Community Division of Heterogeneous Networks

Tsuyoshi Murata

Tokyo Institute of Technology
2-12-1 W8-59 Ookayama, Meguro, Tokyo, 152-8552 Japan
murata@cs.titech.ac.jp

Abstract. Many real world data can be represented as heterogeneous networks that are composed of more than one types of nodes, such as paper-author networks (two types) and user-resource-tag networks (three types) of social tagging systems. Discovering communities from such heterogeneous networks is important for finding similar nodes, which are useful for information recommendation and visualization. Although modularity is a famous criterion for evaluating division of given networks, it is not applicable to heterogeneous networks. This paper proposes new modularity for bipartite networks, as the first step for heterogeneous networks. Experimental results using artificial networks and real networks are shown.

Keywords: modularity, community, bipartite networks.

1 Introduction

There are various relations among entities that are represented as networks, such as citations of papers and friendships of SNS (Social Network Service) users. In order to recognize their overall structures and to recommend similar entities, discovering communities from networks attracts many researchers from physics, computer science, and sociology. Quality of the divisions of networks is often measured by modularity, which is a scalar value that measure the density of edges inside communities as compared to edges between communities. As the strategy for finding divisions of given networks, modularity optimization is often employed.

Modularity is, however, appropriate for homogeneous networks that are composed of only one type of vertices (such as papers and SNS users in the above examples). In real-world situations, there are many heterogeneous networks that are composed of more than one types of vertices, such as paper-author networks and movie-actor networks (Figure 1). Modularity is not appropriate for community division of such heterogeneous networks since the density of edges inside communities of same type of vertices is sparser than that of edges between communities of different types of vertices.

As the first step for generalizing the definition of modularity, this paper proposes a new definition of modularity for bipartite networks, which we call bipartite modularity. As far as the author knows, this is the first attempt for

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 1011–1022, 2009.

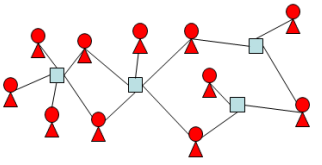

Fig. 1. Example of Bipartite Network

generalizing the definition of modularity for bipartite networks. Experimental results using artificial network data show that our bipartite modularity can clearly detect the existence of community structures of bipartite networks rather than original modularity. Another experiments using real network data show that bipartite modularity is useful also for characterizing each community.

The structure of this paper is as follows: related works about modularity and bipartite networks are reviewed in Section 2. Definition of our new modularity for bipartite networks is shown in Section 3. Experiments using artificial and real network data are shown in Section 4, followed by conclusion in Section 5.

2 Related Work

In this section, the definition of Newman's modularity [9] is reviewed as the basis of the following discussion. Research on bipartite networks are also described.

2.1 Modularity

Modularity is a quantitative measurement for the quality of a particular division of a 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:

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

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 $e_{ij} = a_i a_j$. Therefore modularity is defined as follows:

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

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. If the number of edges inside communities is no better than random, we will get $Q = 0$. Values approaching the maximum ($Q = 1$) indicate strong community structures.

There are many related work regarding modularity. Clauset [3] proposes fast modularity algorithm for efficient search for division of high modularity. Newman [10] propose a spectral algorithm for improving the quality of community division. Wakita [13] and Blondel [2] attempt the division of large-scale networks. Danon [4] performs comparison of several network division methods. Fortunato [6] clarifies resolution limits of modularity-based network division methods.

2.2 Research on Bipartite Networks

Although most of the research for social networks focus on homogeneous networks, the following research focus on bipartite networks.

Neighborhood Formation. Sun [11] proposes algorithms for computing the neighborhood of the nodes of bipartite networks using random walk with restarts and network partitioning. Algorithms for identifying abnormal nodes are also proposed, and their effectiveness and efficiency are confirmed by the experiments on several real datasets.

Co-ranking. Zhou [15] proposes a framework for co-ranking authors and documents in heterogeneous networks. The framework is based on coupling two random walks that separately rank authors and documents. As the result of the coupling, both document ranking and author ranking are improved since both ranking depend on each other in a mutually reinforcing way.

Projection. Zhou [16] proposes a method for projecting bipartite networks to weighted homogeneous networks. Bipartite networks are regarded as resource allocation processes between X-vertices and Y-vertices. Initially assigned weights on X-vertices are propagated to Y-vertices and then back to X-vertices in order to obtain weighted homogeneous networks.

Community variance. Murata [8] proposes a criterion for evaluating correspondence between communities of two types of vertices. Although the criterion (community variance) is useful for clarifying structural properties of communities, it has no relation with modularity.

Although the goals of these research are different from ours, these research put stress on the importance of processing bipartite networks appropriately. Our goal in this paper is to propose new criterion for evaluating division of bipartite networks.

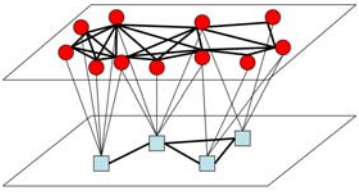

Fig. 2. Projection of Bipartite Network

3 Bipartite Modularity

3.1 Bipartite Networks

In general, social networks can be divided into the following categories: 1) direct connection between persons (such as MySpace or Twitter) and 2) indirect connection through different types of entities (such as film co-starring or paper co-authoring). We call the former "homogeneous networks", and the latter "heterogeneous networks". There are bipartite, tripartite and n-partite networks as the examples of heterogeneous networks. As the first step for processing heterogeneous networks, we focus on bipartite networks composed of two types of vertices. Zhou [16] claims that there are two types of bipartite networks: collaboration network and opinion network. The former is generally defined as a network of users connected by common collaboration acts. The latter is defined as a network of users connected by common objects.

As a naive approach for transforming bipartite networks into homogeneous networks, projection is often used. Suppose a bipartite network is composed of X-vertices $\{x_0, x_1, ...\}$ and Y-vertices $\{y_0, y_1, ...\}$, and y_i is connected to both x_j and x_k. Projection is a transformation of such $x_j - y_i - x_k$ connection into $x_j - x_k$ connection so that a network composed of only X-vertices is obtained. However, projection loses information about the correspondence between X-vertices and Y-vertices, which is often quite valuable.

3.2 Bipartite Modularity

In the case of community division of bipartite networks, finding the correspondence between communities of different types of vertices is often important. Let us suppose that communities of papers and communities of authors are discovered 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 3). 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 4).

In order to evaluate the division of such bipartite networks, we define modularity for bipartite network, which we call bipartite modularity. Bipartite modularity is for measuring the degree of correspondence between communities of different types of vertices.

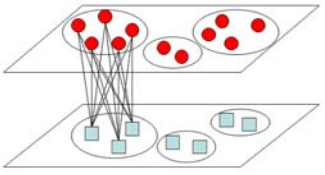

Fig. 3. One-to-one Correspondence between Communities

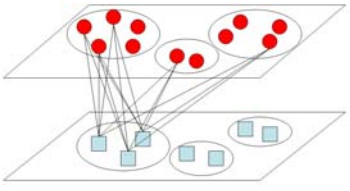

Fig. 4. One-to-many Correspondence between Communities

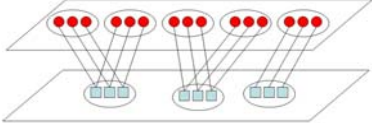

Fig. 5. Communities in a Bipartite Network

Newman's modularity is not appropriate for evaluating community divisions of bipartite networks. Let us suppose that a bipartite network composed of X-vertices and Y-vertices is given, and both X-vertex communities and Y-vertex communities are specified. Since bipartite network does not have any direct edge between X-vertices (and between Y-vertices), $e_{ii} = 0$ for all X-vertex (Y-vertex) communities V_i, and modularity for community division is quite low. For example, modularity of the community division of the bipartite network shown in Figure 5 is -0.14.

Our definition of bipartite modularity is as follows. 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, and the numbers of the communities are L^+ and L^-, respectively. V^+ and V^- are the sets of the communities of X-vertices and Y-vertices, and V_l^+ and V_m^- are the individual communities that belong to the sets ($V^+ = \{V_1^+, ..., V_{L^+}^+\}$, $V^- = \{V_1^-, ..., V_{L^-}^-\}$). $A(i, j)$ 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 (which means $(V_l \in V^+ \wedge V_m \in V^-) \vee (V_l \in V^- \wedge V_m \in V^+)$), 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 homogeneous networks, if edge connections are made at random, we would have $e_{ij} = a_i a_j$. Bipartite modularity Q_B is defined as follows:

$$Q_B = \sum_i (e_{ij} - a_i a_j), \quad j = \operatorname*{argmax}_k (e_{ik})$$

As shown in section 2.1, original 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. 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. If given network is not bipartite, you can see that $Q_B = Q$, which means that bipartite modularity is a straightforward generalization of original modularity.

If the connection between X-vertices and Y-vertices is no better than random, we will get $Q_B = 0$. High Q_B value indicates strong community structure in a bipartite network. Bipartite modularity of the network shown in Figure 5 is 0.66. If you take a closer look at the expression of Q_B, you will find that the value is the sum of bipartite modularities of different directions ($V^+ \to V^-$ and $V^- \to V^+$). Q_B can be divided as follows:

$$Q_{B\pm} = \sum_{i \in V^+} (e_{ij} - a_i a_j), \quad j = \operatorname*{argmax}_{k \in V^-} (e_{ik})$$

$$Q_{B\mp} = \sum_{i \in V^-} (e_{ij} - a_i a_j), \quad j = \operatorname*{argmax}_{k \in V^+} (e_{ik})$$

$$Q_B = Q_{B\pm} + Q_{B\mp}$$

$Q_{B\pm}$ is the bipartite modularity for $V^+ \to V^-$ direction, and $Q_{B\mp}$ is the bipartite modularity for $V^- \to V^+$ direction. In the example shown in Figure 5, $Q_{B\pm} = 0.41$ and $Q_{B\mp} = 0.25$, which means that downward connections are relatively focused rather than upward connections in the figure.

The matrix E composed of e_{ij} as its (i,j) element is represented as follows if rows and columns are reordered appropriately.

$$\left(\begin{array}{ccc|ccc} 0 & \cdots & 0 & e_{1,L^+ +1} & \cdots & e_{1,L^+ +L^-} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & e_{L^+,L^+ +1} & \cdots & e_{L^+,L^+ +L^-} \\ \hline e_{L^+ +1,1} & \cdots & e_{L^+ +1,L^+} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ e_{L^+ +L^-,1} & \cdots & e_{L^+ +L^-,L^+} & 0 & \cdots & 0 \end{array}\right)$$

The upper right quarter of the matrix (E_{UR}) corresponds to $Q_{B\pm}$, and the lower left quarter of the matrix (E_{LL}) corresponds to $Q_{B\mp}$. Since E is a symmetric matrix, it is clear that $E_{UR}^T = E_{LL}$. But $Q_{B\pm} \neq Q_{B\mp}$ in general. This is because a set of (i, j) under the condition that $i \in V^+, j = \underset{k \in V^-}{\text{argmax}}(e_{ik})$ is different from a set of (i, j) under the condition that $i \in V^-, j = \underset{k \in V^+}{\text{argmax}}(e_{ik})$.

When two upper-left communities in Figure 5 are merged, Q_B increases to 0.67. But if all upper communities are merged into one community, Q_B decrease to 0.35. By maximizing bipartite modularity, unobvious community structure will be obtained from bipartite networks.

4 Experiments

4.1 Artificial Four-Community Networks

In order to clarify the properties of our bipartite modularity, modularity and bipartite modularity are compared in the following experiments. Networks with known community structure are used to see whether our bipartite modularity has abilities of detecting the structure.

We have generated many networks with 128 vertices, divided into four communities of 32 vertices each. Edges are placed independently at random with probability p_{in} for an edge to fall between vertices in the same community and p_{out} to fall between vertices in different communities. Such artificial network data are used by Newman [9] and Danon [4]. Figure 6 illustrates an example of the networks. Figure 7 shows the average values of modularity and bipartite modularity of 100 artificial networks.

You can see from the figure that modularity and bipartite modularity are the same for networks of high p_{in}. This is obvious from the definition of bipartite modularity. For the networks with high p_{in}, diagonal elements of matrix E are the biggest among the all elements in the same row ($\forall j \ e_{ii} \geq e_{ij}$). Therefore $j = \underset{k}{\text{argmax}}(e_{ik}) = i$ and $Q_B = Q$.

For networks of smaller p_{in} ($p_{in} < p_{out}$), diagonal elements of matrix E are *not* the biggest ($\exists j \ e_{ii} \leq e_{ij}$) and their modularities are below zero. On the other hand, bipartite modularities of the networks are positive because $j = \underset{k}{\text{argmax}}(e_{ik})$ is set to the community that is densely connected with community i.

The above networks are not bipartite because four communities are connected to each other. For the next experiment, we have generated bipartite networks

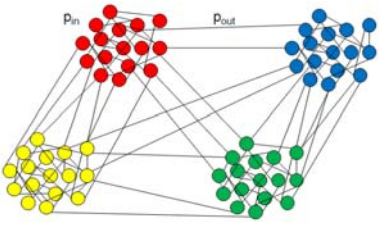

Fig. 6. Network with Four Communities

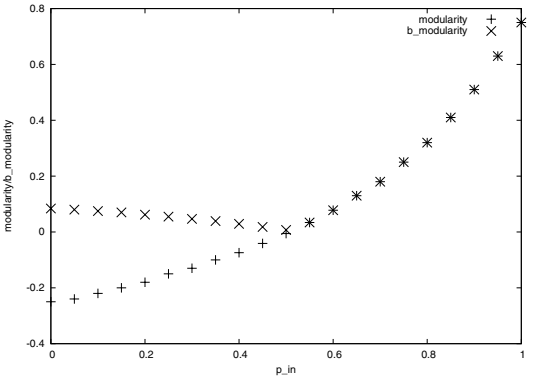

Fig. 7. Modularity and Bipartite Modularity of Four-community Networks

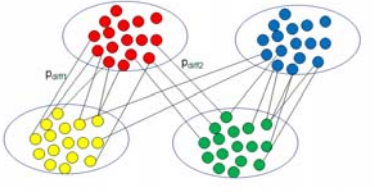

Fig. 8. Bipartite Network with Four Communities

with 128 vertices, divided into four communities of 32 vertices each. Edges are placed independently at random with probability p_{in} for an edge to fall between vertices in the same community, p_{same} to fall between vertices in the communities of same type of vertices, and p_{diff1} $p_{diff2}$to fall between vertices in the communities of different types of vertices. Suppose there are two communities for each type of vertices. p_{in} and p_{same} are set to zero because there are no edges between vertices of the same type in bipartite networks. Figure 8 illustrates an example of such networks. Networks with various p_{diff1} and p_{diff2} are generated and their modularity and bipartite modularity are calculated. Figure 9 shows the average values of modularity and bipartite modularity of 100 artificial bipartite networks.

Figure 9 shows that original modularity is not appropriate for bipartite networks because there is no edge between vertices of the same type. Bipartite modularity is effective for detecting the existence of community structures for

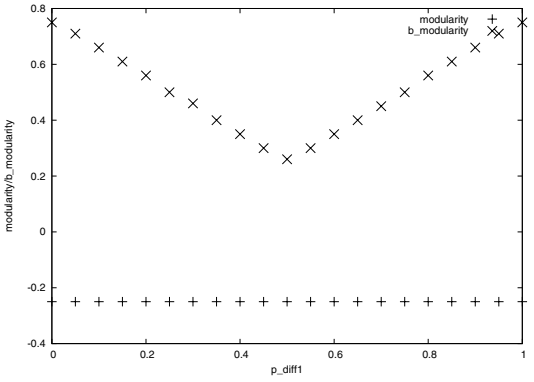

Fig. 9. Modularity and Bipartite Modularity for Bipartite Networks

bipartite networks, and it also shows the degree of correspondence between communities of different types of vertices. In the case of networks with $p_{diff1} = 1.0$ or $p_{diff2} = 1.0$, there are complete one-to-one correspondence between communities of different types of vertices, and the values of bipartite modularity are the highest.

4.2 Real Online Social Networks

Bipartite modularity described above is applied also for real-world networks. We have generated heterogeneous social networks composed of users and boards from the data of Yahoo! Chiebukuro (Japanese Yahoo! Answers, http://chiebukuro.yahoo.co.jp). The site is one of the most popular question-answering forums in Japan. The network of Yahoo! Chiebukuro is summarized in Table 1. From the heterogeneous networks, user communities and board communities are discovered by 1) projecting the heterogeneous networks into homogeneous ones (user networks and board networks), and 2) applying Clauset's fast modularity algorithm [3] for finding community divisions of high modularities. Details of the discovery method is described in [8]. Both user communities and board communities are extracted from the network data. Bipartite modularity and original modularity of the division are as follows. Modularity (Q) and bipartite modularity (Q_B) of the network are -0.1021 and 0.2919, respectively. This shows that the

Table 1. Statistics of the Network of Yahoo! Chiebukuro

number of vertices	6,309,737
number of edges	357,834
average degree	0.1134
clustering coefficient	0
average path length	7.7587

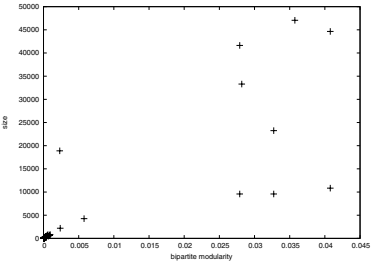

Fig. 10. Bipartite Modularities and the Sizes of Discovered Communities

community division is not good in the sense of homogeneous network division, but it is not bad in the sense of bipartite network division.

Bipartite modularity is for evaluating the whole division of a bipartite network into communities. In addition to that, bipartite modularity of each community (Q_{B_i}) can be used for measuring the degree of "close-knitness" to the communities of the other type of vertices. Figure 10 shows a distribution of bipartite modularity Q_{B_i} (X-axis) and the sizes (Y-axis) of discovered communities. Communities of upper half of the figure (more than 15,000 vertices) are board communities. Their bipartite modularities are high except the one located at middle left position. This community is like the one in Figure 4: the main topics of the community (such as "entertainment and hobby" and "health and fashion") attract many users and thus its bipartite modularity is low. On the other hand, other communities of high bipartite modularity are relatively focused (such as "child care", "mental health", and "cars"), like the one in Figure 3.

This paper focus on the criterion for evaluating given community division for bipartite networks. Finding the best community division is NP-complete [7]. In the case of homogeneous networks, many approaches are proposed for finding appropriate division with the smallest computational cost possible by modularity optimization. Bipartite modularity can be used in the same manner for finding and evaluating division of bipartite networks.

Biclustering algorithms [7][12][5] also aim at finding division of incident matrices. These algorithms are mainly for the purpose of bioinformatics and document clustering, and the size of the incident matrices are at most thousands times tens of thousands. One of the weakness of these algorithms is that most of these algorithms do not scale to large networks. Finding community division of high bipartite modularity from given large-scale networks is another challenging research topic, which is left for our future work.

5 Conclusion

This paper proposes a new criterion for community division of bipartite networks. As far as the author knows, this is the first attempt for generalizing the definition of modularity for bipartite networks. Bipartite modularity is a straightforward generalization of Newman's modularity. Experimental results show that

our bipartite modularity is appropriate for detecting the existence of community structures from bipartite networks. In addition to that, bipartite modularity for each community is the degree of correspondence to the communities of the other type of vertices, which can be used for analyzing the characteristics of the community.

Bipartite modularity proposed in this paper is the first step for intelligent processing of heterogeneous networks in the Web. There are several bipartite, tripartite, and n-partite networks in the Web. Social tagging systems can be represented as tripartite networks composed of three types of vertices (users, URLs and tags). Discovering and evaluating communities of such heterogeneous networks is one of the important and challenging topics of Web mining.

References

1. Adamic, L.A., Zhang, J., Bakshy, E., Ackerman, M.S.: Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In: Proceedings of the 17th International World Wide Web Conference, pp. 665–674 (2008)
2. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast Unfolding of Community Hierarchies in Large Networks, 1–6 (2008) arXiv:0803.0476v1
3. Clauset, A., Newman, M.E.J., Moore, C.: Finding Community Structure in Very Large Networks. Physical Review E 70, 066111, 1–6 (2004)
4. Danon, L., Diaz-Guiler, A., Duch, J., Arenas, A.: Comparing Community Structure Identification. Journal of Statistical Mechanics, P09008, 1–10 (2005)
5. Dhillon, I.S.: Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 269–274 (2001)
6. Fortunato, S., Barthelemy, M.: Resolution Limit in Community Detection. Proceedings of the National Academy of Sciences of the United States of America 104(1), 36–41 (2007)
7. Madeira, S.C., Oliveira, A.L.: Biclustering Algorithms for Biological Data Analysis: A Survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics 1(1), 24–45 (2004)
8. Murata, T., Ikeya, T.: Analysis of Online Question-Answering Forums as Heterogeneous Networks. In: Proceedings of the Second International Conference on Weblogs and Social Media, pp. 210–211 (2008)
9. Newman, M.E.J., Girvan, M.: Finding and Evaluating Community Structure in Networks. Physical Review E 69, 026113, 1–15 (2004)
10. Newman, M.E.J.: Modularity and Community Structure in Networks. Proceedings of the National Academy of Sciences of the United States of America 103, 8577–8582 (2006)
11. Sun, J., Qu, H., Chakrabarti, D., Faloutsos, C.: Neighborhood Formation and Anomaly Detection in Bipartite Graphs. In: Proceedings of the Fifth IEEE International Conference on Data Mining, pp. 418–425 (2005)
12. Tanay, A., Sharan, R., Shamir, R.: Discovering Statistically Significant Biclusters in Gene Expression Data. Bioinformatics 18 (suppl.1), S136–S144 (2002)
13. Wakita, K., Tsurumi, T.: Finding Community Structure in Mega-scale Social Networks. In: Proceedings of the 16th International World Wide Web Conference, pp. 1275–1276 (2007)

14. Xi, W., Zhang, B., Chen, Z., Lu, Y., Yan, S., Ma, W.-Y., Fox, E.A.: Link Fusion: A Unified Link Analysis Framework for Multi-Type Interrelated Data Objects. In: Proceedings of the 13th World Wide Web Conference, pp. 319–327 (2004)
15. Zhou, D., Orchanskiy, S.A., Zha, H., Giles, C.L.: Co-Ranking Authors and Documents in a Heterogeneous Network. In: Proceedings of the Seventh IEEE International Conference on Data Mining, pp. 739–744 (2007)
16. Zhou, T., Ren, J., Medo, M., Zhang, Y.-C.: Bipartite network projection and personal recommendation. Physical Review E 76, 046115, 1–7 (2007)