A Comparative Study of Community Structure Based Node Scores for Network Immunization

Yuu Yamada and Tetsuya Yoshida

Graduate School of Information Science and Technology,
Hokkaido University
N-14 W-9, Sapporo 060-0814, Japan
{yamayuu,yoshida}@meme.hokudai.ac.jp

Abstract. Network immunization has often been conducted by removing nodes with large network centrality so that the whole network can be fragmented into smaller subgraphs. Since contamination (e.g., virus) is propagated among subgraphs (communities) along links in a network, besides centrality, utilization of community structure seems effective for immunization. We have proposed community structure based node scores in terms of a vector representation of nodes in a network. In this paper we report a comparative study of our node scores over both synthetic and real-world networks. The characteristics of the node scores are clarified through the visualization of networks. Extensive experiments are conducted to compare the node scores with other centrality based immunization strategies. The results are encouraging and indicate that the node scores are promising.

1 Introduction

Contamination (e.g., virus) is usually propagated among subgraphs (communities) along links in a network. For preventing the spread of contamination over the whole network, it is necessary to remove (or, vaccinate) contaminated nodes. Since contamination is propagated among communities in a network, for effective network immunization, it is important to identify nodes which play the role of intermediating or connecting communities.

Most previous work on network analysis considers the community structure of a network in terms of links in a network (e.g., graph cut) [7]; however, we consider it in terms of nodes in a network, and proposed community structure based node scores for network immunization [11]. Based on a quality measure of communities for node partitioning [4], a vector representation of nodes in a network is constructed, and the community structure in terms of the distribution of node vectors is utilized for calculating node scores. Two types of node score are proposed based on the direction and the norm of the constructed node vectors.

In this paper we report a comparative study of our node scores over both synthetic and real-world networks. The characteristics of node scores are clarified through network visualization, and they are compared with other centrality

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 328–337, 2012.

based immunization strategies. Comparison with other centrality based immunization strategies shows that our node scores are promising, since they can exploit the community structure of a network without relying on the externally supplied community labels of nodes.

Section 2 explains network immunization and centralities. Section 3 describes our community structure based node scores. Section 4 reports a comparative study and discusses the results. Section 5 summarizes our contributions.

2 Network Immunization

2.1 Preliminaries

We use a bold italic lowercase letter to denote a vector, and a bold normal uppercase letter to denote a matrix. $\mathbf{X}_{ij}$ stands for the element in a matrix $\mathbf{X}$, and $\mathbf{X}^T$ stands for the transposition of $\mathbf{X}$. $\mathbf{1}_n \in \mathbb{R}^n$ stands for a vector where each element is 1.

Let n stands for the number of nodes in a network G, and m stands for the number of links in G [1]. Since most social networks are represented as undirected graph without self-loops [6], we focus on this type of networks in this paper.

The connectivity of a network is represented as a square matrix $\mathbf{A} \in \{0,1\}^{n \times n}$, which is called an adjacency matrix. $\mathbf{A}_{ij} = 1$ if the pair of vertices (i,j) is connected; otherwise, 0. For an undirected graph without self-loops, its adjacency matrix $\mathbf{A}$ is symmetric and its diagonal elements are set to zeros.

2.2 Network Immunization

Epidemics (e.g, virus) are often propagated through the interaction between nodes (e.g., individuals, computers) in a network. If a contaminated node interacts with other nodes, contamination can spread over the whole network. In order to protect the nodes in the network as much as possible, it is necessary to disconnect (or, remove) the contaminated node so that the major part of the network. For instance, the largest connected component (LCC) of a network can be prevented from the contamination by removing several nodes (see Fig. 1).

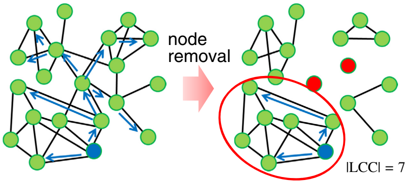

Fig. 1. Network immunization

2.3 Network Centrality

Various notions of "network centrality" have been studied in social network analysis [6,8]. Since nodes with many links can be considered as a hub in a network, the degree (number of links) of a node is called degree centrality. On the other hand, betweenness centrality focuses on the shortest path along which

[1] We also call a network as a graph, a node as a vertex, and a link as an edge.

information is propagated over a network. By enumerating the shortest paths between each pair of nodes, betweenness centrality of a node is defined as the number of shortest paths which go through the node.

Similar to the famous Page Rank, eigenvector centrality utilizes the leading eigenvector of the adjacency matrix $\mathbf{A}$ of a network, and each element (value) of the eigenvector is considered as the score of the corresponding node. Based on the approximate calculation of eigenvector centrality via perturbation analysis, another centrality (called dynamical importance) was also proposed in [9].

By assuming that community labels of nodes in a network can be provided, perturbation analysis of node centrality is utilized for exploiting the relation among communities in [5]. However, although various methods have been proposed for community discovery from networks [8,10], it is still difficult to identify the true community labels.

3 Community Structure Based Node Scores

Our node scores consider the community structure in terms of nodes, not links as in most previous approaches [7]. A vector representation of nodes is constructed based on modularity to reflect the community structure of a network.

3.1 Node Vectors Based on Community Structure

Modularity has been utilized as a standard for community discovery in network analysis[4]. It was shown that the maximization of modularity can be sought by finding the eigenvector for the largest eigenvalue of the following matrix [7]:

$$\mathbf{B} = \mathbf{A} - \mathbf{P} \tag{1}$$

where $\mathbf{P} = \boldsymbol{k}\boldsymbol{k}^T/2m$ for $\boldsymbol{k} = \mathbf{A}\mathbf{1}_n$ ($\boldsymbol{k}$ is the degree vector).

By utilizing several eigenvectors of $\mathbf{B}$ in eq.(1) with several largest positive eigenvalues, the modularity matrix $\mathbf{B}$ can be approximately decomposed as:

$$\mathbf{B} \simeq \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T \tag{2}$$

where $\mathbf{U}=[\boldsymbol{u}_1, \cdots, \boldsymbol{u}_q]$ are the eigenvectors of $\mathbf{B}$ with the descending order of eigenvalues, and $\boldsymbol{\Lambda}$ is the diagonal matrix with the corresponding eigenvalues. Based on eq.(2), the following data representation was proposed [7]:

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}^{1/2} \tag{3}$$

In our approach, the i-th row of $\mathbf{X}$ in eq.(3) is used as the vector representation of the i-th node, and is called as a node vector.

3.2 Inverse Vector Density

A node score was proposed in terms of the mutual angle between node vectors in eq.(3) [11]. The number of “near-by” node vectors is utilized for identifying

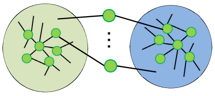

Fig. 2. Synthetic network (CL_*_*_)

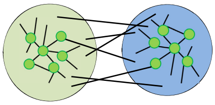

Fig. 3. Synthetic network (CL_*)

border nodes. This node score is defined as:

$$\mathbf{D} = diag(\|\boldsymbol{x}_1\|, \ldots, \|\boldsymbol{x}_n\|) \tag{4}$$

$$\mathbf{X}_1 = \mathbf{D}^{-1}\mathbf{X} \tag{5}$$

$$\boldsymbol{\Theta} = \cos^{-1}(\mathbf{X}_1{\mathbf{X}_1}^T) \tag{6}$$

$$f(\boldsymbol{\Theta}_{ij}, \theta) = \begin{cases} 1 & (\boldsymbol{\Theta}_{ij} < \theta) \\ 0 & (otherwise) \end{cases} \tag{7}$$

$$ivd(\boldsymbol{x}_i) = \frac{1}{\sum_j^n f(\boldsymbol{\Theta}_{ij}, \theta)} \tag{8}$$

where $\mathbf{D}$ in eq.(4) is a diagonal matrix with elements $\|\boldsymbol{x}_1\|, \ldots, \|\boldsymbol{x}_n\|$, and θ is a threshold. The value of $ivd(\cdot)$ in eq.(8) corresponds to the score.

The function f in eq.(7) checks if the angle θ_{ij} in eq.(6) is less than the specified threshold θ. Finally, since border nodes have relatively small number of near-by node vectors, the node score is calculated by taking the inverse of the number of near-by vectors. This node score is called IVD (inverse vector density).

3.3 Community Centrality Based Inverse Vector Density

Removal of hub nodes, which act as mediators of information diffusion over the network, also seems effective for network immunization. However, the node score of a hub node gets rather small with IVD. One of the reasons is that, the direction of each node vector is utilized in IVD, but its norm is not yet utilized.

The square norm of a node vector was regarded to what extent the node is central to a community [7], and was named as community centrality: $cc(\boldsymbol{x}_i)=\boldsymbol{x}_i^T\boldsymbol{x}_i$. This was reflected on IVD, and another node score was proposed as [11]:

$$ccivd(\boldsymbol{x}_i) = cc(\boldsymbol{x}_i) \times ivd(\boldsymbol{x}_i) \tag{9}$$

This is called CCIVD (Community Centrality based Inverse Vector Density).

4 Evaluations

4.1 Experimental Settings

Networks Extensive experiments were conducted over both synthetic and real-world networks. Utilized networks are shown in Table 1 and Table 2.

Table 1. Synthetic networks

dataset	#nodes	#links (ave.)
CL_2_5_1	105	708
CL_3_5_1	165	1064
CL_2	100	355.7
CL_3	150	548.1
CL_4	200	744.3
CL_5	250	931.3

Table 2. Real-world networks

dataset	#nodes	#links
karate	34	78
dolphins	62	159
lesmis	77	254
polbooks	105	441
netscience	379	914
celegansneural	297	2148

When constructing synthetic networks, each component (community) was generated using Barabási-Albert (BA) model [1] by setting the degree distribution $p(k) \propto k^{-3}$, where k denotes the degree of a node. The initial degree was set to 4 in order to generate sparse networks. After generating communities, they were connected either through five intermediate nodes (as illustrated in Fig. 2) or directly with links (Fig. 3). Since synthetic networks are generated based on random networks, we constructed 10 networks for each type and report the average result.

As real world networks, we used three networks in Table 2, which are available as GML (graph markup language) format [2].

Quality Measures. Following the quality measure in [5], the relative size S of the largest connected component (LCC) in a network was measured against the node occupation probability p. After removing some nodes from a network with n nodes, these are calculated as:

$$S = \frac{|\mathrm{LCC}|}{n}, \quad p = \frac{\#remaining\ nodes}{n} \tag{10}$$

where |LCC| is the number of nodes in LCC. The smaller S is, the better a immunization strategy of networks is, since it can prevent the spreading of contamination over the whole network (see Fig. 1).

Compared Methods. For comparison, network immunization based on various concepts of network centrality in Section 2.3 were compared. The node with the maximum centrality was repeatedly selected and removed in each method:

D : degree centrality
B : betweenness centrality
EVC : eigenvector centrality
CC : community centrality
ModC : meta-graph based centrality [5]

In our methods, parameters (q and θ) were set based on preliminary experiments.

[2] `http://www-personal.umich.edu/~mejn/netdata/`

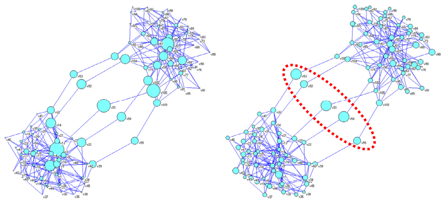

Fig. 4. Visualization result (CL_2_5_1)

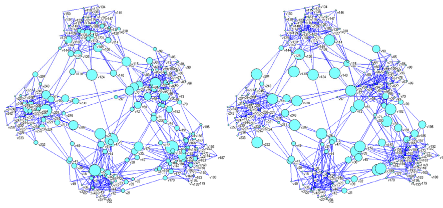

Fig. 5. Visualization result (CL_5)

Immunization Strategies Network immunization was conducted by removing the node with the maximal node score (e.g., centrality). The following strategies were evaluated for the calculation of node scores:

single : scores were calculated only once with respect to the whole network.
recalc : scores were re-calculated when a node is removed from a network.

The strategy recalc can utilize up-to-date node scores even after some nodes are removed, in compensation for the additional computational cost. Since ModC requires re-calculation of centrality, it was not evaluated for single strategy.

4.2 Visualization of Node Scores

In order to verify that our node scores can identify nodes which connect communities in a network, we compared the node scores and the score with **B** (betweenness centrality). The size of each node is depicted proportional to its node score in each method. The results with single strategy are shown in Fig. 4(CL_2_5_1), Fig. 5(CL_5), and Fig. 6 (dolphins). The left-hand side in these figures corresponds to B (betweenness centrality), and the right-hand side to IVD (the result of CCIVD was similar to IVD and thus not shown here).

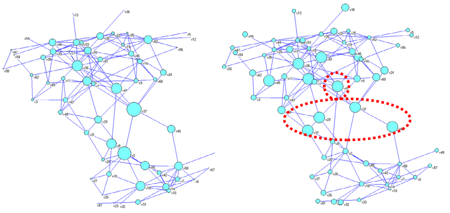

Fig. 6. Visualization result (dolphins)

By comparing the visualized networks, we can see that IVD could identify nodes which connect communities (encompassed with dotted red circles the figures). Since the visualized networks with IVD is similar to those with B, which is known to be effective for immunization despite its rather large time complexity, the node score can be said as effective for identifying intermediating nodes among communities.

4.3 Results of Synthetic Networks

Results of synthetic networks (average of 10 runs) are shown in Fig. 7 and Fig. 8. The horizontal axis is the node occupation probability p, and the vertical one is the relative size S of LCC in eq.(10). In the legend, gray lines with "x" are

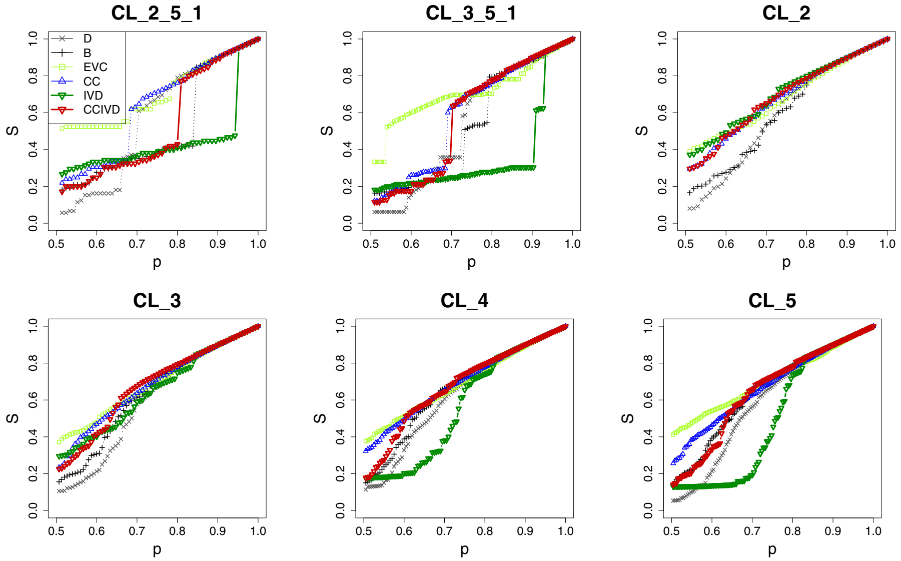

Fig. 7. Results on synthetic networks (single)

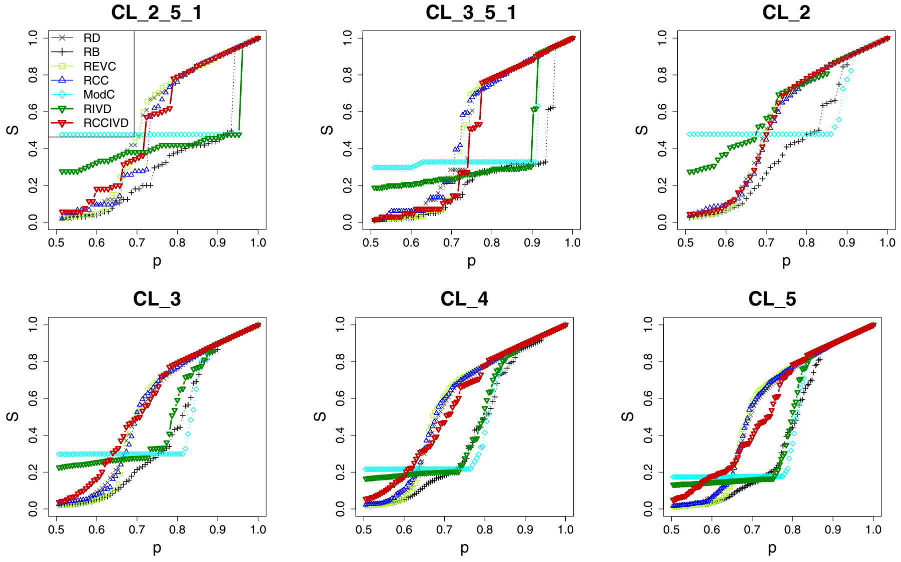

Fig. 8. Results of synthetic networks (recalc)

for D, black lines with "+" for B, yellow lines with squares for EVC, blue lines with upper triangles for CC, and water blue lines with diamonds for ModC. The proposed node scores are shown with lower triangle (IVD (green lines) and CCIVD (red lines)) in Fig. 7. For recalc strategy (Fig. 8), prefix "R" is put on the method name (except for ModC).

As shown in Fig. 7 and Fig. 8, IVD effectively immunized the networks in both single and recalc strategies. Especially, it showed the best performance for single strategy (rapid decrease of S with respect to p), and showed similar result with RB for recalc strategy around $p \geq 0.8$. On the other hand, unfortunately, CCIVD did not outperform B for single strategy, but it showed good performance for recalc strategy when p gets small (i.e., after large number of nodes are removed from a network). Compared with ModC, which also utilizes the community structure of a network (but requires community labels), RIVD outperformed ModC for all p, and RCCIVD showed better performance when p gets small.

4.4 Results of Real-World Networks

Results of real world networks in Table 2 are shown in Fig. 9 and Fig. 10. Both IVD and CCIVD showed almost equivalent performance with B for single strategy. For recalc strategy, the performance of RCCIVD was similar to RB (especially for karate and lesmis), and CCIVD outperformed IVD for real-world networks. This would be because the removal of hub nodes (with large community centrality) is effective for immunization of real-world networks. As in synthetic networks, the performance of ModC saturated after a network was divided into disconnected components. On the other hand, with the proposed methods, the value of S continued to fall even after a network was divided into disconnected components.

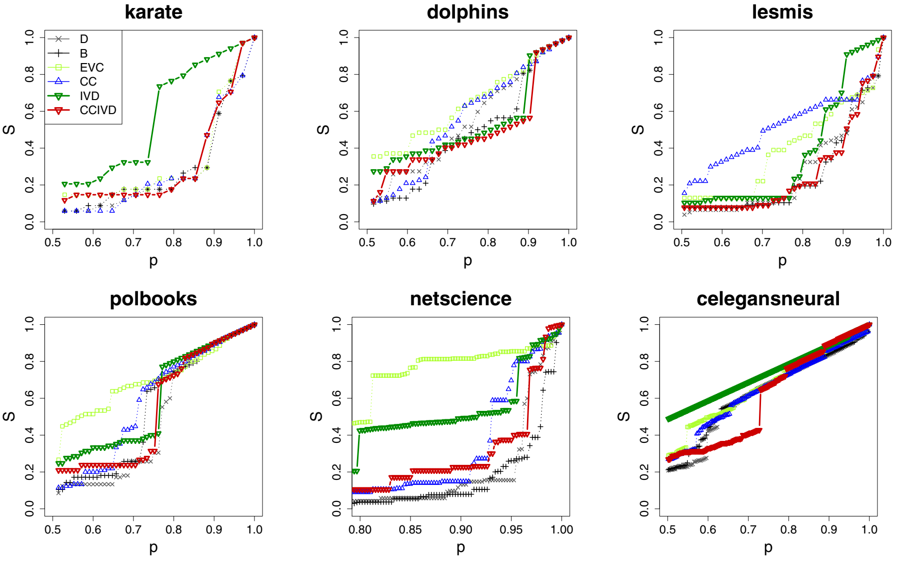

Fig. 9. Results of real-world networks (single)

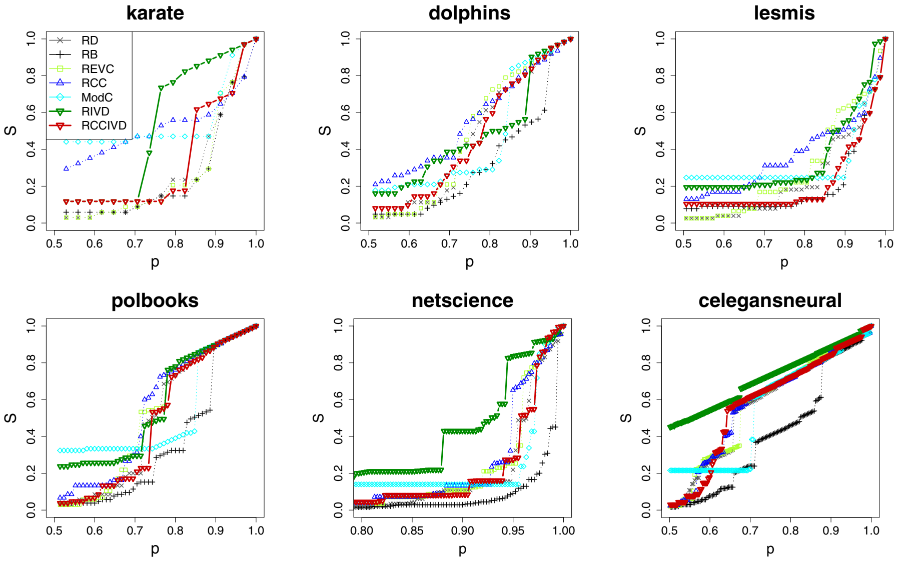

Fig. 10. Results of real-world networks (recalc)

4.5 Discussions

Our node scores (IVD and CCIVD) showed comparable performance with B (betweenness centrality), which is known to be effective for network immunization, in most networks. As in RB, the performance of these methods improved with recalc strategy, albeit this strategy requires much more computational effort. In addition, CCIVD showed better performance than CC, which is solely based on the norm of node vectors. This indicates the effectiveness of IVD for reflecting the community structure of a network in terms of the distribution of node vectors.

As shown in [5], utilization of the community structure of a network is effective for network immunization. However, finding community labels of nodes with maximum modularity is NP-complete [3]. The proposed approach can exploit the community structure of a network in terms of the distribution of node vectors, without relying on the externally supplied community labels of nodes.

5 Concluding Remarks

This paper reported a comparative study of community structure based node scores over both synthetic and real-world networks. Since contamination is propagated among groups of nodes (communities) through intermediating nodes in a networks, such nodes are identified based on the community structure of a network *without* requiring community labels of nodes. The characteristics of the proposed node vectors was analyzed. Extensive experiments were conducted to compare the node scores with other centrality based immunization strategies. The results are encouraging, and indicate that the node scores are promising.

Immediate future work includes more in-depth analysis of the node vectors and their relations. Especially, we plan to conduct the analysis of kernel density in terms of the histogram of node vectors for determining the appropriate parameter (e.g., θ) in our approach.

Acknowledgements. This work is partially supported by the grant-in-aid for scientific research (No. 24300049) funded by MEXT, Japan, Casio Science Promotion Foundation, and Toyota Physical & Chemical Research Institute.

References

1. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
2. Brandes, U.: A faster algorithm for betweenness centrality. Journal of Mathematical Sociology, 163–177 (2001)
3. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On modularity clustering. IEEE Transactions on Knowledge and Data Engineering 20(2), 172–188 (2008)
4. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70(6), 066111 (2004)
5. Masuda, N.: Immunization of networks with community structure. New Journal of Physics 11, 123018 (2011), doi:10.1088/1367-2630/11/12/123018
6. Mika, P.: Social Networks and the Semantic Web. Springer (2007)
7. Newman, M.: Finding community structure using the eigenvectors of matrices. Physical Review E 76(3), 036104(2006)
8. Newman, M.: Networks: An Introduction. Oxford University Press (2010)
9. Restrepo, J.G., Ott, E., Hunt, B.R.: Characterizing the dynamical importance of network nodes and links. Physical Review Letters 97, 094102 (2006)
10. Yoshida, T.: Toward finding hidden communities based on user profile. Journal of Intelligent Information Systems (2011) (accepted)
11. Yoshida, T., Yamada, Y.: Community Structure Based Node Scores for Network Immunization. In: Anthony, P., Ishizuka, M., Lukose, D. (eds.) PRICAI 2012. LNCS, vol. 7458, pp. 899–902. Springer, Heidelberg (2012)