the topological potential community method to determine community node importantance degree. Experiments show that, compared with the method of SNC, the NSNC method can not only realize the compression effect is equivalent with the compression ratio, but also can specify arbitrary.

REFERENCES

[1] Tian Y, Hankins R A, Patel J M. Efficient aggregation for graph summarization. Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2008: 567-579P

[2] Hauguel S, Zhai Chengxiang, Han Jiawei. Parallel PathFinder algorithms for mining structures from graphs. 2009 Ninth IEEE International Conference on Data Mining. Miami: Institute of Electrical and Electronics Engineers Inc., 2009: 812-817P

[3] Navlakha S, Rastogi R, Shrivastava N. Graph summarization with bounded error. 2008 ACM SIGMOD International Conference on Management of Data. New York: Association for Computing Machinery, 2008: 419-432P

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[4] Zhang Ning, Tian Yuanyuan, Patel J M. Discovery-driven graph summarization. 26th IEEE International Conference on Data Engineering. Long Beach: IEEE Computer Society, 2010: 880-891P

[5] Toivonen H, Zhou Fang, Hartikainen A, et al. Compression of weighted graphs. The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. San Diego: Association for Computing Machinery, 2011: 965-973P

[6] Toivonen H, Mahler S, Zhou F. A frame work for path-oriented network simplification. Advances in Intelligent Data Analysis IX, 2010, 6065(2010): 220-231P