A New Method for Identifying Detected Communities Based on Graph Substructure

Shumei Kameyama[1] Makoto Uchida[1] Susumu Shirayama[2]
[1]*School of Engineering, the University of Tokyo* [2]*RACE, the University of Tokyo*
E-mail shumei@race.u-tokyo.ac.jp uchida@race.u-tokyo.ac.jp sirayama@race.u-tokyo.ac.jp

Abstract

Many methods have been developed that can detect community structures in complex networks. The detection methods can be classified into three groups based on their characteristic properties. In this study, the inherent features of the detection methods were used to develop a method that identifies communities extracted using a given community detection method. Initially, a common detection method is used to divide a network into communities. The communities are then identified using another detection method from a different class. In this paper, the community structures are first extracted from a network using the method proposed by Newman and Girvan. The extracted communities are then identified using the proposed detection method that is an extension of the vertex similarity method proposed by Leicht et al. The proposed method was used to identify communities in a blog network (blogosphere) and in a Wikipedia word network.

1. Introduction

A large number of clustering methods have been developed and implemented for various purposes in many different fields. For example, in complex network analysis, clustering methods have been used to analyze different social and biological networks using pattern extraction and recognition [1]. It has been shown that groups of similar vertices in a network can be used to characterize the network itself. In this field, detection and extraction of patterns or clusters from a network are often referred to as community detection. Although the community is sometimes connected to a real community in the case of social networks, in other networks, the community is often treated as a subgraph.

Many methods for detecting community structures in complex networks have been developed. The detection methods can be broadly classified into three groups:

(i) Methods that are based on the structural similarity of the global or local graph;
(ii) Methods that are based on the criterion where the vertices within the communities have a higher density of edges than vertices between communities;
(iii) Methods that are based on the eigenspace of the adjacent matrix or Laplacian matrix of the graph.

These methods can be viewed as different approaches that use different topological methods to determine the same result: the communities in a network. These methods all have the drawback that certain parameters, such as the number of communities, must be determined before hand. Despite this drawback, it has been proven that a network can still be successfully subdivided into its communities. However, these communities may not precisely delimit the different communities actually found in the network. In order to perform this task, another procedure is required; that is, a different property must be used to identify the communities. In many cases, a possible choice is to use the semantics of a community as a unique property. For example, with the TF-IDF technique, an attempt was made to characterize the communities detected in a blogosphere [2]. Specific topics discussed in each community were found, and it was shown that the communities can be identified by such specific topics. Thus, it can be considered that semantics is a useful property for identifying a community. However, to extract a structure using semantics in a network is complicated. Therefore, a different characterization method is required.

It has been shown that the information obtained from a single-stage detection (extraction) is insufficient to express the whole community structure of the network. This implies that another property is required to characterize the networks based on their topological structures. One possible approach is to use a different class of detection methods to delineate the communities present in a given network. Using this assumption, a new method is proposed to identify communities that had already been detected using a

0-7695-3028-1/07 $25.00 © 2007 IEEE
DOI 10.1109/WI-IATW.2007.102

given community detection method. The inherent features of a detection method are used to identify each community. This proposed method is then, applied to identify the communities in two real data sets: a blog network (*blogosphere*) and a Wikipedia word network. It will be shown that each community can be identified by the pattern of clusters that appears in the communities.

2. Proposed method

Each community will have different characteristic properties depending on the type of detection method that is being used. This inherent feature of the detection methods is used to identify each community. Each community that has been detected using a certain method is further subdivided into subcommunities based on a different type of detection method. The subdivision pattern is used to characterize each community. Thus, it must be determined which method is suitable for each detection stage. All of the methods are not suitable for huge networks because of their time complexity.

First, based on the time complexity of some of the methods for large networks, a detection method with a small time complexity was required. The best method that was found was the detection method first proposed by Newman and Girvan [3][4], which belongs to class (ii). This method was used for the first stage of detection. Second, the detection method for the second subdivision was chosen from either class (i) or (iii). In this paper, a class (i) method was used. Most class (i) methods, such as CONCOR [5], are based on the structural similarity of the graph. HITS [6] can divide a graph into groups of similar vertices. However, due to the presence of arbitrary constants in the grouping process and the inability to use graph distance as a distinguishing parameter, there are few methods that can directly be used to extract a specific pattern from the communities. Although there are many vertices with a small graph distance in each community detected using the Newman and Grivan method, the distribution of graph distances is not uniform. Therefore, it may be assumed that graph distance plays a key role in extracting a specific pattern.

Leicht et al. proposed an algorithm for quantifying the amount of similarity between vertices [7]. Similarity between two arbitrary vertices can be evaluated iteratively using only a knowledge of the adjacency matrix of the network. The difference in graph distance among the vertices is reflected in the similarity of vertices. This method is the best algorithm that uses structural equivalence. However, this method was not extended to detect communities in a network. Thus, this paper presents a method that can be used to extend this method. An overview of the proposed method is as follows:

1) Divide a network into some communities using the community detection method by Newman and Girvan.
2) All communities corresponding to subgraphs of the network are subdivided into some clusters using the proposed similarity-matrix-clustering method.
3) Each community is identified by the pattern of the clusters obtained in 2).

Figure 1 shows a schematic of the proposed method.

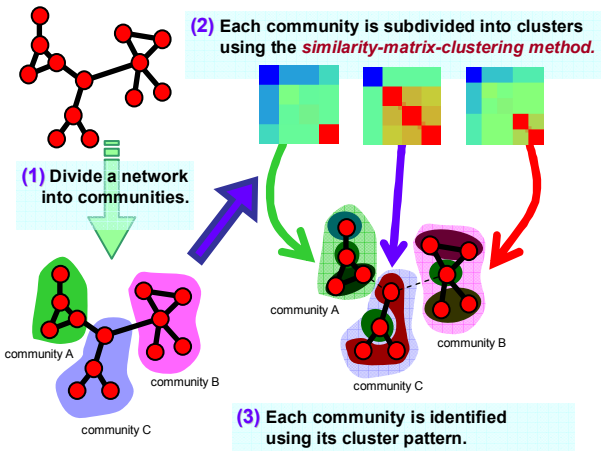


Figure 1. Schematic of the proposed method

2.1 Similarity-Matrix-Clustering (SMC)

The similarity-matrix-clustering method is a method that clusters the vertices by re-arranging the rows and columns of the similarity matrix **S**, whose component $S_{i,j}$ represent the similarity between the vertices i and j. This idea of using vertex similarity was proposed by Leicht et al. [7]. The algorithm can be described as follows.

First, obtain the similarity matrix **S**, which is defined as,

$$\mathbf{S} = \varphi\mathbf{AS} + \psi\mathbf{I}. \qquad (1)$$

where **A** is the adjacency matrix, **I** is the identity matrix, and φ and ψ are some constants. The matrix has size N x N, where N is the number of vertices. Matrix **S** is computed using the following iterative procedure [7]:

$$\mathbf{DSD} = (\alpha / \lambda_1)\mathbf{A}(\mathbf{DSD}) + \mathbf{I}, \qquad (2)$$

where **D** is a matrix whose components are given by $D_{i,j}$: $D_{i,j} = k_i \delta_{ij}$, k_i is the degree of vertex i, λ_1 is the largest eigenvalue of **A**, and δ_{ij} is the Kronecker's

delta. Let **B** represent **DSD**. The iterative procedure can be rewritten as

$$\mathbf{B}^{n+1} = (\alpha / \lambda_1)\mathbf{A}\mathbf{B}^n + \mathbf{I}, \quad \mathbf{B}^1 = \mathbf{I}, \qquad (3)$$

where n denotes the iteration step. Equation (3) is repeated until $| \mathbf{B}^{n+1} - \mathbf{B}^n | < \varepsilon$ is satisfied or n equals $nmax$, where ε is a small positive number and $nmax$ is the maximum number of iterations. The similarity matrix **S** can be obtained by

$$\mathbf{S} = \mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1}. \qquad (4)$$

After the similarity matrix is obtained, the components of the matrix are re-arranged to determine the clustering of the vertices. Let s_i be the row sum of the similar matrix, defined as $s_i = \sum_j S_{i,j}$. Let $r_{i,j}$ be the squared sum of the differences between the i^{th} row and the j^{th} row ($j > i$), defined as

$$r_{i,j} = \sum_{n \neq i, n \neq j} (\mathbf{S}_{i,n} - \mathbf{S}_{j,n})^2 \qquad (5)$$

The re-arrangement procedure consists of the following three steps:

1. Exchange the first row of **S** with the row that has the smallest row sum. Augment by 1 the value of i.
2. For the i^{th} row, compute $r_{i,j}$ ($j > i$). Find the row k which makes $r_{i,k}$ the smallest, and exchange the $(i + 1)^{th}$ row with the k^{th} row.
3. Set i equal to $I + 1$, and repeat step 2 until i equals n -1.

Thus, the transformed similarity matrix **S'** is computed. Figure 2 shows that similar rows in **S'** are located adjacent to each other. The vertical 3-dimensional bars show the value of each component of the matrices.

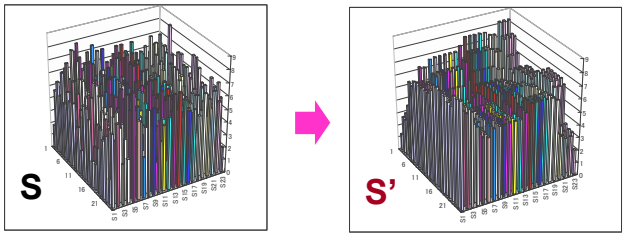


Figure 2. Transformation of the similarity matrix

The components of the i^{th} row vector of **S'** represent the similarity between the i^{th} vertex and the other vertices. The closeness of two adjacent rows of **S'** are used to determine the amount of clustering. Let the gap G be the square sum of the two adjacent rows of **S'**. The gap is defined as

$$G_j = \sum_n (S'_{j,n} - S'_{j+1,n})^2 . \qquad (6)$$

The deviation of the gap (GV) is calculated by

$$GV_j = 10\frac{G_j - \bar{G}}{\sigma_G} + 50, \quad j = 1, \cdots, N-1, , \qquad (7)$$

where $\bar{G}$ is the average of the gap distance and σ_G is the standard deviation. The vertices are clustered by dividing the two adjacent rows of **S'**. First, GV is ranked in descending order. Secondly, a GV_k is found such that $GV_k > Gt$, where Gt is a certain threshold value. Finally, the vertices are divided into two groups based on GV_k.

2.2 Characteristics of SMC

Using SMC, the clustering of vertices is first demonstrated. The network to be clustered is shown in Figure 3 (left). Figure 3 (right) shows the magnitude of the values in Matrix **S'**. The color red represents highest intensity, while blue represents the lowest intensity. The numbers along the left side of the grid represent the vertex number. For Gt = 60, 3 GV_k were obtained. The largest GV_k is GV_9. Vertices are divided into two groups: (1, 2, 3, 4, 5, 10, 11, 12, 13) and (6, 7, 8, 9) as shown in Figure 4(a). The second largest cluster is GV_{10}. Vertices (6, 7, 8, 9) are subdivided into (6, 7, 8) and (9) (Figure 4(b)). The third largest cluster is GV_4. Vertices (1, 2, 3, 4, 5, 10, 11, 12, 13) are subdivided into (1, 2, 3, 4) and (5, 10, 11, 12, 13) (Figure 4(c)).

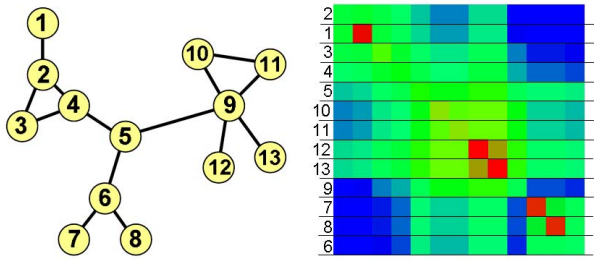
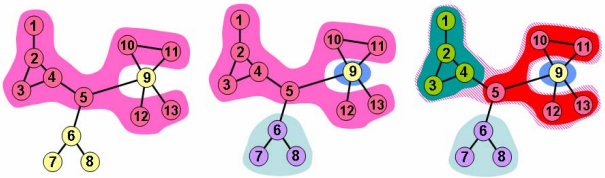


Figure 3. Network and rearranged similarity matrix



(a) (b) (c)

Figure 4. Clustering the network using the SMC method

Next, the characteristic features of SMC are examined by considering the resultant clusters from a network equivalent to the tripartite graph and comparing the result with that of the Newman-Girvan method. The difference between the two methods is shown in Figure 5. SMC can also be tested by using it to cluster an n-partite graph. In the case of n-partite graph clustering, it is found that SMC is superior to the

Newman-Girvan method. As well, SMC has different characteristic features than those of class (ii) methods.

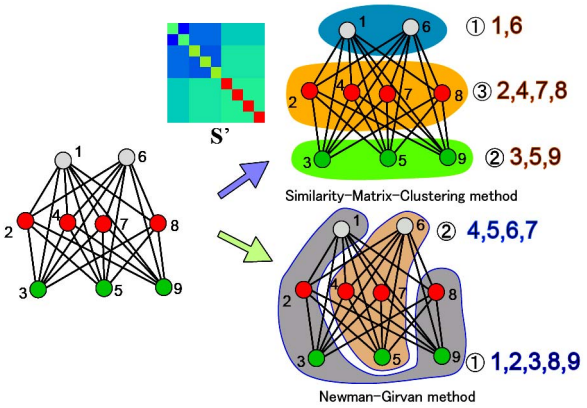


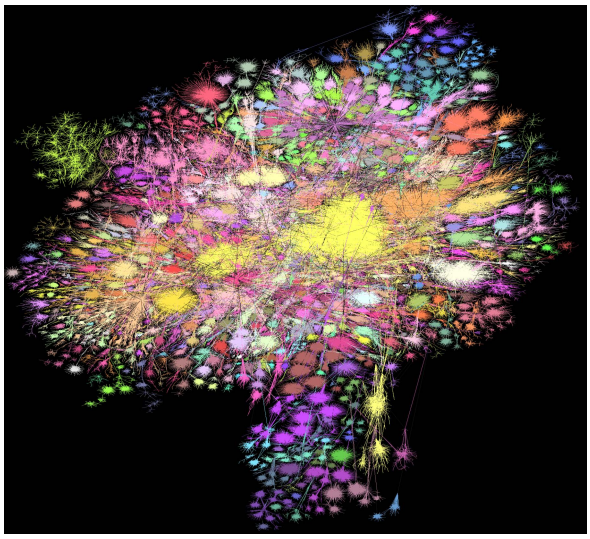
Figure 5. Comparison between SMC and Newman-Girvan methods

3. Results

The proposed method is used to identify communities in a blogosphere and in a Wikipedia word network.

In the blogosphere case, the number of vertices, which correspond to each individual page, is 181,714. A link represents a trackback. The total number of links is 626,365. 392 communities were detected using the Newman-Girvan method. The communities are visualized in the upper part of Figure 6. The links that are components of communities are colored by 392 different colors; See the detail in [8]. The SMC method is also used to identify each of the 392 communities. The lower part of Figure 6 shows some of the communities, which are denoted by the colored **S**'. It can be shown that some patterns appear in each community. The subgraphs corresponding to the patterns are shown in Figure 7.

In the Wikipedia word network case, a vertex corresponds to each individual word. The total number of words is 2,046. A link is a hyperlink among the words. The total number of links is 14,081. Figure 8 shows the different cluster patterns in the community. For example, a pattern in which all of the vertices in a community have the same homogeneous nature appears in community (l), while an *n*-partite graph structure can be seen in community (c).



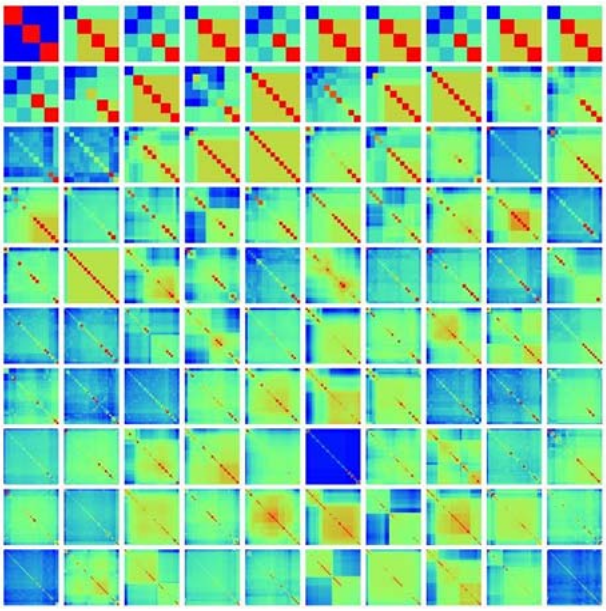


Figure 6. Blogosphere colored according to the communities(*upper figure*). Cluster patterns in the communities (*lower figure*)

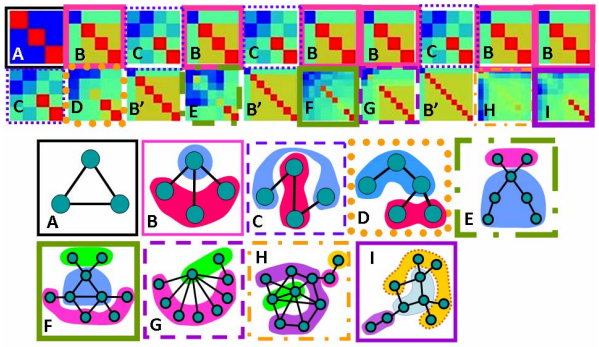


Figure 7. Cluster subgraphs that correspond to the patterns

Table 1. Number of vertices and edge density in community

	# of vertices	edge density		# of vertices	edge density
(a)	34	0.137	(k)	3	0.667
(b)	48	0.160	(l)	128	0.211
(c)	13	0.154	(m)	2	1.0
(d)	59	0.038	(n)	4	0.5
(e)	608	0.011	(o)	7	0.381
(f)	11	0.236	(p)	6	0.467
(g)	103	0.226	(q)	230	0.028
(h)	673	0.020	(r)	34	0.121
(i)	46	0.524	(s)	31	0.146
(j)	5	0.4			

Table 1 gives the number of vertices and edge density in each community. By comparing Figure 8 (lower part) with Table 1, it can be seen that the communities with a similar number of vertices and edge density have different visualization patterns, for example communities (r) and (s). On the other hand, communities with a similar visualization pattern have a different number of vertices and edge density, for example communities (b) and (q).

This implies that applying the SMC method to the communities detected by Newman-Girvan method can show the characteristic pattern for each community. Thus, it can be concluded that the communities can be characterized and identified using this inherent pattern for each community.

4. Conclusions

A new method for identifying detected communities based on the graph substructure is presented. The proposed method is applied to the identification of communities in a blog network and in a Wikipedia word network.

Despite being only qualitative, the results show that this method can successfully identify communities using only topological information.

Acknowledgments

This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 17300029.

References

[1] D. Gfeller, P. De Los Rios, A. Caflisch and F. Rao, "Complex network analysis of free-energy landscapes", Proc. Natl. Acad. Sci. USA 104, pp.1817–1822 (2007)

[2] M. Uchida, N. Shibata and S. Shirayama,"Identification and Visualization of Emerging Trends from Blogosphere", Proceedings of ISWSM, pp. 305-306 (2007)

[3] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks", Phys. Rev. E 69, 026113 (2004).

[4] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, "Complex networks: Structure and dynamics", Physics Reports, 424, pp. 175-308 (2006)

[5] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, Cambridge University Press (1994)

[6] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment", Journal of the ACM, 46(5), pp. 604–632 (1999)

[7] E. A. Leicht, et al., "Vertex similarity in networks", Phys. Rev. E 73, 026120 (2006)

[8] M. Uchida and S. Shirayama, "Formation of patterns from complex network", Journal of Visualization, 10(3), pp.253-255 (2007)

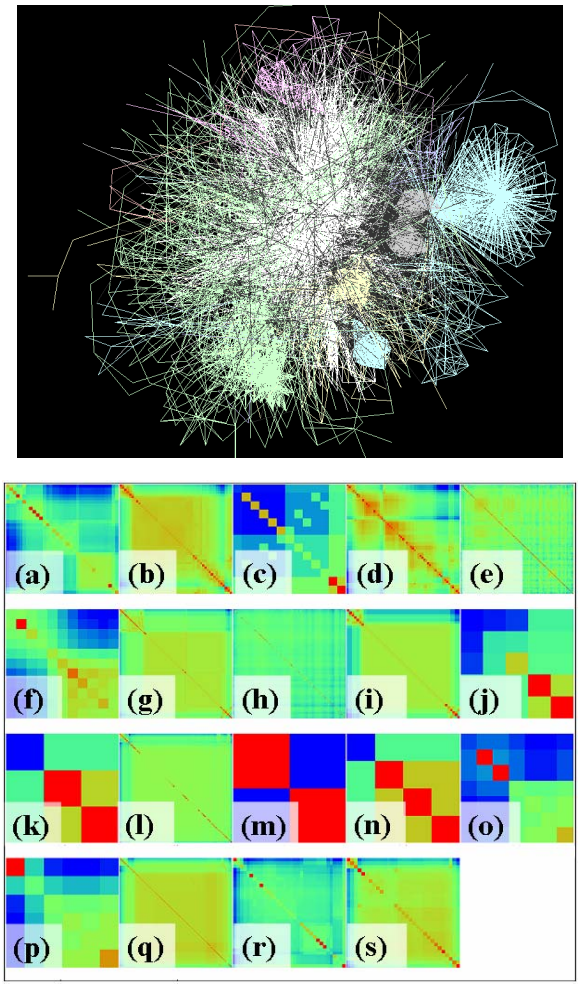


Figure 8. Wikipedia word network (*upper figure*). Patterns of clusters in communities (*lower figure*)