New Modularity for Evaluating Communities in Bipartite Networks

Tsuyoshi Murata MURATA@CS.TITECH.AC.JP
Tokyo Institute of Technology, W8-59 2-12-1 Ookayama, Meguro, Tokyo, 152-8552 Japan

Keywords: modularity, communities, bipartite networks

Abstract

Detecting dense subnetworks (communities) and evaluating them are practically important for finding similar vertices in a network. Although Newman's modularity is widely used for evaluating network division, it is for unipartite networks composed of only one vertex type. As the attempts for evaluating divisions of bipartite networks, Guimera and Barber propose bipartite modularities. The author proposes another bipartite modularity that allows one-to-many correspondence of communities of different vertex types.

1. Introduction

Analysis of networked data attracts many researchers from computer science, physics, and sociology. Several relations in real-world can be represented as bipartite networks composed of two types of vertices, such as paper-author networks and event-attendee networks. Extracting communities from such bipartite networks and evaluating their qualities are practically important for understanding real-world networks.

Newman's modularity(Newman & Girvan, 2004) is widely used for evaluating the quality of divisions of unipartite networks, although Fortunato (Fortunato & Barthelemy, 2007) claims resolution limits of modularity-based division methods. Modularity is a scalar value that measures the density of edges inside communities as compared to edges between communities. As for bipartite networks, Guimera et al. (Guimera et al., 2007) and Barber (Barber, 2007) propose bipartite modularities. However, these are not appropriate for practical applications. Guimera's bipartite modularity focuses on the connectivity of only one vertex type, and Barber's bipartite modularity is based on an assumption that there is one-to-one correspondence between communities of different vertex types. The author proposes another bipartite modularity that allows one-to-many correspondence between communities of different vertex types (Murata, 2009). Preliminary experimental results using the above bipartite modularities are shown.

Presented at *ILP-MLG-SRL*, Leuven, Belgium, 2009.

2. Previous Modularities

2.1. Newman's Modularity

Let us consider a particular division of a network into k communities. M is the number of edges in a network, V is a set of all vertices in the network, and V_l and V_m are the communities. $A(i,j)$ is an adjacency matrix of the network. We can define e_{lm}, the fraction of all edges in the network that connect vertices in community l to vertices in community m as follows:

$$e_{lm} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V_m} A(i,j)$$

We further define a $k \times k$ symmetric matrix E composed of e_{ij} as its (i,j) element, and its row sums a_i:

$$a_i = \sum_j e_{ij} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V} A(i,j)$$

In a network in which edges fall between vertices without regard for the communities they belong to, we would have $e_{ij} = a_i a_j$. Modularity Q is defined as follows:

$$Q = \sum_i (e_{ii} - a_i^2)$$

2.2. Guimera's Bipartite Modularity

Guimera's bipartite modularity is defined as the cumulative deviation from the random expectation of the number of the communities of vertex type Y in which two vertices of type X are expected to be together:

$$M_B = \sum_{s=1}^{N_M} \left\{ \frac{\sum_{i \neq j \in s} c_{ij}}{\sum_a m_a(m_a - 1)} - \frac{\sum_{i \neq j \in s} t_i t_j}{(\sum_a m_a)^2} \right\}$$

where s is a community of vertex type X, N_M is the number of community of type X, a is a community of vertex type Y, m_a is the number of edges that are connected to the vertices in community a, c_{ij} is the number of communities of vertex type Y in which vertices i and j are connected, and t_i and t_j are total number of communities of type Y to which vertices i and j are connected, respectively.

Two vertex types are not symmetrical in the above definition. Guimera's bipartite modularity focus on the connectivities of only one vertex type (via the vertices of the other type). Connectivities of the other vertex type (M'_B) can be defined by swapping s and a in the formula of M_B.

2.3. Barber's Bipartite Modularity

Modularity is a deviation from null model, and bipartite networks have specific constraints that should be reflected in the null model. Barber (Barber, 2007) takes the constraints into consideration and formalizes bipartite modularity. Since there is no edge between the vertices of same type, the adjacency matrix of a bipartite network is as follows

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}_{p\times p} & \tilde{\mathbf{A}}_{p\times q} \\ (\tilde{\mathbf{A}}^{\mathbf{T}})_{q\times p} & \mathbf{0}_{q\times q} \end{bmatrix},$$

where $\mathbf{0}_{i\times j}$ is the all-zero matrix with i rows and j columns. Probabilities in the null model that an edge exists between vertices i and j are represented as follows:

$$\mathbf{P} = \begin{bmatrix} \mathbf{0}_{p\times p} & \tilde{\mathbf{P}}_{p\times q} \\ (\tilde{\mathbf{P}}^{\mathbf{T}})_{q\times p} & \mathbf{0}_{q\times q} \end{bmatrix},$$

Barber's bipartite modularity is defined as follows:

$$Q = \frac{1}{m} \sum_{i=1}^{p} \sum_{j=1}^{q} (\tilde{A}_{ij} - \tilde{P}_{ij}) \delta(g_i, g_{j+p}).$$

where g_i is the community that vertex i is assigned to, and δ_{ij} is the Kronecker's delta. This definition implicitly indicates that the numbers of communities of both types are equal. The weaknesses of Barber's bipartite modularity are: 1) the number of communities have to be searched in advance, and 2) the numbers of communities of both vertex types have to be equal.

3. Our New Bipartite Modularity

Let us suppose that communities of papers and communities of authors are detected from a paper-author network. If there is one-to-one correspondence between a paper community and an author community, it shows that the topics of the papers attract only limited authors (Figure 1). On the other hand, if there is one-to-many correspondence between a paper community and author communities, it shows that the topics of the papers attract several communities of authors (Figure 2).

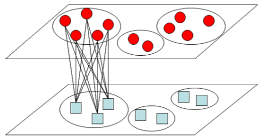

Figure 1. One-to-One Correspondence

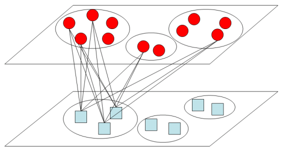

Figure 2. One-to-Many Correspondence

The constraint of one-to-one correspondence between communities of both types is removed in our definition of bipartite modularity. Let us suppose that M is the number of edges in a bipartite network, and V is a set of all vertices in the bipartite network. Consider a particular division of the bipartite network into X-vertex communities and Y-vertex communities. $A(i, j)$ is an adjacency matrix of the network whose (i, j) element is equal to 1 if vertices i and j are connected, and is equal to 0 otherwise.

Under the condition that the vertices of V_l and V_m are different types we can define e_{lm} (the fraction of all edges that connect vertices in V_l to vertices in V_m) and a_i (its row sums) just the same as those in section 2.1.

$$e_{lm} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V_m} A(i, j)$$

$$a_i = \sum_j e_{ij} = \frac{1}{2M} \sum_{i \in V_l} \sum_{j \in V} A(i, j)$$

As in the case of unipartite networks, if edge connections are made at random, we would have $e_{ij} = a_i a_j$. Our new bipartite modularity Q_B is defined as follows:

$$Q_B = \sum_i (e_{ij} - a_i a_j), \quad j = \underset{k}{\mathrm{argmax}}(e_{ik})$$

Newman's modularity measures the fraction of the edges in the network that connect vertices within the same community minus the expected value of the same quantity in a network with the same community divisions but random connection between vertices. Our new bipartite modularity measures the fraction of the edges in the bipartite network that connect vertices of the corresponding X-vertex communities and Y-vertex communities minus the expected value of the same quantity with random connections between X-vertices and Y-vertices.

4. Experiments

As an example for comparing different bipartite modularities, southern women dataset is used. Davis et al. (Davis et al., 1941) collected the data around Mississippi during the 1930s as part of an extensive study of class and race in the Deep South. The dataset describes the participation of 18 women in 14 social events. The women and social events constitute a bipartite network whose vertices are women and social events, and whose edges are the participation in the events.

Experiments of network divisions by the following strategies are performed : (1) optimization of Guimera's bipartite modularity (M_B), (2) optimization of Guimera's bipartite modularity for the other vertex type (M'_B), and (3) optimization of our new bipartite modularity (Q_B). In addition, (4) results of Barber's BRIM algorithm are also discussed later.

As an initial state of the network division, each woman/event is assigned to its own community. Then greedy searches for the optimization of bipartite modularities ((1), (2), (3)) are performed by merging a pair of women/event communities. The results of network divisions by the above strategies are shown in Table 1. Each row of the Table shows the number of discovered communities, values of M_B, M'_B, and Q_B, respectively. Each column shows the strategies (1), (2), and (3), respectively. Although communities are surely obtained with strategy (1), its division is good only for Guimera's bipartite modularity for one vertex type. Strategy (2) does not work for network division. We performed an additional experiment that combines the strategies (1) and (2), but its result is not better than the result of strategy (1). The result of strategy (3) shows that the obtained network division is good for Q_B, of course, and also for M'_B, although its M_B value is worse than the result of strategy (1). This means that our new bipartite modularity is appropriate for obtaining good network divisions from the viewpoint of connectivities of both vertex types, as well as the degree of correspondence between the communities of different types, which is our main objective.

Table 1. Communities from Southern Women Network

	(1)	(2)	(3)
number of communities	13	32	4
M_B	0.140	0.000	0.0025
M'_B	-0.00797	0.000	0.0109
Q_B	0.354	0.138	0.575

According to Barber's paper, strategy (4) (optimization of Barber's bipartite modularity) results in the discovery of coarse division composed of only two communities. However, the division is obtained as the results of 500,000 trials from random community assignment as its initial state.

5. Conclusion

A new bipartite modularity for community extraction from bipartite networks is proposed in this paper. Each bipartite modularity is based on different assumption on the communities in bipartite networks. Our bipartite modularity is for communities of one-to-many correspondence, and it is more flexible than other bipartite modularities.

References

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, *76*, 1–9.

Davis, A., Gardner, B. B., & Gardner, M. R. (1941). *Deep south.* Chicago: University of Chicago Press.

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

Guimera, R., Sales-Pardo, M., & Amaral, L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review*, *76*, 1–8.

Murata, T. (2009). Modularities for bipartite networks. *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia.*

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*, 1–15.