3C4-IOS-4a-5

# Community Detection on Heterogeneous Networks

Weichu Liu*1 Tsuyoshi Murata*1 Xin Liu*1

*1 Department of Computer Science
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Many real-world systems can be described as networks, where the nodes represent the objects of a system and the edges represent relationships between them. Among them there are many heterogeneous systems which contain different types of objects and edges. Previous researches on community detection often focus on homogeneous networks, while combining nodes of different types provides richer information for us to understand the structure of a network better. In this paper, we propose a method to perform community detection on heterogeneous networks, and explain some experiments on synthetic networks.

## 1. Introduction

In recent years, data mining from networks has attracted many researchers due to its broad application in many fields such as physics, bioinformatics, sociology, etc. A big topic is to detect communities from networks. Relationships in the real world can be abstracted as a network, whose nodes represent objects and edges represent the relationship/interaction between the nodes. A community is a set of nodes in which its nodes have strong connections to each other. By community detection, we can obtain knowledge about the structure and the correlation between nodes.

The former researches on community detection overwhelmingly focus on homogeneous networks. A homogeneous network is composed of single type of nodes, which is suitable to describe a homogeneous system. Nevertheless, in the real world, there are many systems containing different types of objects and relationships. For example, in a SNS site, users and their friendship can be regarded as a homogeneous network. But there may be much more useful information contained in the posting - replying system for us to understand the relationships between users. By considering nodes to represent post and edges to represent who replied it, we get a bipartite network. Also, there could be hyperedges linking more than 2 nodes such as user-tag-article links, which appear in many site using tagging system.

A method often used in community detection is to optimize modularity. The concept of modularity is firstly proposed by Newman-Girvan[1]. It is a function that measures the quality of a division of a network into groups or communities. By optimizing the modularity, a division of network can be obtained.

In addition to a unipartite network, the concept of modularity has also been extended into a bipartite network. A bipartite network is a network composed of two types of nodes and there are edges only when it connects to the nodes of both sides. Bipartite networks are often used to express relationship between customers and goods they bought or users and the website they visited. Using unipartite modularity to perform community detection on bipartite network is insufficient. Consider a community obtained using unipartite modularity definition; there should be a dense connection inside it, while there are no edges between nodes of same type. Therefore, the result is wrong. The definition of modularity has been extended to bipartite networks by several researchers[2][3].

Similarly, it is insufficient to detect community on a heterogeneous network using methods for homogeneous networks. What is more, the unipartite modularity cannot handle hyperedges. Also, heterogeneous networks are different from k-partite networks since there may be edges connecting nodes of the same type. Therefore, we proposed a new transforming - detecting method to perform community detection on heterogeneous network.

Contact: Weichu Liu, liu.w@ai.cs.titech.ac.jp, W8-59 2-12-1 Ookayama, Meguro, Tokyo, 152-8552

## 2. Related Work

### 2.1 Modularity

Modularity proposed by Newman-Girvan[1] is a quality function to measure how good a particular partition of a network is. It is based on the idea to define the quality of a community by comparing the edge density with their null model and calculate the deviation. Further, the quality of a division of a network is the sum of modularity of all communities.

Consider a network composed of M edges and node set V and divided into communities. Its adjacency matrix is A. Consider two communities $l$ and $m$, the fraction of links between them over all edges is defined as $e_{lm}$:

$$e_{lm} = \sum_{i \in V_l} \sum_{j \in V_m} \frac{A_{i,j}}{2M}$$

On the other hand, the sum of degrees of all nodes in community $l$ is expressed as $a_l$. Therefore, the deviation between the number of edges in a community and its null model is $e_{ll} - a_l^2$. By adding modularity of all communities, modularity of the division can be calculated.

$$Q = \sum_l (e_{ll} - a_l^2).$$

### 2.2 Bipartite Modularity

The definition of modularity has been extended by many researchers respectively. Barber's definition[2] considers bipartite communities in which nodes of different sides both exist. Therefore, it is similar to the original definition and the communities of each side have an one-to-one relationship. While in the Murata's definition, the communities only contain either side of nodes and it describes a many-to-many correspondence.

### 2.3 BiNetFinder Method

Liu[4] developed a community detection method called BiNetFinder for bipartite networks based on the idea that the description of a graph as communities can be viewed as a lossy compression of the graph's structure, and the community detection problem as a problem of finding an efficient compression of structure.

Therefore, Liu proposed a quality function based on the insight of information compression method. Suppose a signaler who aims to transmit much of information of a network $H$ in a reduced fashion to a receiver over a lossless channel. To do so, the signaler makes a partition of H and transfer the structural information of communities to the Receiver. Then the Receiver tries to recover the original structure information of the network. The quality function is defined by the loss of information during the compression. Liu also developed a method to optimize the quality function. By minimizing the information loss, a division can be obtained. With such method, nodes connected by parallel edges, i.e., the nodes that are similar to one another as regards their relations to nodes of other types, will be merged into the same community.

## 3. Heterogeneous Network Transformation and Community Detection

### 3.1 Heterogeneous Network Transformation

Given a heterogeneous network, firstly we transform it into a bipartite network by using the following procedure:

- Build a bipartite network, which contains two different types of nodes — the vertex nodes and the link nodes. For each node in the original network, we put a corresponding vertex node in the new network, for each edge in the original network, we put a link node in the new network to represent it.

- Therefore, each node and edge/hyperedge in the original heterogeneous network is, respectively, mapped into a vertex node and a link node in the bipartite network.

- A vertex node and a link node are connected if and only if in the original network the vertex is connected by the link. Thus far, we transformed the original network into a new bipartite network, as shown in Fig 1.

Unlike traditional approaches such as projection, our transformation reserves all information of the original network. Consequently, community detection on heterogeneous networks is turned into community detection on bipartite networks.

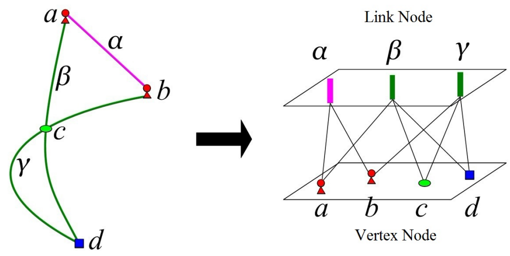


Figure 1: Transforming a heterogeneous network to a bipartite network. Latin letters a, b, c, d denote nodes of heterogeneous network, or the vertex nodes of the bipartite network. Greek letters $\alpha$, $\beta$, $\gamma$ denote edges/hyperedges of the heterogeneous network, or link nodes of the bipartite network.

### 3.2 Community Detection

After transformation, community detection is performed on the corresponding network. However, different from normal bipartite networks, such corresponding bipartite networks is relatively sparse and asymmetric. From the perspective of number of nodes, in the original network, the number of edges is usually much more than the number of nodes. Additionally, the degrees of link nodes side is very low for that one link node only connects where one hyperedge connects, while all vertex nodes remain the same degrees as its corresponding node. Such feature suggests that general bipartite modularity optimization method is inefficient to detect communities from it. On the other hand, community detection method based on grouping up "nodes with similar linking pattern" makes sense since it in turn groups up link nodes, which is an edge in the original network, having similar linking pattern.

We performed the community detection based on Murata's bipartite modularity optimization and the BiNetFinder algorithm on the bipartite network in the experiment. And the result also shows that the BiNetFinder is efficient to detect communities from the corresponding bipartite network.

## 4. Experiment and the Result

### 4.1 Synthetic Network

An synthetic heterogeneous network is build to test our method. The model of it is shown below as Fig 2. The network is build to simulate the scenario of SNS website that there are not only interactive between users but also interactive through posts and tags. In the network, there are 3 types of nodes, 2 types of edges and 7 communities with 15 nodes each. Both hyperedges and edges connecting same type of nodes exist in the network.

In our synthetic network, community 1, 3 and 5 are strongly connected, while community 2, 4, 6 and 7 are strongly connected. Edges connecting communities are randomly connected to nodes inside community. Except for edges appeared on Fig 2, we added random edges as noise. The number of noise edges is from 5% to 25% of the total edges, and add 5% for each level.

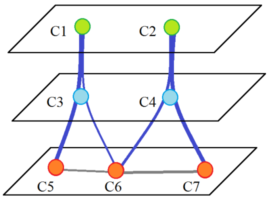


Figure 2: An synthetic heterogeneous network which contains 3 types of nodes, hyperedges and edges between the same type of nodes

### 4.2 Result

We tested three different methods on our synthetic networks with different level of noise. After transformation, we perform BiNetFinder and bipartite modularity optimization based on Murata's definition on the corresponding network. Besides, as a control, we perform Newman's modularity optimization on the original network. Since it cannot handle hyperedge, we transform each hyperedge into 3 edges linking 3 nodes the hyperedge links. The NMI (Normalized Mutual Information) is used to measure the goodness of partition. The result is showed as below as Fig 3.

As it shows, our method, which firstly transform the heterogeneous network and then perform community detection using BiNetFinder algorithm, gives a good result compared to Newman's method performed directly on the original network. A reason is that clustering edges and nodes based on the idea to merge nodes with similar linking pattern is intuitive in this case. Also, Newman's method's accuracy is lower because a hyperedge must be transformed into a clique, which permuted the information of original network. On the other hand, the method based on bipartite modularity optimization doesn't perform well. It shows that the characteristics of the transformed bipartite network make it not suitable for generalized bipartite modularity optimization methods.

## 5. Conclusions

In the paper, we proposed a new method to perform community detection on the heterogeneous network. By transforming an original network into a bipartite network and perform community detection on it, our method succeeds to detect community on synthetic heterogeneous networks and obtained better results compared to homogeneous method. In the experiment, we notice that the transformed networks is not suitable for methods based on bipartite modularity.

However, our method is time consuming and is not suitable for large scale networks yet. We hope to find a method for large-scale heterogeneous network in the future.

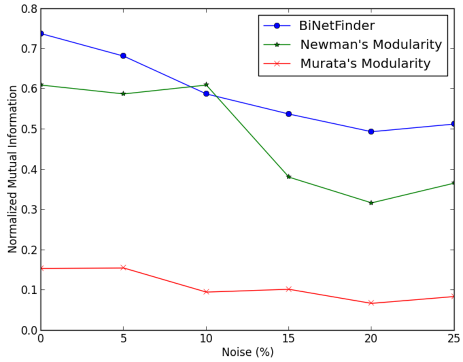


Figure 3: Comparison of the results of different methods based on NMI

## References

[1] Newman, M. E. J. and Girvan, M.: Finding and evaluating community structure in networks, Phys. Rev. E, 69(026113), American Physical Society, (2004)

[2] Barber, Michael J.: Modularity and community detection in bipartite networks, Phys. Rev. E, 76(066102), American Physical Society, (2007)

[3] Murata, T.: Detecting Communities from Bipartite Networks Based on Bipartite Modularities, Proceedings of the 2009 IEEE International Conference on Social Computing (SocialCom 09), pp. 50-57, IEEE Computer Society, (2009)

[4] Liu Xin, Tsuyoshi Murata: Detecting Communities in K-Partite K-Uniform (Hyper)Networks, Journal of Computer Science and Technology, Springer, 26(5), pp. 778-791, (2011)