Behaviormetrika
Vol.40, No.1, 2013, 1–15

# COMMUNITY STRUCTURE DISCOVERY IN DIRECTED GRAPHS BY ASYMMETRIC CLUSTERING

Wolfgang Gaul*, Rebecca Klages*, and Akinori Okada**

A familiar problem with respect to the analysis of network data (in which relations between objects can be described by links between the vertices of a graph) is the discovery of so-called community structures, i.e., the detection of subgraphs of closely connected vertices with comparatively few links joining vertices of different subgraphs. For this task modularity is a popular goodness-of-fit-index. While undirected graphs restrict considerations to basically symmetric relations, more realistic situations can be described by directed graphs. In this paper we consider shortest walk lengths between all pairs of vertices as dissimilarities instead of just using the adjacency information given by the directed edges of the graphs, which enables us to suggest a new approach in which the application of asymmetric clustering is a main step. This enrichment of the underlying adjacency matrix to a walk-length based dissimilarity matrix together with asymmetric hierarchical clustering are the keys of our proposed approach to community structure discovery in directed graphs. We use example graphs from the literature with known modularity values and apply computer-generated directed benchmark graphs for the evaluations. The findings show that our approach compares favourably with results available from the literature.

## 1. Introduction

Based on information by which objects are described the task of classifying a set of objects into subsets or classes such that similar objects belong to the same class is known as a standard problem of cluster analysis. How similarity between objects can be measured depends on the information provided. When objects can be interpreted as vertices of a graph and relations between objects (as links between the vertices of the graph) describe the only information available, the discovery of so-called community structures, i.e., the detection of subgraphs of closely connected vertices with comparatively few links joining vertices of different subgraphs, is a challenge for which quite a number of solution approaches have already been proposed in the literature (see, e.g., the overviews of Schaeffer (2007)[35], Fortunato (2010)[11]). Most contributions use undirected graphs in order to describe the underlying situation and modularity as popular goodness-of-fit-index for finding best solutions. Against this background basic notations with respect to community structure discovery, a summary of related research concerning modularity optimization, and the motivation for this paper are presented in section 2, which allows us to argue why traditional hierarchical clustering applied to shortest walk length dissimilarities as enrichment of the underlying adjacency information is an option for the determination of optimal community structures. In section 3 our new approach for directed graphs is described

*Key Words and Phrases:* Asymmetric Clustering, Community Structure Discovery, Modularity
* KIT, Karlsruhe, Germany
** Tama University, Tokyo, Japan

within a three-step framework: 1. Shortest walk computations are performed which enrich the given adjacency information. 2. Asymmetric clustering is applied to detect suitable subgraphs connected via shortest walks. 3. Vertex mover techniques are used to check whether the exchange of vertices between subgraphs can lead to better modularity values. Results based on examples from the literature with known modularity values and directed benchmark graphs are provided in section 4 to show that our approach compares favourably with earlier attempts to discover community structures in directed graphs. Finally, a concluding discussion and a brief outlook to further activities are provided in section 5.

## 2. Basic Notations, Summary of Related Work, and Motivation for this Paper

Let $G = (V, E)$ denote a graph with $V = V(G)$ as the set of vertices and $E = E(G)$ as the set of edges where each edge $e \in E$ links a pair of vertices $i, j, \in V$. As it depends on the application situation whether directed graphs with directed edges (arcs) or undirected graphs with undirected edges are used, the basic notations are given for both types of graphs. In all cases we restrict ourselves to graphs without parallel edges (the number of edges between a pair of vertices is always less than or equal to one) and without loops (for a pair of vertices an edge can only exist if the vertices are different). With $|M|$ as notation for the cardinality of a set $M$ we assume $|V| = n$ and $|E| = m$. In the undirected case the degree $d(i) = |N(i)|$ for vertex $i \in V$ is needed where $N(i)$ is the set of neighbours of $i$ and contains all vertices that are linked to $i$. In the directed case the set of vertices from which directed edges point to $i \in V$ is denoted by $N^-(i)$ and called the set of predecessors of $i$ while the set of vertices that are pointed to by directed edges with starting vertex $i \in V$ is denoted by $N^+(i)$ and called the successor set of $i$. $d^+(i) = |N^+(i)|$ gives the out-degree and $d^-(i) = |N^-(i)|$ the in-degree of $i \in V$.

The matrix $A = (A_{ij})$ with $A_{ij} = 1$, if an edge links $i, j, \in V$, $A_{ij} = 0$, otherwise, is called adjacency matrix of graph $G$ and describes the only information known in the underlying situation.

Given this adjacency information, the community structure discovery problem aims at determining a partition $\mathcal{C} = (C_1, C_2, \ldots, C_{|\mathcal{C}|})$ of $V$ into subsets $C_k \subset V, k = 1, \ldots, |\mathcal{C}|$, or – equivalently – finding subgraphs $G_k$ of $G$ with $V(G_k) = C_k$, that are closely connected while the number of edges between different subgraphs is small. As goodness-of-fit index the modularity measure

$$Q_{\mathcal{C}} = \frac{1}{2m} \sum_{i,j} \left( A_{ij} - \frac{d(i) \cdot d(j)}{2m} \right) \cdot \delta(C_i, C_j) \tag{1}$$

respectively

$$Q_{\mathcal{C}} = \frac{1}{m} \sum_{i,j} \left( A_{ij} - \frac{d^+(i) \cdot d^-(j)}{m} \right) \cdot \delta(C_i, C_j) \tag{2}$$

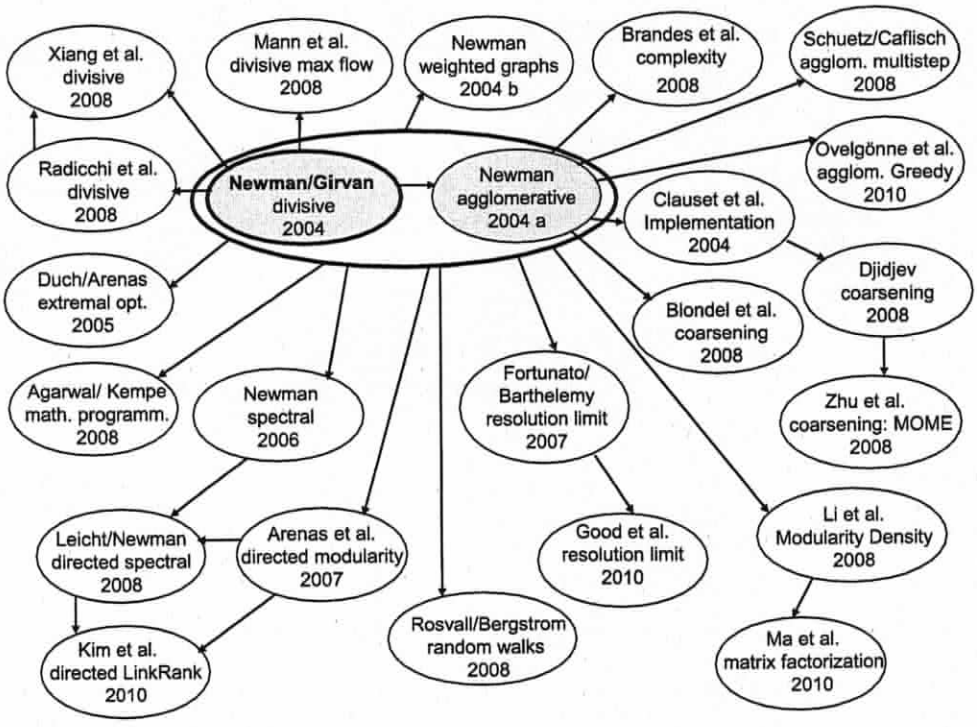


Figure 1: Activities in the Frame of Modularity Optimization.

is used, where in this notation $C_i$ describes the subset of partition $\mathcal{C}$ that contains vertex $i$ and $\delta(C_i, C_j)$ is the Kronecker-Delta which is equal to 1, if $C_i = C_j$, and equal to 0, otherwise.

Formula (1) describes the modularity measure for the undirected case, formula (2) adapts modularity to the directed case. No discussion of other suggestions to measure community structures will be provided in this paper (see, e.g., Brandes/Erlebach (Eds.) (2005)[4] for an introduction to graph clustering and a description of corresponding measures), instead figure 1 attempts to depict different aspects that have been tackled in the frame of modularity optimization.

Within a summary of related research two papers describe the starting point with respect to modularity optimization activities. In Newman/Girvan (2004)[29] modularity was introduced as a new goodness-of-fit measure together with a divisive hierarchical graph clustering procedure while an agglomerative version was suggested in Newman (2004a)[26].

There are various divisive (e.g., Radicchi et al. (2004)[32], Mann et al. (2008)[22], Xiang et al. (2008)[39]) and agglomerative (e.g., Schuetz/Caflisch (2008)[36], Ovelgönne et al. (2010)[31]) techniques. Furthermore, several methods iteratively coarsen the graphs to find communities within the coarsened copies of the original graph and refine these structures afterwards using the original graph (e.g., Arenas

et al. (2007)[2], Blondel et al. (2008)[3], Djidjev (2008)[8], Zhu et al. (2008)[40]). Other approaches apply well-known heuristics or reduce the problem to questions that have been solved in other contexts: Duch/Arenas (2005)[9] use extremal optimization, while Agarwal/Kempe (2008)[1] employ linear and vector programming. More recent approaches include the application of probabilistic flows on random walks in graphs (Rosvall/Bergstrom (2008)[34]) and matrix factorization (Ma et al. (2010)[21]). The application of spectral approaches was also discussed by several authors including Newman (2006)[28] (see Nascimento/de Carvalho (2010)[25] for a detailled survey concerning spectral approaches). Moreover, aspects such as implementation (e.g., Clauset et al. (2004)[6]), complexity (e.g., Brandes et al. (2008)[5]), alternate definitions (e.g., Li et al. (2008)[20]) and the extension of modularity to weighted graphs (Newman (2004b)[27]) have been considered. The concept has also been reviewed critically (e.g., Fortunato/Barthélemy (2007)[12], Good et al. (2010)[13]). While in earlier papers graph clustering based on the modularity measure was performed for undirected graphs (see formula (1)), considerations how modularity could be adapted to community structures in directed graphs have already been presented. The first approach to directly derive a goodness-of-fit measure for directed graphs (see formula (2)) from the modularity index for undirected graphs was given by Arenas et al. (2007)[2]. Leicht/Newman (2008)[19] encourage the used modification of modularity to directed graphs and show that the spectral algorithm proposed by Newman (2006)[28] to detect communities in undirected graphs can be altered in order to be applied to directed graphs using the directed version of modularity. Kim et al. (2010)[17] criticize this directed modularity formula (2) showing that the directions of the edges may not be adequately considered when this measure is used, and propose a different goodness-of-fit index to evaluate the quality of clusterings in directed graphs based on random walks, called LinkRank, which can be used in any algorithm that aims to find community structures.

Given the above mentioned activities in the frame of modularity optimization, the motivation for this paper is straightforward: The summary of related research based on modularity optimization has shown that – up to now – standard cluster analysis methodology (e.g., hierarchical clustering applied to (dis)similarities between objects) has not been used for solving the problem of detecting community structures. A reason may be that the adjacency information mentioned before is not of the form of a (dis)similarity matrix. However, if in a first step the adjacency information is enriched via shortest walk calculations, the derived shortest walk dissimilarities can be used as starting point for the application of standard cluster analysis methodology. In the case of undirected graphs the shortest walk dissimilarity matrix is symmetric and traditional clustering techniques can be applied. For directed graphs the shortest walk lengths can show asymmetric values, however, a recently published approach concerning asymmetric clustering (Takeuchi et al. (2007)[37]) can be used to find – as we will see in the next sections – promising solutions for the community structure detection problem. Of course, solutions obtained via standard cluster analysis methodology depend on the techniques chosen and the goodness-of-fit measures con-

sidered for their application. But if we calculate the corresponding modularity values for our "best" standard clustering solutions as comparative values (remember that all our examples mentioned in the summary of related research used modularity as measure for comparisons) we can bridge between modularity optimization and standard cluster analysis.

Thus, the idea to enrich the adjacency information between vertex pairs via shortest walk length values and apply hierarchical clustering enables us to suggest an additional alternative concerning community structure discovery which can be seen as a new approach put together with the help of techniques well-known in graph theory and cluster analysis.

In the following, we concentrate on directed graphs.

## 3. New Approach to Community Structure Discovery in Directed Graphs

As the discussion of the foregoing section and the lower left part of figure 1 have shown, first papers that deal with community structure discovery in directed graphs within the frame of modularity optimization have already appeared (Arenas et al. (2007)[2], Leicht/Newman (2008)[19], Kim et al. (2010)[17]). Our new approach adds at least two aspects not covered in earlier contributions: (I) The enrichment of the given adjacency information by shortest walk dissimilarities (Consider two non-adjacent vertices which could – nevertheless – be "similar" because they could have joint neighbours or the successor of the first vertex is the predecessor of the second vertex.) and (II) the application of asymmetric clustering (In directed graphs walk lengths from vertices $i$ to $j$ and from $j$ to $i$ are not necessarily equal, thus, the dissimilarity matrix is not symmetric (which would be needed for "traditional" clustering techniques) and, hence, asymmetric variants of cluster analysis techniques have to be considered.).

### *3.1 Shortest Walk Dissimilarities*

In a directed graph $G = (V, E)$ with adjacency matrix $A = (A_{ij})$, where a directed edge that links a starting vertex $i \in V$ with an end vertex $j \in V$ is denoted by $e = (i, j)$, a walk from $v_1 \in V$ to $v_2 \in V$ is a subgraph of $G$, described by $W_{v_1 v_2}$, with $E(W_{v_1 v_2}) = \{e_1, \ldots, e_s, \ldots, e_l\}$ as sequence of directed edges, $V(W_{v_1 v_2}) = \{v_1, j_1, j_2, \ldots, j_{l-1}, v_2\}$ as the corresponding set of vertices, and $e_1 = (v_1, j_1)$, $e_s = (j_{s-1}, j_s)$, $s \in \{2, \ldots, l-1\}$, $e_l = (j_{l-1}, v_2)$. $l$ is the number of directed edges of $W_{v_1 v_2}$ and called length of $W_{v_1 v_2}$. As there might be several walks between a pair of vertices, we denote by $l(W_{ij})$ the length of $W_{ij}$ and by

$$L_{ij} = \min_{W_{ij}} \; l(W_{ij})$$

the shortest walk length from $i$ to $j$. The determination of shortest walks has been

known for many years (e.g., Dijkstra (1959)[7], Floyd (1962)[10], Warshall (1962)[38]). In our situation

$$L_{ij} = \begin{cases} 1 & A_{ij} = 1, \\ \min\limits_{W_{ij}} \; l(W_{ij}) > 1 \quad \text{if} & A_{ij} = 0 \text{ but a walk } W_{ij} \text{ exists}, \\ L^* & A_{ij} = 0 \text{ and no walk } W_{ij} \text{ exists}, \end{cases} \tag{3}$$

with $L^*$ being a sufficiently large constant, enriches the adjacency information and allows (dis)similarity values to be assigned in cases in which a pair of vertices $i, j, \in V$ is not directly linked, i.e., $A_{ij} = 0$.

Thus, instead of the adjacency matrix $A = (A_{ij})$, the shortest walk length matrix $L = (L_{ij})$ will be used in the following.

### *3.2 Asymmetric Clustering*

Traditional clustering is based on symmetric (dis)similarity matrices that describe information to which extent objects belong together. But – of course – in some very common applications the most natural measure of similarity is asymmetric like the shortest walk lengths in directed graphs (see Hubert (1973)[14] who introduced asymmetric clustering). As the analysis of asymmetric data has extensively been studied by Japanese researchers (see, e.g., Okada/Iwamoto (1996)[30] for an early contribution and Takeuchi et al. (2007)[37] for a more recent paper on asymmetric clustering), we apply asymmetric agglomerative hierarchical clustering as described in Takeuchi et al. (2007)[37] to the shortest walk length matrix $L = (L_{ij})$ for the solution of the community structure discovery problem in directed graphs. Although we checked various variants, all results reported in the underlying paper are based on the choice of $W = \max$ (step 1 of definition 2 in the paper by Takeuchi et al. (2007)[37] (which supports the tendency that clusters with fairly short distances in both directions are joint first)) and the asymmetric updating formula for the asymmetric weighted average algorithm (step 2 of definition 2 in the paper by Takeuchi et al. (2007)[37] (which yielded the most convincing results)) (see Takeuchi et al. (2007)[37] for a detailed description).

### *3.3 Vertex Mover Refinement*

A common practice in traditional cluster analysis is to start with a hierarchical technique and apply a so-called exchange algorithm to the hierarchical solution that was considered most promising, i.e., the selected goodness-of-fit measure is checked for improvements when objects from a cluster of the chosen solution are removed and added to one of the remaining clusters. In the frame of modularity optimization this corresponds to a neighborhood search in the clusters of the currently best community structure discovery solution which Schuetz/Caflisch (2008)[36] referred to as vertex mover refinement. While these authors have described their refinement technique for undirected graphs, the idea can easily be adapted to directed graphs.

*3.4 Three steps approach*

Given the explanations presented above, our approach can be described within three steps:

Step 1: Shortest walk computations are performed to enrich the given adjacency information: $A = (A_{ij}) \rightarrow L = (L_{ij})$.

Step 2: Asymmetric agglomerative hierarchical clustering is applied to $L = (L_{ij})$: The partition $\mathcal{C} = (C_1, C_2, \ldots, C_{|\mathcal{C}|})$ with highest modularity value is selected.

Step 3: The solution from step 2 is checked for vertex mover refinements. The final solution is denoted as $\mathcal{C}_{SWAC}$ where SWAC is the acronym for Shortest Walks Asymmetric Clustering.

## 4. Examples and Results

Three types of examples are used to demonstrate that the SWAC approach compares favorably with results known from the literature. In example 1 (one directed ringgraph) and example 2 (30 computer-generated directed benchmark graphs with 500 and 1000 vertices) the spectral approach by Leicht/Newman (2008)[19], which uses the modularity measure for directed graphs by Arenas et al. (2007)[2] in combination with the spectral cluster procedure by Newman (2006)[28], has been incorporated into our examinations to give a feeling how the results obtained by the SWAC approach can be assessed. In these two examples, the Rand Index (RI) has been calculated (Rand (1971)[33]) in order to compare different community structure solutions.

In example 3 – due to suggestions from a referee to provide more comprehensive results – 10 repetitions of computer generations for prespecified parameters of directed benchmark graphs with 1000 and 2000 vertices (in total 360 graphs) have been checked to further help to assess the new approach. Again, for the comparison of the community structure solutions provided by the benchmark graphs and by our new approach we used the Rand Index (RI) although other indices related to RI exist (e.g., the earlier published Mirkin Index (Mirkin/Chernyi (1970)[24]) or the Adjusted Rand Index (Hubert/Arabie (1985)[15])). Additionally, the Variation of Information (VI) has been calculated (see, e.g., Meilă (2007)[23]) as this measure has also been applied in the literature concerning community structure discovery (see, e.g., Karrer et al. (2008)[16]).

Note (as a referee asked us to mention this) that RI is a similarity index with RI$\in [0, 1]$ and 1 as best value and VI is a metric on the space of partitions describing the dissimilarity between two partitions with 0 as best value.

Example 1:

As examples to demonstrate the advantages of their LinkRank measure Kim et al. (2010)[17] used directed graphs consisting of rings of $k$ subgraphs in which subgraphs are connected by a single directed edge of weight $w$ and where each subgraph itself is a ring with $n_k$ vertices and $m_k = n_k$ directed edges, that are either all clockwise or

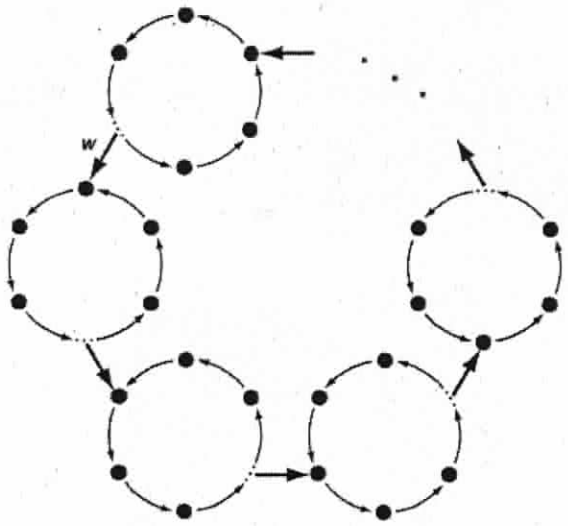


Figure 2: Example for a Directed Ringgraph (see Kim et al. (2010)).

Table 1: Results on the Ringgraph Discussed by Kim et al. (2010).

| Benchmark Graph | | Spectral Approach | | | Shortest Walk Approach | | |
|---|---|---|---|---|---|---|---|
| $Q_{Ring}$ | $\lvert\mathcal{C}_{Ring}\rvert$ | $Q_{Spec}$ | $\lvert\mathcal{C}_{Spec}\rvert$ | RI | $Q_{SWAC}$ | $\lvert\mathcal{C}_{SWAC}\rvert$ | RI |
| 0.7639 | 8 | 0.7058 | 6 | 0.9122 | 0.7639 | 8 | 1 |

anti-clockwise oriented. This kind of graph construction as depicted in figure 2 is very specific, thus, we only report results for the directed ringgraph with $w = 1$, $k = 8$ and $m_k = n_k = 8$ as also discussed in Kim et al. (2010)[17].

The optimal partition of the benchmark graph has $|\mathcal{C}_{Ring}| = 8$ subgraphs and a modularity value $Q_{Ring} = 0.7639$ which was also found by the SWAC approach (identical outcomes for $Q_{SWAC}$ and $|\mathcal{C}_{SWAC}|$ with a Rand index value 1), while the spectral approach finds a solution with only $|\mathcal{C}_{Spec}| = 6$ clusters and $Q_{Spec} = 0.7058$, see table 1.

Example 2:

Some of the directed graphs known from the literature that could be used for comparisons of community structure discovery algorithms are very specific as the ringgraph by Kim et al. (2010)[17] in example 1 has shown. Fortunately, Lancichinetti/Fortunato (2009)[18] have introduced a possibility to construct directed benchmark graphs with built-in community structures that account for the heterogeneity in the distributions of vertex degrees and community sizes, the values of which are computed by drawing random numbers from power laws with exponents $\tau_1$ and $\tau_2$. Weights for the directed edges and overlappings of communities can also be considered, but are not needed for the comparisons in our case.

With $n = |V|$ as number of vertices, $d_{av}$ as average vertex degree, $\mu$ as mixing parameter that describes the percentage of arcs incident to a vertex and linking this vertex to vertices in other clusters as the community to which the underlying vertex

Table 2: Results on Directed Benchmark Graphs with 500 Vertices.

| Benchmark Graphs | | | | | | Spectral Approach | | | Shortest Walk Approach | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d_{av}$ | $\mu$ | $m$ | $Q_{Bench}$ | $\lvert\mathcal{C}_{Bench}\rvert$ | $Q_{Spec}$ | $\lvert\mathcal{C}_{Spec}\rvert$ | RI | $Q_{SWAC}$ | $\lvert\mathcal{C}_{SWAC}\rvert$ | RI |
| 500 | 15 | 0.1 | 7896 | 0.8262 | 17 | 0.8262 | 17 | 1 | 0.8262 | 17 | 1 |
| 500 | 15 | 0.2 | 7789 | 0.7273 | 16 | 0.7016 | 13 | 0.9629 | 0.7273 | 16 | 1 |
| 500 | 15 | 0.3 | 7718 | 0.6190 | 14 | 0.6075 | 12 | 0.9813 | 0.6190 | 14 | 1 |
| 500 | 15 | 0.4 | 7837 | 0.5341 | 17 | 0.5248 | 13 | 0.9752 | 0.5341 | 17 | 1 |
| 500 | 15 | 0.5 | 7707 | 0.4292 | 15 | 0.4088 | 11 | 0.9499 | 0.4292 | 15 | 1 |
| 500 | 20 | 0.1 | 9332 | 0.8263 | 16 | 0.8263 | 16 | 1 | 0.8263 | 16 | 1 |
| 500 | 20 | 0.2 | 10081 | 0.7362 | 18 | 0.7362 | 18 | 1 | 0.7362 | 18 | 1 |
| 500 | 20 | 0.3 | 9600 | 0.6331 | 17 | 0.6193 | 13 | 0.9737 | 0.6331 | 17 | 1 |
| 500 | 20 | 0.4 | 10130 | 0.5358 | 17 | 0.5226 | 14 | 0.9752 | 0.5358 | 17 | 1 |
| 500 | 20 | 0.5 | 9813 | 0.4323 | 16 | 0.3992 | 11 | 0.9189 | 0.4323 | 16 | 1 |
| 500 | 25 | 0.1 | 12554 | 0.8235 | 16 | 0.8235 | 16 | 1 | 0.8235 | 16 | 1 |
| 500 | 25 | 0.2 | 12474 | 0.7317 | 17 | 0.7258 | 16 | 0.9924 | 0.7317 | 17 | 1 |
| 500 | 25 | 0.3 | 12633 | 0.6408 | 18 | 0.6356 | 17 | 0.9933 | 0.6408 | 18 | 1 |
| 500 | 25 | 0.4 | 12850 | 0.5301 | 16 | 0.5301 | 16 | 1 | 0.5301 | 16 | 1 |
| 500 | 25 | 0.5 | 12865 | 0.4362 | 17 | 0.4185 | 12 | 0.9554 | 0.4362 | 17 | 1 |

Table 3: Results on Directed Benchmark Graphs with 1000 Vertices.

| Benchmark Graphs | | | | | | Spectral Approach | | | Shortest Walk Approach | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | $d_{av}$ | $\mu$ | $m$ | $Q_{Bench}$ | $\lvert\mathcal{C}_{Bench}\rvert$ | $Q_{Spec}$ | $\lvert\mathcal{C}_{Spec}\rvert$ | RI | $Q_{SWAC}$ | $\lvert\mathcal{C}_{SWAC}\rvert$ | RI |
| 1000 | 15 | 0.1 | 15427 | 0.8633 | 33 | 0.8540 | 27 | 0.986 | 0.8633 | 33 | 1 |
| 1000 | 15 | 0.2 | 15395 | 0.7648 | 33 | 0.7436 | 21 | 0.9685 | 0.7648 | 33 | 1 |
| 1000 | 15 | 0.3 | 15520 | 0.6660 | 32 | 0.6538 | 22 | 0.9764 | 0.6660 | 32 | 1 |
| 1000 | 15 | 0.4 | 15831 | 0.5660 | 32 | 0.5474 | 18 | 0.9608 | 0.5660 | 31 | 0.9988 |
| 1000 | 15 | 0.5 | 14963 | 0.4641 | 29 | 0.4499 | 17 | 0.8845 | 0.4641 | 29 | 1 |
| 1000 | 20 | 0.1 | 19665 | 0.8630 | 33 | 0.8502 | 27 | 0.9833 | 0.8630 | 33 | 1 |
| 1000 | 20 | 0.2 | 20055 | 0.7645 | 35 | 0.7422 | 24 | 0.9636 | 0.7645 | 35 | 1 |
| 1000 | 20 | 0.3 | 19530 | 0.6635 | 31 | 0.6295 | 19 | 0.9424 | 0.6628 | 30 | 0.9985 |
| 1000 | 20 | 0.4 | 19463 | 0.5650 | 31 | 0.5430 | 19 | 0.9542 | 0.5650 | 31 | 1 |
| 1000 | 20 | 0.5 | 19160 | 0.4677 | 35 | 0.4340 | 12 | 0.9203 | 0.4666 | 33 | 0.9963 |
| 1000 | 25 | 0.1 | 25921 | 0.8625 | 30 | 0.8526 | 26 | 0.9874 | 0.8625 | 30 | 1 |
| 1000 | 25 | 0.2 | 25828 | 0.7630 | 31 | 0.7541 | 27 | 0.9856 | 0.7630 | 31 | 1 |
| 1000 | 25 | 0.3 | 25321 | 0.6651 | 32 | 0.6612 | 28 | 0.9926 | 0.6651 | 32 | 1 |
| 1000 | 25 | 0.4 | 25404 | 0.5660 | 32 | 0.56 | 26 | 0.9873 | 0.5635 | 29 | 0.9951 |
| 1000 | 25 | 0.5 | 25283 | 0.4647 | 31 | 0.4394 | 18 | 0.9392 | 0.4647 | 31 | 1 |

belongs, and $m = |E|$, we used graphs with 500 vertices (see table 2) and 1000 vertices (see table 3), average vertex degrees $d_{av}$ of 15, 20, and 25 and mixing parameters $\mu \in \{0.1, 0.2, \ldots, 0.5\}$. More precisely, in all cases the in-degree $d^-(i)$ was sampled from a power law with parameter $\tau_1 = 2$, the mixing parameters for in ($\mu^-$) - and out ($\mu^+$) - links were set equal to $\mu = \mu^+ = \mu^-$, for the community sizes a power law parameter $\tau_2 = 1$ was chosen and an interval $[20, 50]$ for the minimum respectively maximum number of vertices in a community was selected as either suggested in the description of the software package for the generation of benchmark graphs or applied by the authors Lancichinetti/Fortunato (2009)[18] in their demonstrations of how the

benchmark graph construction works.

In the following tables 2 and 3 the outcomes of the built-in structures of the directed benchmark graphs $Q_{Bench}$ and $|\mathcal{C}_{Bench}|$, the spectral approach $Q_{Spec}$ and $|\mathcal{C}_{Spec}|$ adapted to these directed graphs, and the corresponding SWAC approach values are depicted. For the spectral approach, which was recommended by Leicht/Newman (2008)[19] for directed graphs, we also applied the possibility of vertex mover refinements to establish equal opportunities for the comparisons of the results.

Table 2 shows that for all directed benchmark graphs with 500 vertices the SWAC approach succeeds in finding the optimal built-in community structure (the modularity values and the cardinalities of the partitions coincide, the Rand Index has value 1). For table 3 one recognizes that for three benchmark graphs with 1000 vertices (with the parameter constellations 1000, 20, 0.3, 19530 and 1000, 20, 0.5, 19160 as well as 1000, 25, 0.4, 25404) the SWAC approach does not compute the community structure with the highest modularity value although (in terms of modularity value and cardinality of the computed partition) the differences are small and always better than the outcomes of the spectral approach. For the benchmark graph with parameter constellation 1000, 15, 0.4, 15831 the SWAC approach finds a solution with the same modularity value $Q_{SWAC} = 0.566$ as the built-in community structure ($Q_{Bench} = 0.566$) but only $|\mathcal{C}_{SWAC}| = 31$ communities (instead of $|\mathcal{C}_{Bench}| = 32$), which indicates that optimal community structures don't need to be unique.

Example 3:

Again, computer-generated benchmark graphs with built-in community structures were constructed according to the Lancichinetti/Fortunato (2009)[18] instructions. This time, with the same notations and explanations as already described in example 2, $n \in \{1000, 2000\}$, $d_{av} \in \{15, 20, 25\}$, and $\mu \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$ were selected as parameters. For each of these 36 parameter constellations 10 directed benchmark graphs were sampled using the power laws with, again, $\tau_1 = 2$ and $\tau_2 = 1$. This time, table 4 and table 5 show the averaged results of these repetitions. Here, it does not make sense to report the numbers $m$ of edges for the 10 repetitions of the benchmark graph constructions (as these numbers differ). Here, $Q_{Bench}$, $Q_{SWAC}$, and $|\mathcal{C}_{Bench}|$, $|\mathcal{C}_{SWAC}|$ describe the averages of the corresponding values of the 10 repetitions based on the built-in structures of the computer-generated benchmark graphs and the solutions of the new approach. This time, the average values of the Rand Index RI and the Variation of Information VI as another measure for comparing clusterings also used in the area of community structure detection (see, e.g., Meilă (2007)[23]) are reported and $\mu = 0.6$ (60% of the arcs linked to a vertex in a cluster are incident to vertices in other clusters) was added to the range of mixing parameter values used in example 2 to show the effect that the comparison measures start to deteriorate to a larger extent if the mixing parameter indicates that the cluster structure in the underlying graph worsens too much. For different average vertex degrees $d_{av}$ the average RI values (figure 2 and figure 4) and the average VI values (figure 3 and figure 5) are depicted dependent on the mixing parameter $\mu$. For $\mu > 0.4$ both comparison

Table 4: Results on Directed Benchmark Graphs with 10 Repetitions per Parameter Constellation $(1000, d_{av}, \mu)$.

| Benchmark Graphs | | | | | Shortest Walk Approach | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $d_{av}$ | $\mu$ | $Q_{Bench}$ | $\lvert\mathcal{C}_{Bench}\rvert$ | $Q_{SWAC}$ | $\lvert\mathcal{C}_{SWAC}\rvert$ | RI | VI |
| 1000 | 15 | 0.1 | 0.8612 | 29.3 | 0.8612 | 29.3 | 1 | 0 |
| 1000 | 15 | 0.2 | 0.7635 | 30.1 | 0.7635 | 30.1 | 1 | 0 |
| 1000 | 15 | 0.3 | 0.6642 | 30.8 | 0.6642 | 30.8 | 1 | 0 |
| 1000 | 15 | 0.4 | 0.5643 | 29.9 | 0.5643 | 29.7 | 0.9995 | 0.0066 |
| 1000 | 15 | 0.5 | 0.4826 | 30.1 | 0.4628 | 28.6 | 0.9965 | 0.0333 |
| 1000 | 15 | 0.6 | 0.3647 | 30.4 | 0.3629 | 27.4 | 0.9946 | 0.1260 |
| 1000 | 20 | 0.1 | 0.8617 | 29.8 | 0.8617 | 29.8 | 1 | 0 |
| 1000 | 20 | 0.2 | 0.7627 | 29.5 | 0.7627 | 29.5 | 1 | 0 |
| 1000 | 20 | 0.3 | 0.6627 | 29.4 | 0.6627 | 29.4 | 1 | 0 |
| 1000 | 20 | 0.4 | 0.5638 | 29.9 | 0.5638 | 29.9 | 1 | 0 |
| 1000 | 20 | 0.5 | 0.4645 | 30.6 | 0.4638 | 29.7 | 0.9979 | 0.0416 |
| 1000 | 20 | 0.6 | 0.3649 | 30.9 | 0.3627 | 26.0 | 0.9908 | 0.2007 |
| 1000 | 25 | 0.1 | 0.8618 | 30.2 | 0.8618 | 30.2 | 1 | 0 |
| 1000 | 25 | 0.2 | 0.7633 | 29.9 | 0.7632 | 29.9 | 1 | 0 |
| 1000 | 25 | 0.3 | 0.6646 | 30.9 | 0.6635 | 30.0 | 0.9989 | 0.0246 |
| 1000 | 25 | 0.4 | 0.5649 | 30.9 | 0.5648 | 30.8 | 0.9999 | 0.0037 |
| 1000 | 25 | 0.5 | 0.4639 | 30.0 | 0.4637 | 29.8 | 0.9995 | 0.0095 |
| 1000 | 25 | 0.6 | 0.3643 | 30.2 | 0.3629 | 27.0 | 0.9940 | 0.0648 |

Table 5: Results on Directed Benchmark Graphs with 10 Repetitions per Parameter Constellation $(2000, d_{av}, \mu)$.

| Benchmark Graphs | | | | | Shortest Walk Approach | | | |
|---|---|---|---|---|---|---|---|---|
| $n$ | $d_{av}$ | $\mu$ | $Q_{Bench}$ | $\lvert\mathcal{C}_{Bench}\rvert$ | $Q_{SWAC}$ | $\lvert\mathcal{C}_{SWAC}\rvert$ | RI | VI |
| 2000 | 15 | 0.1 | 0.8807 | 59.8 | 0.8807 | 59.8 | 1 | 0 |
| 2000 | 15 | 0.2 | 0.7814 | 59.5 | 0.7813 | 58.9 | 0.9999 | 0.0024 |
| 2000 | 15 | 0.3 | 0.6817 | 61.8 | 0.6817 | 61.7 | 0.9999 | 0.0032 |
| 2000 | 15 | 0.4 | 0.5816 | 60.3 | 0.5815 | 59.6 | 0.9997 | 0.0172 |
| 2000 | 15 | 0.5 | 0.4826 | 60.6 | 0.4803 | 54.7 | 0.9961 | 0.0335 |
| 2000 | 15 | 0.6 | 0.3828 | 61.8 | 0.3813 | 53.0 | 0.9943 | 0.2482 |
| 2000 | 20 | 0.1 | 0.8808 | 61.0 | 0.8808 | 61.0 | 1 | 0 |
| 2000 | 20 | 0.2 | 0.7826 | 62.8 | 0.7826 | 62.8 | 1 | 0 |
| 2000 | 20 | 0.3 | 0.6828 | 63.1 | 0.6825 | 62.6 | 0.9999 | 0.0029 |
| 2000 | 20 | 0.4 | 0.5819 | 59.6 | 0.5814 | 59.1 | 0.9999 | 0.0037 |
| 2000 | 20 | 0.5 | 0.4828 | 62.0 | 0.4821 | 59.2 | 0.9977 | 0.0823 |
| 2000 | 20 | 0.6 | 0.3830 | 61.9 | 0.3822 | 60.7 | 0.9949 | 0.0942 |
| 2000 | 25 | 0.1 | 0.8811 | 62.2 | 0.8811 | 62.2 | 1 | 0 |
| 2000 | 25 | 0.2 | 0.7819 | 60.5 | 0.7819 | 60.5 | 1 | 0 |
| 2000 | 25 | 0.3 | 0.6819 | 59.9 | 0.6819 | 59.9 | 1 | 0 |
| 2000 | 25 | 0.4 | 0.5818 | 61.0 | 0.5812 | 57.8 | 0.9998 | 0.0084 |
| 2000 | 25 | 0.5 | 0.4827 | 61.8 | 0.4816 | 57.7 | 0.9994 | 0.0192 |
| 2000 | 25 | 0.6 | 0.3823 | 60.8 | 0.3811 | 55.1 | 0.9969 | 0.1297 |

measures indicate the above mentioned deterioration effect which is reasonable as it is questionable whether the detection of communities makes sense in graphs with only vague cluster structure. For $\mu \leq 0.4$, our new approach behaves remarkably well.

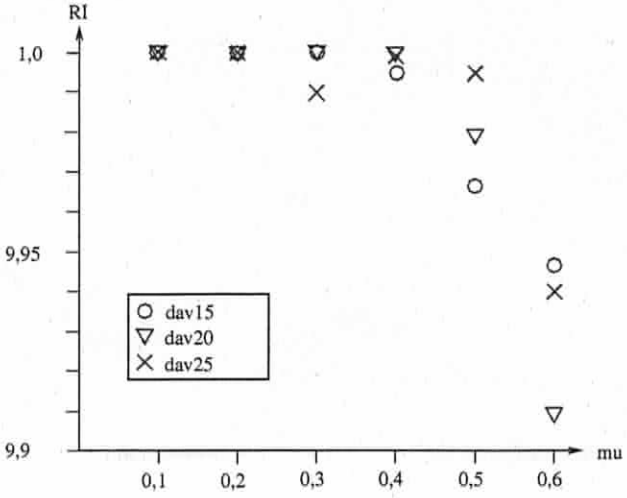


Figure 3: Average Rand Indices RI concerning $\mathcal{C}_{SWAC}$ compared to $\mathcal{C}_{Bench}$ for Directed Benchmark Graphs with 1000 Vertices (10 Repetitions per Parameter Constellation $(1000, d_{av}, \mu)$).

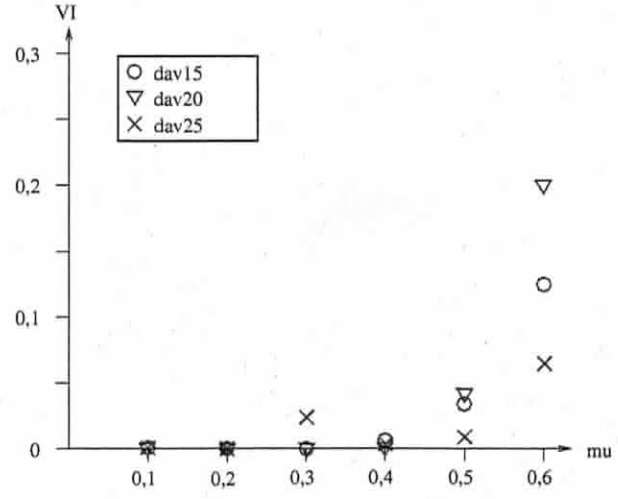


Figure 4: Average Variation of Information VI concerning $\mathcal{C}_{SWAC}$ compared to $\mathcal{C}_{Bench}$ for Directed Benchmark Graphs with 1000 Vertices (10 Repetitions per Parameter Constellation $(1000, d_{av}, \mu)$).

## 5. Concluding Remarks

For the community structure discovery problem in directed graphs a new solution technique, called SWAC (Shortest Walks Asymmetric Clustering) approach, has been proposed that combines graph theory with results known from asymmetric cluster analysis of (dis-)similarity data. In order to find a partition of the graph's vertex set that has an optimal value of the modularity measure in the directed case, our

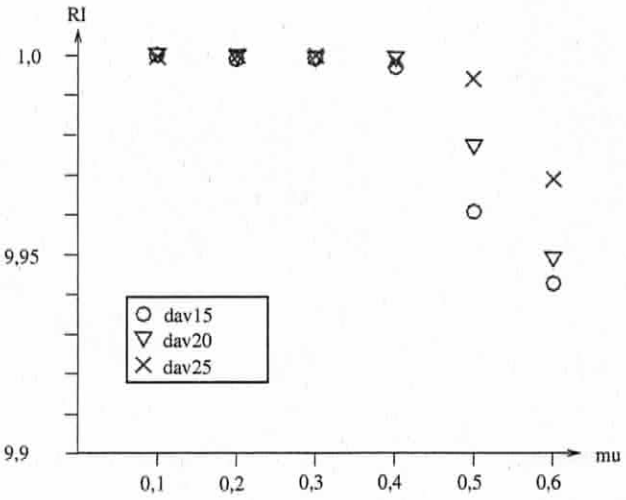


Figure 5: Average Rand Indices RI concerning $\mathcal{C}_{SWAC}$ compared to $\mathcal{C}_{Bench}$ for Directed Benchmark Graphs with 2000 Vertices (10 Repetitions per Parameter Constellation $(2000, d_{av}, \mu)$).

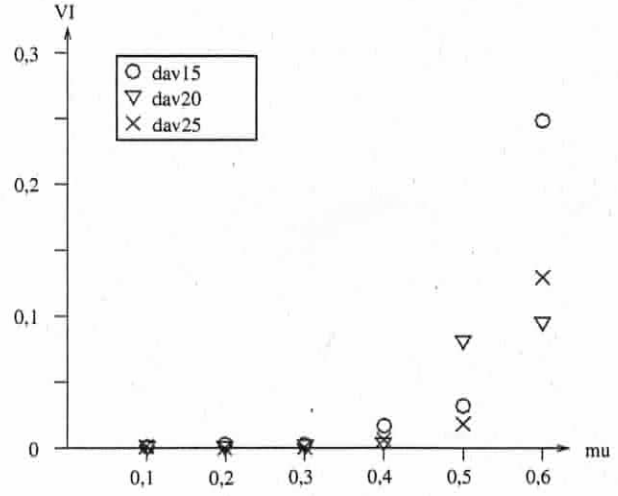


Figure 6: Average Variation of Information VI concerning $\mathcal{C}_{SWAC}$ compared to $\mathcal{C}_{Bench}$ for Directed Benchmark Graphs with 2000 Vertices (10 Repetitions per Parameter Constellation $(2000, d_{av}, \mu)$).

approach computes the lengths of shortest walks between the vertices in the directed graph as enrichment of the given adjacency information and applies hierarchical cluster analysis for asymmetric dissimilarity data to the walk lengths. From the obtained hierarchy the clustering with highest modularity is selected as starting solution for a neighborhood search within the graph clusters.

To test our proposal we mainly constructed directed benchmark graphs computer-generated according to the Lancichinetti/Fortunato (2009)[18] instructions and com-

pared the outcomes of the SWAC approach in two ways: We used a variant of a spectral clustering technique recommended for community structure discovery in directed graphs and applied two comparison measures, the Rand Index (RI) and the Variation of Information (VI), for assessing the community structure recovery properties of our approach with convincing results.

So far we have applied the SWAC approach to directed, unweighted graphs. While standard clustering can be applied for undirected unweighted graphs, the treatment of weighted graphs with possibly overlapping communities is a next challenge.

## REFERENCES

[1] Agarwal, G., Kempe, D. (2008). Modularity-Maximizing Graph Communities via Mathematical Programming. *European Physical Journal B 66*, 409–418.

[2] Arenas, A., Duch, J., Fernández, A., Gómez, S. (2007). Size Reduction of Complex Networks Preserving Modularity. *New Journal of Physics* 9, 176.

[3] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). Fast Unfolding of Community Hierarchies in Large Networks. *Journal of Statistical Mechanics*, P10008.

[4] Brandes, U., Erlebach, T. (Ed.) (2005). Network Analysis: Methodological Foundations. *Lecture Notes in Computer Science* 3418, Springer-Verlag, Berlin-Heidelberg.

[5] Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z., Wagner, D. (2008). On Modularity Clustering. *IEEE Transactions on Knowledge an Data Engineering* 20(2), 172–188.

[6] Clauset, A., Newman, M. E., Moore, C. (2004). Finding Community Structure in Very Large Networks. *Physical Review E* 70, 066111.

[7] Dijkstra, E. W. (1959). A Note on two Problems in Connexion With Graphs. *Numerische Mathematik* 1, 269–271.

[8] Djidjev, H. N. (2008). A Scalable Multilevel Algorithm for Graph Clustering and Community Structure Detection. *Lecture Notes in Computer Science* 4936, 117–128, Springer-Verlag, Berlin-Heidelberg.

[9] Duch, J., Arenas, A. (2005). Community Detection in Complex Networks Using Extremal Optimization. *Physical Review E* 72, 027104.

[10] Floyd, R. W. (1962). Algorithm 97: Shortest Path. *Communications of the ACM* 5(6), 345–345.

[11] Fortunato, S. (2010). Community Detection in Graphs. *Physics Reports* 486, 75–174.

[12] Fortunato, S., Barthelemy, M. (2007). Resolution Limit in Community Detection. *Proceedings of the National Academy of Sciences* 104(1), 36–41.

[13] Good, B. H., de Montjoye, Y.-A., Clauset, A. (2010). The Performance of Modularity Maximization in Practical Contexts. *Physical Review E* 81, 046106.

[14] Hubert, L. (1973). Min and Max Hierarchical Clustering Using Asymmetric Similarity Measures. *Psychometrika* 38(1), 63–72.

[15] Hubert, L., Arabie, P. (1985). Comparing Partitions. *Journal of Classification* 2, 193–218.

[16] Karrer, B., Levina, E., Newman, M. E. (2008). Robustness of Community Structure in Networks. *Physical Review E* 77(4), 046119.

[17] Kim, Y., Son, S.-W., Jeong, H. (2010). LinkRank: Finding Communities in Directed Networks. *Physical Review E* 81, 016103.

[18] Lancichinetti, A., Fortunato, S. (2009). Benchmarks for Testing Community Detection Algorithms on Directed and Weighted Graphs with Overlapping Communities. *Physical Review E* 80, 016118.

[19] Leicht, E. A., Newman, M. E. (2008). Community Structure in Directed Networks. *Physical*

*Review Letters* 100, 118703.

[20] Li, Z., Zhang, S., Wang, R.-S., Zhang, X.-S., Chen, L. (2008). Quantitative Function for Community Detection. *Physical Review E* 77, 036109.

[21] Ma, X., Gao, L., Yong, X., Fu, L. (2010). Semi-Supervised Clustering Algorithm for Community Structure Detection in Complex Networks. *Physica A* 389, 187–197.

[22] Mann, C. F., Matula, D. W., Olinick, E. V. (2008). The Use of Sparsest Cuts to Reveal the Hierarchical Community Structure of Social Networks. *Social Networks* 30, 223–234.

[23] Meilă, M. (2007). Comparing Clusterings – An Information Based Distance. *Journal of Multivariate Analysis* 98(5), 873–895.

[24] Mirkin, B. G., Chernyi, L. B. (1970). Measurement of the Distance Between Distinct Partitions of a Finite Set of Objects. *Automation and Remote Control* 31(5), 786–792.

[25] Nascimento, M. C., de Carvalho, A. C. (2010). Spectral Methods for Graph Clustering – A Survey. *European Journal of Operational Research* 211(2), 221–231.

[26] Newman, M. E. (2004a). Fast Algorithm for Detecting Community Structure in Networks. *Physical Review E* 69, 066133.

[27] Newman, M. E. (2004b). Analysis of Weighted Networks. *Physical Review E* 70, 056131.

[28] Newman, M. E. (2006). Finding Community Structure in Networks Using the Eigenvectors of Matrices. *Physical Review E* 74, 036104.

[29] Newman, M. E., Girvan, M. (2004). Finding and Evaluating Community Structure in Networks. *Physical Review E* 69, 026113.

[30] Okada, A., Iwamoto, T. (1996). University Enrollement Flow Among the Japanese Prefectures: A Comparison Before and After the Joint First Stage Achievement Test by Asymmetric Cluster Analysis. *Behaviometrika* 23(2), 169–185.

[31] Ovelgönne, M., Geyer-Schulz, A., Stein. M. (2010). Randomized Greedy Modularity Optimization for Group Detection in Huge Social Networks. In *SNA-KDD'10: Proceedings of the 4th Workshop on Social Network Mining and Analysis, ACM*, 1–9.

[32] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D. (2004). Defining and Identifying Communities in Networks. *Proceedings of the National Academy of Sciences* 101(9), 2658–2663.

[33] Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association* 66(336), 846–850.

[34] Rosvall, M., Bergstrom, C. T. (2008). Maps of Random Walks on Complex Networks Reveal Community Structure. *Proceedings of the National Academy of Sciences* 105, 1118–1123.

[35] Schaeffer, S. E. (2007). Graph Clustering. *Computer Science Review* 1, 27–64.

[36] Schuetz, P., Caflisch, A. (2008). Multistep Greedy Algorithm Identifies Community Structure in Real-World and Computer-Generated Networks. *Physical Review E* 78, 026112.

[37] Takeuchi, A., Saito, T., Yadohisa, H. (2007). Asymmetric Agglomerative Hierarchical Clustering Algorithms and Their Evaluations. *Journal of Classification* 24, 123–143.

[38] Warshall, S. (1962). A Theorem on Boolean Matrices, *Journal of the ACM* 9(1), 11–12.

[39] Xiang, J., Hu, K., Tang, Y. (2008). A Class of Improved Algorithms for Detecting Communities in Complex Networks. *Physica A* 387, 3327–3334.

[40] Zhu, Z., Wang, C., Ma, L., Pan, Y., Ding, Z. (2008). Scalable Community Discovery of Large Networks. *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management*, 381–388.

(Received October 29 2012, Revised February 1 2013)