Detecting Communities from Social Tagging Networks Based on Tripartite Modularity

Tsuyoshi Murata
Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology
Tokyo, Japan
murata@cs.titech.ac.jp

Abstract

Online social media such as delicious and digg are represented as tripartite networks whose vertices are users, tags, and resources. Detecting communities from such tripartite networks is practically important. Newman-Girvan modularity is often used as the criteria for evaluating the goodness of network divisions into communities. Murata has extended Newman-Girvan modularity in order to evaluate the quality of the division of tripartite networks. This paper shows the results of community detection from large-scale real social tagging networks based on Murata's tripartite modularity.

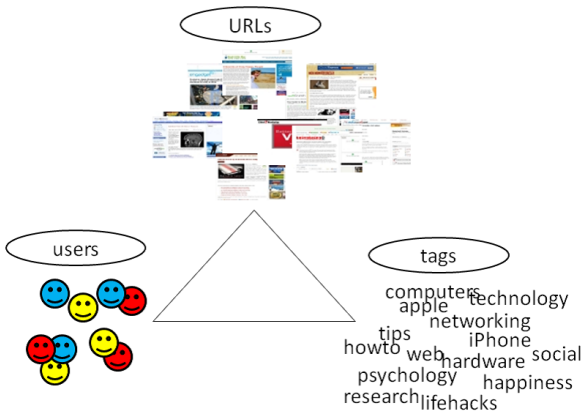

Figure 1: Social Media as a Tripartite Network

1 Introduction

Relations among real-world entities are often represented as n-partite networks that are composed of n types of vertices. Paper-author networks and event-attendee networks are the examples of bipartite networks, and user-resource-tag networks of social tagging systems are the examples of tripartite networks. Detecting communities (subnetworks that are densely connected inside and sparsely connected outside) from such n-partite networks is practically important for finding similar entities and understanding the structure of social media. (Figure 1)

As a naive approach for transforming n-partite networks into unipartite networks, projection is often employed for the sake of convenience. However, it is pointed out that qualities of the communities obtained from projected networks are worse than those from original non-unipartite networks [Guimera *et al.*, 2007].

As a metric for evaluating the goodness of detected communities, Newman-Girvan modularity [Newman and Girvan, 2004] is often employed. Optimizing the modularity is one of the popular strategies for detecting communities from networks. Since it is not suitable for n-partite networks, some researchers extend its definition for bipartite networks, such as the definitions given by Barber[Barber, 2007], Guimera[Guimera *et al.*, 2007], Murata[Murata, 2009] and Suzuki[Suzuki and Wakita, 2009].

Defining suitable tripartite modularity and optimizing it for detecting communities are practically important for the networks of social tagging systems, which are composed of users, resources and tags. As an attempt to extend modularity for tripartite networks, Neubauer proposes a tripartite modularity [Neubauer and Obermayer, 2009] based on Murata's bipartite modularity. His approach is to project a tripartite network into three bipartite networks and then apply Murata's bipartite modularity. However, Neubauer's tripartite modularity still needs projection, and projection will lose some of the information that original tripartite network has. Murata therefore proposes a new tripartite modularity for tripartite networks[Murata, 2010a][Murata, 2010b], which will be explained in Section 2.2. There are some other attempts for detecting communities from tripartite networks[Neubauer and Obermayer, 2010][Ghosh *et al.*, 2011].

In general, detecting communities from real tripartite networks is computationally expensive. This paper employs an approximate method for optimizing Murata's tripartite modularity. Our method employs spectral partitioning, and it has abilities of detecting communities from tripartite networks that are composed of thousands of nodes and tens of thousands of hyperedges. The contribution of this paper is that optimization of the tripartite modularity is attempted for real tripartite networks which are much larger than the ones used in previous research.

2 Related Works

2.1 Community Detection from Heterogeneous Networks

Several attempts have been made for detecting communities from heterogeneous networks. For example, Lin et al. propose MetaFac[Lin *et al.*, 2009], an algorithm for community detection based on tensor decomposition. Sun et al. proposes NetClus, an algorithm for clustering star networks[Sun *et al.*, 2009]. Tang et al. propose an algorithm for clustering based on evolutionary clustering [Tang *et al.*, 2008].

Our approach is based on modularity optimization, which is one of the most popular methods for community detection from unipartite networks. If we can define suitable modularity for heterogeneous networks, the know-hows of modularity optimization can be used for heterogeneous networks. In addition to that, our approach is different from the above approaches in that each community is composed of the vertices of the same type. Correspondences of the communities of different vertex types will give insights to the structures of heterogeneous networks.

2.2 Modularity

Newman-Girvan modularity [Newman and Girvan, 2004] is a quantitative measurement for the quality of a particular division of unipartite network. Let us consider a particular division of a network into k communities. Let us suppose 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 whose (i,j) element is equal to 1 if there is an edge between vertices i and j, and is equal to 0 otherwise. Then we can define e_{lm}, the fraction of all edges in the network that connect vertices in community l to vertices in community m:

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

We further define a $k \times k$ symmetric matrix E composed of e_{lm} as its (l,m) element, and its row sums a_l:

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

If a network had edges between vertices regardless of the communities they belong to, we would have $e_{lm} = a_l a_m$ for this network. Newman-Girvan modularity is thus defined as follows:

$$Q = \sum_l (e_{ll} - a_l^2) \tag{3}$$

2.3 Murata's Tripartite Modularity

Murata defines a new tripartite modularity [Murata, 2010a][Murata, 2010b] in a way that the correspondence of the communities of three vertex types are clearly indicated.

Let us suppose that a tripartite network G is described as (V,E), where V is a set of vertices, and E is a set of hyperedges. V is composed of three types of vertices: V^X, V^Y, and V^Z. A hyperedge connects triples of the vertices (i,j,k), where $i \in V^X$, $j \in V^Y$, and $k \in V^Z$, respectively. Suppose that $deg(i)$ is the number of hyperedges that connect to vertex i.

$A(i,j,k)$ is an adjacency matrix for a tripartite network. The element $A(i,j,k)$ of the adjacency matrix is 1 if vertices i, j, and k are connected with a hyperedge, otherwise it is 0. A community in a tripartite network is defined as a subset of vertices of a single type in this paper, although Barber defines it as a subset of all types of vertices. We employ the above definition since there can be one-to-many correspondence among the communities of different vertex types.

M is the number of hyperedges in a tripartite network, and that V is a set of all vertices in the tripartite network. Consider a particular division of the tripartite network into X-vertex communities, Y-vertex communities, and Z-vertex communities, and the numbers of the communities are L^X, L^Y, and L^Z, respectively. V^X, V^Y, and V^Z are the sets of the communities of X-vertices, Y-vertices, and Z-vertices, and V_l^X, V_m^Y, and V_n^Z are the individual communities that belong to the sets ($V^X = \{V_1^X, ..., V_{L^X}^X\}$, $V^Y = \{V_1^Y, ..., V_{L^Y}^Y\}$, $V^Z = \{V_1^Z, ..., V_{L^Z}^Z\}$). E^{XY}, E^{YZ}, and E^{ZX} are the sets of the edges that connect vertex pairs (X and Y), (Y and Z), and (Z and X), respectively. The number of edges in these sets are equal. ($|E^{XY}| = |E^{YZ}| = |E^{ZX}|$)

Under the condition that the vertices of V_l^X, V_m^Y, and V_n^Z are of different types, we can define e_{lmn} (the fraction of all edges that connect vertices in V_l^X, V_m^Y, and V_n^Z) and its sums over three dimensions, such as a_l, a_m, and a_n.

$$e_{lmn} = \frac{1}{M} \sum_{i \in V_l^X} \sum_{j \in V_m^Y} \sum_{k \in V_n^Z} A(i,j,k) \tag{4}$$

$$a_l^X = \sum_m \sum_n e_{lmn}$$

$$= \frac{1}{M} \sum_{i \in V_l^X} \sum_{j \in V^Y} \sum_{k \in V^Z} A(i,j,k) \tag{5}$$

(6)

a_m^Y and a_n^Z are defined in the same manner. Suppose $s_X = \sum_l a_l^X$, $s_Y = \sum_m a_m^Y$, and $s_Z = \sum_n a_n^Z$. From the above definitions, it is obvious that $s_X = s_Y = s_Z = \sum_l \sum_m \sum_n e_{lmn} = 1$. As in the case of unipartite networks, if hyperedge connections are made at random, we would have $e_{lmn} = a_l^X a_m^Y a_n^Z$. Therefore, $Q_l^X = \sum_m \sum_n (e_{lmn} - a_l a_m a_n)$, where $m, n = \underset{j,k}{\operatorname{argmax}}(e_{ljk})$, will be zero. On the other hand, if hyperedges from X-vertices are mainly from the vertices in community V_l^X, the value of Q_l^X will be greater than zero. The sum over all communities of V^X is as follows.

$$Q^X = \sum_l Q_l^X$$

$$= \sum_l \sum_m \sum_n (e_{lmn} - a_l a_m a_n) \tag{7}$$

$$m, n = \underset{j,k}{\operatorname{argmax}}(e_{ljk})$$

Q^X means the deviation of the number of hyperedges that connect l-th X-vertex community and the corresponding (m-th) Y-vertex community and (n-th) Z-vertex community, from

the expected number of randomly-connected hyperedges. A larger Q^X value means stronger correspondence from the l-th community to the m-th Y-vertex community and the n-th Z-vertex community. Q^Y and Q^Z are defined in the same manner.

Murata's tripartite modularity Q_M is defined as the average of Q^X, Q^Y and Q^Z.

$$Q_M = \frac{1}{3}(Q^X + Q^Y + Q^Z) \tag{8}$$

The main advantages of Murata's new tripartite modularity over Neubauer's tripartite modularity are: 1) The former does not employ projection, and 2) the former can be extended to n-partite modularity.

3 Experiments

Community detection based on optimizing modularity is often employed for unipartite networks. However, optimizing modularity is computationally expensive in general. Therefore, several approaches (such as greedy techniques, simulated annealing, external optimization, spectral optimization, and so on) have been proposed for optimization [Fortunato, 2010].

Optimizing tripartite modularity is more computationally expensive since the partition of only one vertex type affects the goodness of overall partition. In order to optimize the above tripartite modularity, the following approximate method (Figure 2) is employed in our experiment.

function $MaxQ_M$:
var $A(i,j,j)$: adjacency matrix;
$maxn$: maximum number of division;
begin
% similarity matrices of V^X, V^Y, and V^Z
$A^X_{i,j} := |\Gamma(V^X_i)| \cup |\Gamma(V^X_j)|$
$A^Y_{i,j} := |\Gamma(V^Y_i)| \cup |\Gamma(V^X_j)|$
$A^Z_{i,j} := |\Gamma(V^Z_i)| \cup |\Gamma(V^X_j)|$
$maxqm := 0$;
(L^X, L^Y, L^Z):= (1,1,1);
repeat
% spectral partitioning
bipartition V^X based on $A^X_{i,j}$ (L^X times)
bipartition V^Y based on $A^Y_{i,j}$ (L^Y times)
bipartition V^Z based on $A^Z_{i,j}$ (L^Z times)
compute Q_M for the above partition
if $maxqm < Q_M$ **then** $maxqm := Q_M$
increment (L^X, L^Y, L^Z);
until $(L^X, L^Y, L^Z) > (maxn, maxn, maxn)$;
$MaxQ_M := maxqm$
end:

Figure 2: Algorithm for optimizing tripartite modularity

1. Similarity matrices ($A^X_{i,j}$, $A^Y_{i,j}$, and $A^Z_{i,j}$) are generated from given tripartite network based on (extended) common neighbors.
2. Then each vertex set is divided into communities using spectral partitioning.
3. Tripartite modularity is computed for the division.
4. The above procedure is repeated for each L^X, L^Y, and L^Z. At the present stage, Q_M is computed for every combinations of L^X, L^Y, and L^Z. The maximum tripartite modularity is returned as the final result.

The reason for employing spectral partitioning is that it is a divisive approach and relatively faster than agglomerative ones. Other fast method (such as [Clauset *et al.*, 2004]) can be used for this procedure.

We use the data [Wetzker *et al.*, 2008] of delicious, a popular social tagging system that allows users to collaboratively tag resources in the form of URLs. The 10000 tag assignments posted on September 2003 are used for our experiments. The number of users (X), URLs (Y) and tags (Z) are 820, 4750, and 2417, respectively.

In our experiment, each vertex set (V^X, V^Y, and V^Z) are divided into L^X, L^Y, L^Z communities, and the tripartite modularity (Q_M) for the division is calculated. The numbers of communities (L^X, L^Y, L^Z) are set from $(1,1,1)$ to $(15,15,15)$. Figure 3 shows the average values of Q_M for each L^X, L^Y, and L^Z.

Tripartite modularity (Q_M) takes its maximum value when (L^X, L^Y, L^Z) are $(3,9,2)$. We therefore set the numbers of suitable communities for users (X), URLs (Y), and tags (Z) as 3, 9, and 2, respectively.

Since users (X) are anonymized in the dataset, and there are only two tag communities (Z), we show some examples of the terms contained in URL communities (Y). The sizes of nine URL communities are 211, 1433, 1154, 489, 71, 188, 464, 300, and 440. Although characterizing all these communities are not easy, some communities are surely characteristic. For example, technical terms frequently appear in the URLs of communities 4 and 8, and terms of shopping and entertainments appear in the URLs of communities 2 and 5.

com 1 (211) de, uk, ca,...

com 2 (1433) burkesbackyard, milkandcookies, article...

com 3 (1154) uk, fr, it,...

com 4 (489) codeproject, java, linux,...

com 5 (71) links, milkandcookies, shiptheweb...

com 6 (188) de, fr, ru...

com 7 (464) library, research, article...

com 8 (300) sourceforge,wiki, blog...

com 9 (440) circuits, electronic, filter...

As far as the author knows, this is one of the first attempts for detecting communities from large-scale real tripartite networks. Most of the previous research [Neubauer and Obermayer, 2010][Ghosh *et al.*, 2011] use small-size synthetic tripartite networks that are composed of at most hundreds of nodes. Scalability is quite important for community detection since there are many real large-scale heterogeneous networks.

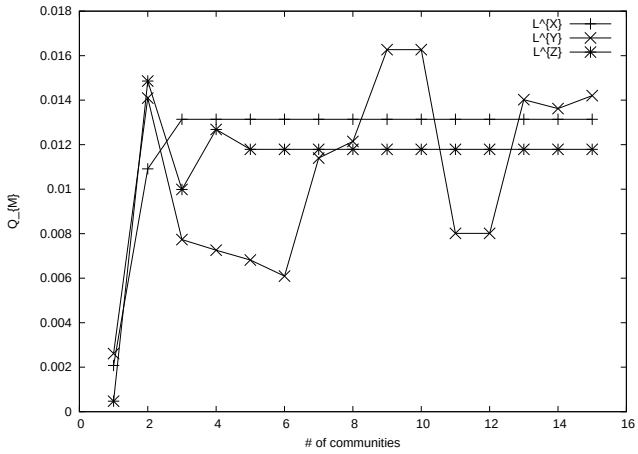

Figure 3: Users(X),URLs(Y),Tags(Z), and their tripartite modularities

4 Concluding Remarks

This paper explains previous tripartite modularity and shows the results of community detection using the data of real tripartite networks. The result shown in this paper is the first step for processing real heterogeneous networks that are available in the social networks. Murata's tripartite modularity is based on an assumption that a community in certain vertex type corresponds to one or more communities in other vertex types. However, this assumption may not be true for some synthetic heterogeneous networks. Possibilities and limitations of our method has to be analyzed in detail as our future research.

References

[Barber, 2007] Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(066102):1–9, 2007.

[Clauset *et al.*, 2004] Aaron Clauset, Mark E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(066111):1–6, 2004.

[Fortunato, 2010] San Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[Ghosh *et al.*, 2011] Saptarshi Ghosh, Pushkar Kane, and Niloy Ganguly. Identifying overlapping communities in folksonomies of tripartite hypergraphs. In *Proceedings of the 20th International World Wide Web Conference (WWW2011)*, 2011.

[Guimera *et al.*, 2007] Roger Guimera, Marta Sales-Pardo, and Luis A. Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review*, 76(036102):1–8, 2007.

[Lin *et al.*, 2009] Yu-Ru Lin, Jimeng Sun, Paul Castro, Ravi Konuru, Hari Sundaram, and Aisling Kelliher. Metafac: Community disocovery via relational hypergraph factorization. In *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 527–535, 2009.

[Murata, 2009] Tsuyoshi Murata. Detecting communities from bipartite networks based on bipartite modularities. In *Proceedings of the 2009 IEEE International Conference on Social Computing (SocialCom-09)*, pages 50–57, 2009.

[Murata, 2010a] Tsuyoshi Murata. Detecting communities from tripartite networks. In *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, pages 1159–1160, 2010.

[Murata, 2010b] Tsuyoshi Murata. Modularity for heterogeneous networks. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia (HyperText2010)*, pages 129–134, 2010.

[Neubauer and Obermayer, 2009] Nicolas Neubauer and Klaus Obermayer. Towards community detection in k-partite k-uniform hypergraphs. In *Proceedings of the NIPS 2009 Workshop on Analyzing Networks and Learning with Graphs*, 2009.

[Neubauer and Obermayer, 2010] Nicolas Neubauer and Klaus Obermayer. Community detection in tagging-induced hypergraphs. In *Workshop on Information in Networks*, 2010.

[Newman and Girvan, 2004] Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(026113):1–15, 2004.

[Sun *et al.*, 2009] Yizhou Sun, Yintao Yu, and Jiawei Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–805, 2009.

[Suzuki and Wakita, 2009] Kenta Suzuki and Ken Wakita. Extracting multi-facet community structure from bipartite networks. In *Proceedings of the International Symposium on Social Intelligence and Networking (SIN09)*, pages 312–319, 2009.

[Tang *et al.*, 2008] Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 677–685, 2008.

[Wetzker *et al.*, 2008] Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *In Proceedings of Mining Social Data (MSoDa) Workshop, pp. 26–30*, 2008.