Community Detection Through Optimal Density Contrast of Adjacency Matrix

Tianzhu LIANG, Kwok Yip SZETO

Bioengineering Program, Hong Kong University of Science and Technology
Physics Department, Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
e-mail: henryltz@gmail.com, phszeto@ust.hk

Abstract. Detecting communities in real world networks is an important problem for data analysis in science and engineering. By clustering nodes intelligently, a recursive algorithm is designed to detect community. Since the relabeling of nodes does not alter the topology of the network, the problem of community detection corresponds to the finding of a good labeling of nodes so that the adjacency matrix form blocks. By putting a fictitious interaction between nodes, the relabeling problem becomes one of energy minimization, where the total energy of the network is defined by putting interaction between the labels of nodes so that clustering nodes that are in the same community will decrease the total energy. A greedy method is used for the computation of minimum energy. The method shows efficient detection of community in artificial as well as real world network. The result is illustrated in a tree showing hierarchical structure of communities on the basis of sub-matrix density. Applications of the method to weighted and directed networks are discussed.

Keywords: community detection, node clustering, energy minimization, greedy method, sub-matrix density, hierarchical structure.

1. Introduction

Nowadays community detection is becoming an increasingly important problem in the research of many areas in science and engineering. Biologists find the communities of proteins or cells to reveal the complex but organized system. Social scientists study the communities of social networks to investigate human behaviors. Analyzing communities of the stock market can show the relationships between different stocks. Studying communities in online social networks, such as facebook, can give us more information about our online behaviors. Analysis of ferromagnetic interactions between particles also reveals properties of ferromagnetic materials (Garliauskas, 2009). However, the definition of a community is still diverse (Newman and Girvan, 2004; Reichardt and Bornholdt, 2006; Newman, 2004; Newman, 2006; Li *et al.*, 2008; Hastings, 2006), making the problem of community detection more difficult. Nevertheless, a common feature of various mathematical definitions of community is that a community is a group of nodes that has strong relations, manifested through the higher density of links within the group than the density of links between nodes from different groups. This is the feature that we use in

this paper to define community. Thus, the problem of community detection becomes one of finding groups of connected nodes so that the density of links within group is highest, resulting in an image analysis problem for an adjacency matrix with highest density contrast. Many techniques have been explored to solve the problem of community detection (Newman and Girvan, 2004; Reichardt and Bornholdt, 2006; Newman, 2004, 2006; Li *et al.*, 2008; Hastings, 2006; Šutiene *et al.*, 2010). A common solution is to change the community detection problem into the optimization of some quantitative functions. Examples of these quantitative functions include the modularity introduced by Newman and Girvan in 2004 (Newman and Girvan, 2004) and the Hamiltonian in Potts model introduced by Reichardt and Bornholdt in 2006 (Reichardt and Bornholdt, 2006). Both techniques show effectiveness in community detection, but some research indicates that resolutions of community sizes are limited for these techniques (Kumpula *et al.*, 2007a, 2007b; Fortunato and Barthelemy, 2007). Improvement of quantitative functions has been tried (Li *et al.*, 2008). Special community detection methods are also explored to search for fuzzy communities (Reichardt and Bornholdt, 2004; Nepusz *et al.*, 2008). Specific methods are also designed for special networks (Barber, 2007). In this paper we introduce a new community detection method by defining an energy function for the network. After minimizing the energy, nodes from the same community will be clustered. Then we use a technique of tree building to construct hierarchical structure of communities. The technique is based on the density ordering of subsets. We have tested our method for both the artificial networks and real networks and find it to be very efficient.

2. Node Clustering

A network of N nodes can be described by its adjacency matrix A, which is an $N \times N$ matrix whose elements are defined by $A_{ij} = 1$ when there is a link between node i and j and is zero otherwise. In general, the raw data of the network will have a corresponding adjacency matrix with a distribution of 1 and 0 that is rather uneven, or seemingly random. A common feature for all definitions of community is that a community in the network can be considered as a group of nodes such that the links inside the community is dense. In order to properly define the density of links, we introduce the notion that there are two kinds of links, one between nodes inside the same community and the other kind of links are those that connect nodes inside the community with some nodes outside the community. Since the labeling of the nodes does not alter the topology of the network, we can label nodes belonging to the same community sequentially. In this way, the corresponding adjacency matrix, which depends on our labeling of nodes, will have a block form. Each block will consist of nodes in a common community. In another word, since nodes in the same community are more likely to have links, the corresponding adjacency matrix A_{ij} is more likely to be 1 if node i and node j are in the same community. For example, if nodes 1 to 10 belong to the same community, then in the first 10×10 sub-matrix, we have more likely to find 1s inside this block than outside, corresponding to a sub-matrix with higher density than other parts of the adjacency matrix.

The problem for community detection is then transformed into one of relabeling the nodes in such a way as to reveal the block structures of the adjacency matrix. The best relabeling will provide the highest density contrast, where the density of links is the density of 1s. The quality of the detection algorithm is then measured by the image quality of the adjacency matrix, with highest density contrast being the better description of the community structure of the given network.

Since there are $N!$ ways to label the nodes, exhaustive search will be prohibitive for large N. In this paper, we aim at finding an efficient algorithm to search among these $N!$ ways of relabeling the one that yields the highest density contrast in the adjacency matrix. With this in mind, we introduce the concept of network energy. The network energy is defined on the basis of the adjacency matrix:

$$E = \sum_{i,j} w(i,j) \cdot A_{ij}.$$

Here i, j denotes the index order of nodes, which runs from 1 to N for a network with size N. The weight $w(i,j)$ measures the contribution of the link A_{ij} to the total network energy. Here the weight has the following properties:

(1) $w(i,j) = f(|i-j|)$, which means $w(i,j)$ is only a one-variable function of $|i-j|$.
(2) $f(d_1) > f(d_2)$, when $d_1 > d_2$.

Since when $i = j$, A_{ij} is 0 (no self linkage), we can assume $|i-j| > 0$. Here are two examples of the definition of weight function:

(1) gravitational energy:

$$w_1(i,j) = -\frac{1}{|i-j|},$$
$$E_1 = \sum_{i,j} A_{ij} \cdot \left(-\frac{1}{|i-j|} \right).$$

(2) distance:

$$w_2(i,j) = |i-j|,$$
$$E_2 = \sum_{i,j} A_{ij} \cdot |i-j|.$$

From the conditions of the weight function we can infer that the smaller the distance is between the position of entry and the diagonal line of adjacency matrix, the smaller $|i-j|$ is, and the smaller $w(i,j)$ will be. We then like to minimize the total energy so that nodes in the same community can be clustered. As it is more likely to have links between nodes from the same community, more 1s will appear near the diagonal of the adjacency matrix after the minimization of total energy. This in turn will give a better node clustering.

Note that the weight function is thus far arbitrary as long as condition (1) and (2) are satisfied. The form of the energy function is just a mean to cluster the nodes, and there

is no physical significance on their interpretation. The objective is to cluster nodes by minimizing energy, though which particular choice of energy weight $w(i, j)$ is reflected in the efficiency of the search algorithm for a good labeling of the nodes to reveal highest density contrast. In literature, the idea of node clustering is often used before constructing a hierarchical tree of communities (Sales-Pardo *et al.*, 2007).

3. Computation of Minimum Energy

In our community detection method, we use a recursive procedure to compute the index order that minimize the energy of network. Here a greedy algorithm is applied in order to achieve high efficiency.

First we need to define the neighbor index order for this optimization problem. For a network of size N, a particular index order, I, is a $1 \times N$ array. We define the neighbor index order X as an index order that can be acquired by exchanging two elements of I. So for a network of size N, every index order I has $N \cdot (N-1)/2$ neighbor index orders.

Our algorithm for the minimization of the network energy is a greedy method, which involves the continuous evolution of the index order I to a better neighbor index order X. Suppose the initial index order is denoted by I_0. In the pth evolution, the five steps of our algorithm are,

1. Calculate the energy after the last evolution $E(I_{p-1})$ with the index order I_{p-1}.
2. Search for the best neighbor index order of I_{p-1}, which is denoted by X_{p-1}.
3. Calculate the energy $E(X_{p-1})$ with the index order X_{p-1}.
4. If $E(X_{p-1}) < E(I_{p-1})$, then $I_p = X_{p-1}$. Go to the next evolution.
5. If $E(X_{p-1}) > E(I_{p-1})$, then I_{p-1} is better than its neighbors. Stop the evolution.

In this greedy method, the network energy E always decreases as much as possible. It has the advantage of fast convergence but with the drawback of the danger of being trapped in local energy minimum.

4. Application Tests

In this section, artificial networks with different size and community numbers are used to test our community detection method. In addition, a network of American college football games with 115 nodes as well as a dolphin social network of 62 nodes are also used for the testing of the method. All the tests demonstrate that our method is quite efficient.

4.1. *Artificial Network*

The artificial networks we use are ideal networks with known community structure, with nodes in the same community fully connected and nodes in different communities un-connected. We choose an index order randomly to hide the community structure and then apply the greedy method to search for the best index order. Here the network energy is

Table 1

Test results for ideal networks

Network size	Community number	Community size	Test times	Average iteration	Number of times trapped in local minimum
25	2	21, 4	10	3.9	0%
25	4	12, 8, 3, 2	10	7.5	0%
50	2	28, 22	10	11.2	0%
50	6	20, 12, 12, 3, 2, 1	10	19.2	0%
75	4	46, 20, 6, 3	10	21.2	0%
75	7	35, 23, 7, 4, 3, 2, 1	10	27.9	0%
100	3	86, 11, 3	10	11.9	0%
100	6	44, 25, 19, 7, 4, 1	10	37.3	0%

defined as total distance of nodes. For each artificial network, we perform our algorithm 10 times and the results are shown in Table 1. In these tests, we never run into local minimum, showing that our method is effective for networks with ideal community structure. We can also see that the average iteration required for the detection of the community structure increases with the network size as well as the number of communities.

4.2. *Football Network*

One real network we use to test our community detection method is the American college football network between Division I for the regular season in fall 2000 (Girvan and Newman, 2002). In this network, nodes represent football teams and edges represent existence of games between the two teams. Community structure exists in this network because the teams are divided into conferences, with about 8–12 teams each. Teams in the same conference play much more games than teams from different conferences. Again we randomly choose an index order and apply the greedy method. Figures 1 and 2 are the pictures of the adjacency matrix before and after the adjustment of index order.

Here a white dot denotes a 1 in the matrix and a black dot denotes a 0. In Fig. 2, several blocks with high density appear along the diagonal line. The greedy method works well in this example.
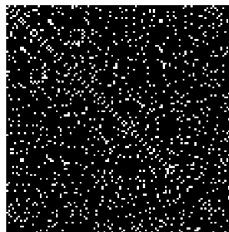

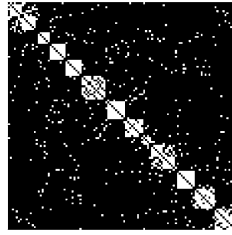Fig. 1. Adjacency matrix of original football network.
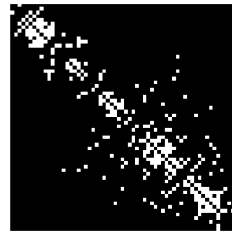
Fig. 2. Adjacency matrix of adjusted football network.

Fig. 3. Adjacency matrix of adjusted dolphin network.

4.3. *Dolphin Network*

The second real world network for the validation of our community detection method is a social organization formed by bottlenose dolphins living off Doubtful Sound, New Zealand (Lusseau *et al.*, 2003). In this network, nodes represent dolphin individuals and links represent that there is a frequent association between two dolphins. Previous work indicates that large communities as well as tiny communities (communities with only two or three dolphins) probably exist in this dolphin network, although there are also some random links between communities to make the whole network more insensitive to sudden disasters (Lusseau, 2003; Lusseau and Newman, 2004). Now we use our greedy method to see whether community structure exists in this network. Results are shown in Fig. 3. A few blocks do form at the diagonal line. However, there are also some 1s far away from the diagonal line in the network, meaning that there are links connecting different blocks in the right bottom corner. This observation is consistent with previous results, which also indicates that some blocks may interact with each other to form a larger block.

Summarizing the above results, we conclude that the greedy method is effective for both artificial and real world networks.

5. Hierarchical Structures of Communities

In this section, we will introduce a technique to build a hierarchical tree of communities based on sub-matrix density. It is an agglomerative method for hierarchical clustering

that works on the adjacency matrix after node clustering. We treat the whole network as a large community, which is composed of several smaller communities. Each community can also be composed of even smaller communities. Since nodes in the same community are more likely to have links, we assume that the density of parent community is smaller than its sub-communities. Suppose the community is of size M. We define density as $D = \frac{1}{M^2} \sum_{i,j} A_{ij}$. The technique of tree building is based on the adjusted adjacency matrix. Note first that the existence of a community in the network corresponds to the existence of a block of dense 1s around the diagonal of the adjacency matrix after index order adjustments. Given a community of size M, we can let a sub-matrix of size M slide along the diagonal line and calculate its density, as is shown in Fig. 4. We then obtain a density curve $\rho(k, M)$ parameterized by the position k of the sliding block of size M. Here k is the index along the diagonal, $0 < k < N - M$, with N denoting the size of the network. It is reasonable that a density peak will appear at the position of the community, since the density of the community should be larger than its neighboring subsets, which is composed of nodes from more than one community. With this insight, we conclude that if there is a community of size M, then there will be a peak in the density curve as we slide down the block. However, a density peak does not always indicate a community. The reason is that we do not know a priori the size of the community, so that a density peak for a particular chosen block size M need not correspond to the existence of a community of size M. Nevertheless, we can try all possible block size M. For a given size M, we can draw a density curve and find the local maxima of the density curve. If there is a local maximum of the density curve at k, we interpret that there is a possible community of size M with nodes inside the block of size M located at the diagonal position k. This particular subset of nodes, is denoted as $V(k, M)$. It contains M nodes. We call $V(k, M)$ the candidate set of nodes forming a possible community of size M. Let's denote the set of candidate set $V(k, M)$, corresponding to a local maximum for the density curve $\rho(k, M)$ for block size M, as $S(M)$.

$$S(M) = \bigcup_{h=1}^{H(M)} V\big(k(h), M\big).$$

Here h is the label for the hth local maximum along the density curve, and $k(h)$ is the corresponding position of this local maximum along the diagonal. $H(M)$ is the total number of local maximum for the density curve $\rho(k, M)$. The collection of sets of local maxima $\{S(M)\colon M = 2, \ldots, N - 1\}$ consists of all possible sets of block of size $M = 2, \ldots, N - 1$ that are candidate communities.

Now we can build a tree of communities. We rank these local maxima to form an ordered list of candidate communities. We examine the candidate sets one by one in the odered list and we first select the candidade set $V(k, M)$ with the largest density. Then decision on selecting candidate set in the list is made based on the following criterion:

1. If none of the nodes in the selected group belongs to a certain community that already exists, then this group forms a new community.
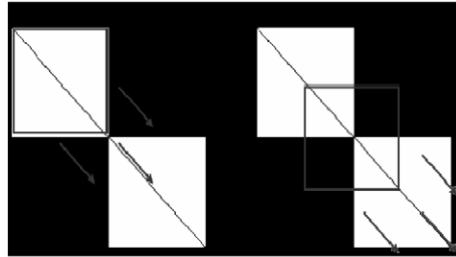
Fig. 4. A block of size M sliding along the diagonal line.

2. If the selected group of nodes includes one or more communities that already exist and it does not overlap with other communities, then a hierarchical structure of communities is built. The selected group of nodes is identified as the parent community of sub-communities included.
3. If the selected group of nodes overlaps with one or more communities that already exist, it is not considered as a community.

The last criterion is based on our naive assumption that communities do not overlap. In our classification, if two possible subsets of nodes overlap, only the group with a larger density is considered as a community.

With this technique, we can build a hierarchical structure of communities with detailed community information, such as the number of communities, the size of each community and the tag of nodes that compose each community. Test results are shown below.

5.1. *Hierarchical Structures of Artificial Network*

The technique of tree building is effective for ideal networks generated by the computer. For example, a network of size 100 with 5 communities is shown in Fig. 5. The information about the groups of nodes is given in Table 2. Node 100 is an independent node and does not belong to any communities. The hierarchical community structure for this ideal network is shown in Fig. 6.

Here '$(x\ y), z$' denotes that the community is consisted of the xth to the yth nodes after the relabeling of index order and the density of links is z. We can see that the community information is shown correctly.
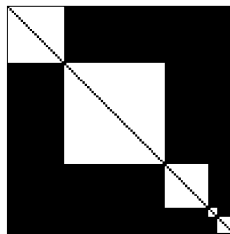

Fig. 5. Ideal network used to test the tree building technique.

Table 2

Groups of nodes in the ideal network

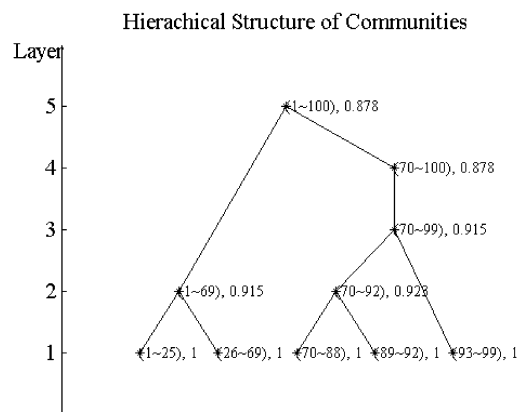| Group number	Nodes
1	1–25
2	26–69
3	70–88
4	89–92
5	93–99

Fig. 6. The tree of ideal network communities.

5.2. *Hierarchical Structures of Football Network*

In this section, we use the American college football network to test the technique of tree building. Results are shown in Fig. 7. The whole network can be divided into 17 communities. Detailed information of these communities is available. As we know the truth of community structure of this football network, we can compare it with the results of tree building, as shown in Table 3. From the results we can see that, more than 95% of nodes are grouped correctly. For some conferences, teams are divided into more than one group, which means sub-community structure is detected in these conferences.

5.3. *Hierarchical Structures of Dolphin Network*

To validate our tree building technique, the dolphin network is also used. Results (Fig. 8) shows that two large communities exist in the dolphin network (nodes 1 to 21 and nodes 23 to 59) while the larger community of the two can be further divided into 3 sub-communities (nodes 23 to 33, nodes 34 to 47 and nodes 48 to 59). In addition, tiny communities composed of two or three dolphins are also detected in our method (see Table 4 for details). Comparing our results with previous results (Lusseau and Newman,
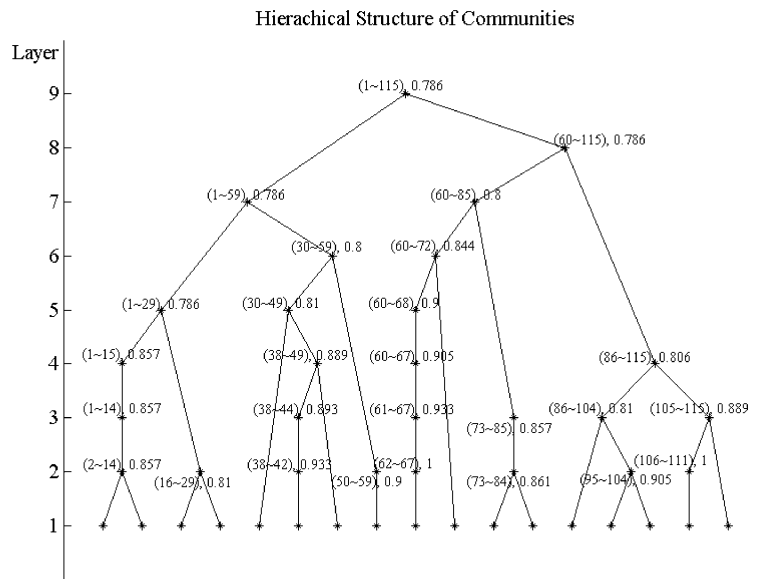
Hierachical Structure of Communities

Fig. 7. The tree of football network communities.

Table 3

Results of tree building for football network

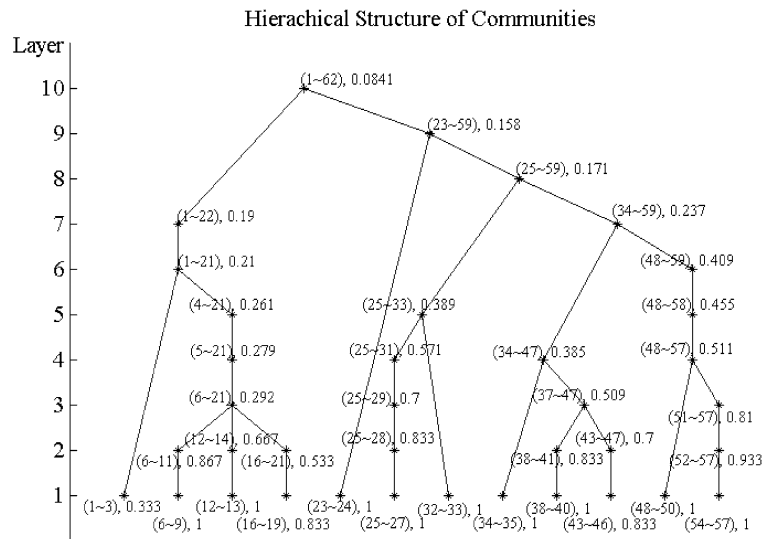| Label of community detected	True community labels in that community
1	6, 6, 6, 6, 6, 6
2	6, 6, 6, 6, 6, 6, 6
3	11, 5, 10, 10, 10, 10
4	1, 1, 1, 1, 1, 1, 1, 1
5	7, 7, 7, 7, 7, 7, 7, 7
6	9, 9, 9, 9
7	9, 9, 9, 9, 9
8	0, 0, 0, 0, 0, 0, 0, 0, 0
9	4, 4, 4, 4, 4
10	10, 10, 10, 11
11	3, 3, 3, 3, 3, 3
12	3, 3, 3, 3, 3, 3
13	11, 11, 11, 11, 11, 11, 4, 11, 11
14	8, 8, 8, 8, 8
15	8, 8, 8, 8, 8
16	2, 2, 2, 2, 2
17	2, 2, 2, 2

Hierachical Structure of Communities

Fig. 8. The tree of dolphin network communities.

Table 4

Results of tree building for dolphin network

Nodes	Name of nodes	Previous results (Lusseau and Newman, 2004)
1–3	Zig, TR82, Ripplefluke	In community 1
6–9	Feather, DN21, Gallatin, Web	In community 1
12–13	Jet, MN23	In community 1
16–19	Mus, Number1, Notch, Beescratch	In community 1
23–24	TR88, TR120	In community 2.3
25–27	PL, Oscar, TR77	In community 2.2
32–33	Zipfel, TSN83	In community 2.3
34–35	Thumper, Shmuddel	In community 2.3
38–40	SN63, Stripes, Scabs	In community 2.3
43–46	Grin, SN4, SN9, SN100	In community 2.3
48–50	Double, CCL, Zap	In community 2.3
54–57	Jonah, Patchback, MN105, Trigger	In community 2.1

2004), we find that only minor differences are present (e.g., tiny communities are grouped a little bit differently to form sub-communities in the two largest communities).

Note that in Table 4, community 2.1, 2.2, 2.3 denote the 1st, 2nd and 3rd sub-community in the community 2, corresponding to the dark grey, light grey and white sub-community in Lusseau and Newman (2004).

6. Discussion

Here we like to discuss the generalization of our method to weighted and directed network.

For weighted network, the elements of adjacency matrix can be any real values. Our energy minimization method can still be used. The elements in our energy function $E = \sum_{i,j} f(|i - j|) \cdot A_{ij}$ with large A_{ij} will be pushed to the diagonal line after index order adjustment, which means nodes that have strong links will be clustered together. On the other hand, the density of subset is still defined as average of elements in the sub-matrix. So subsets that contain strong links will have a larger density, which may be identified as a community with higher probability. We can still apply our tree building method to detect hierarchical structure of communities. We expect that our community detection method will still be effective for weighted networks.

As for directed network, we cannot apply our techniques directly to detect communities. Our definitions for the energy function as well as density of block of size M are defined in reference to a symmetric adjacency matrix, corresponding to undirected network. For a directed network, we have an asymmetric adjacency matrix. However, we can construct its symmetric part by defining

$$A'_{ij} = \frac{1}{2}(A_{ij} + A_{ji}).$$

If we assume that the results of community detection on the symmetric part of the directed network is the same as that of the original network, A_{ij}, then we can apply our technique. The validity of this assumption is yet to be tested on real data.

7. Conclusion

In this paper we introduce a method of community detection, which includes the process of node clustering and tree building. By the introduction of interaction energy between labels of nodes, we can find the minimum energy of the network, corresponding to a re-labeling of the nodes so as to reveal the clustering structure in the adjacency matrix. This method has been tested on idealized network structure with pre-assigned community structures, as well as on real network data. Our results show that the method is quite efficient, as we use the greedy method to minimize energy for index ordering of nodes. Furthermore, with a sliding block algorithm on the density curve, we collect all possible candidate sets for communities. Based on these candidate sets, we can build a tree to reveal hierarchical structure of communities. It is based on the assumption that the density of a community is larger than neighboring subsets composed of nodes from more than one community. This tree building technique is tested on both artificial networks with pre-assigned communities as well as real networks. The results are very good.

Our method is simple and efficient, though the search for local maxima along the density curve using greedy method may sometimes run into errors, the method overall

is quite fast. The problem of tree building is simple and good, though error may occur in situation of overlapping communities. One advantage of tree structure is that it shows more information about communities. We can see which communities have closer connection with each other, indicating that they may form a larger community with a smaller density. Future research will apply our techniques for weighted and directed network, as well as addressing the more theoretical issue of overlapping communities. We hope that this technique also find application in knowledge discovery in data mining on the web.

References

Barber, M.J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6), 066102.

Fortunato, S., Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41.

Garliauskas, A. (2009). Embedded patterns, indirect couplings with randomness, and memory capacity in neural networks. *Informatica*, 20(4), 477–486.

Girvan, M., Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.

Hastings, M.B. (2006). Community detection as an inference problem. *Physical Review E*, 74(3), 035102.

Kumpula, J.M., Saramki, J., Kaski, K., Kertsz, J. (2007a). Limited resolution and multiresolution methods in complex network community detection. In: *Proceedigs of the Society of Photographic Instrumentation Engineers*, Vol. 6601. The Society of Photographic Instrumentation Engineers, USA, 660116.

Kumpula, J.M., Saramki, J., Kaski, K., Kertsz, J. (2007b). Limited resolution in complex network community detection with Potts model approach. *Proceedings of the Society of Photographic Instrumentation Engineers*, 56(1), 41–45.

Li, Z., Zhang, S., Wang, R., Zhang, X., Chen, L. (2008). Quantitative function for community detection. *Physical Review E*, 77(3), 036109.

Lusseau, D. (2003). The emergent properties of a dolphin social network. *Proceedings of the Royal Society B*, 271, S477–S481.

Lusseau, D., Newman, M.E.J. (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society B*, 270, S186–S188.

Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E., Dawson, S.M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54, 396–405.

Nepusz, T., Petrczi, A., Ngyessy, L., Bazso, F., Chen, L. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1), 016107.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 066133.

Newman, M.E.J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 036104.

Newman, M.E.J., Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.

Reichardt, J., Bornholdt, S. (2004). Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93(21), 218701.

Reichardt, J., Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1), 016110.

Sales-Pardo, M., Guimera, R., Moreira, A.A., Amaral, L.A.N. (2007). Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39), 15224–15229.

Šutiene, K., Makackas, D., Pranevičius, H. (2010). Multistage K-means clustering for scenario tree construction. *Informatica*, 21(1), 123–138.

T. Liang studied at Tsinghua University, Beijing, China, since 2004 and received a BS degree in July, 2008. During January of 2008, he was a junior research assistant in the Physics Department, the Hong Kong University of Science and Technology, Hong Kong. Now he is a PhD student in the Bioengineering Program, the Hong Kong University of Science and Technology, Hong Kong.

K.Y. Szeto is an associate professor in the Physics Department, the Hong Kong University of Science and Technology, Hong Kong. He received his BSc (Eng. Sci) from the University of Toronto and a PhD in Physics at M.I.T.

Bendrijų nustatymas per optimalaus tankio kontrasto gretimumo matricas

Tianzhu LIANG, Kwok Yip SZETO

Bendrijų nustatymas realaus pasaulio tinkluose yra aktuali duomenų analizės problema. Darbe sukurtas rekursyvinis algoritmas bendrijoms nustatyti klasteriuojant viršūnes intelektiniu būdu. Kadangi viršūnių pervardijimas nekeičia tinklo topologijos, bendrijų nustatymo problema suvedama į viršūnių sužymėjimą, suformuojant gretimumo matricoje atitinkamus blokus. Įvedus fiktyvius sąryšius tarp viršūnių, viršūnių pervardijimo problema suvedama į energijos minimizavimo uždavinį, apibrėžiant bendrąją tinklo energiją per sąryšius tarp viršūnių tokiu būdu, kad viršūnės, esančios toje pačioje bendrijoje, mažintų energiją. Emergijos minimumui apskaičiuoti yra pritaikytas godusis algoritmas. Metodas leidžia efektyviai nustatyti bendrijas dirbtiniuose ir realiuose tinkluose. Rezultatai iliustruojami medžiams turintiems hierarchinę struktūrą. Aptariami metodo pritaikymai svoriniams ir kryptiniams grafams.