Hierarchical clustering in minimum spanning trees

Meichen Yu,[1,a)] Arjan Hillebrand,[1] Prejaas Tewarie,[2] Jil Meier,[3] Bob van Dijk,[1] Piet Van Mieghem,[3] and Cornelis Jan Stam[1]
[1]*Department of Clinical Neurophysiology and MEG Center, VU University Medical Center, PO Box 1081 HV, Amsterdam, The Netherlands*
[2]*Department of Neurology, VU University Medical Center, Amsterdam, The Netherlands*
[3]*Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, PO Box 5031, 2600 GA, Delft, The Netherlands*

(Received 7 October 2014; accepted 2 February 2015; published online 11 February 2015)

The identification of clusters or communities in complex networks is a reappearing problem. The minimum spanning tree (MST), the tree connecting all nodes with minimum total weight, is regarded as an important transport backbone of the original weighted graph. We hypothesize that the clustering of the MST reveals insight in the hierarchical structure of weighted graphs. However, existing theories and algorithms have difficulties to define and identify clusters in trees. Here, we first define clustering in trees and then propose a tree agglomerative hierarchical clustering (TAHC) method for the detection of clusters in MSTs. We then demonstrate that the TAHC method can detect clusters in artificial trees, and also in MSTs of weighted social networks, for which the clusters are in agreement with the previously reported clusters of the original weighted networks. Our results therefore not only indicate that clusters can be found in MSTs, but also that the MSTs contain information about the underlying clusters of the original weighted network. © *2015 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4908014]

Clustering or community structure has been regarded as one of the most significant features of complex networks. The minimum spanning tree (MST) is the spanning tree in a weighted graph for which the sum of the weights of its constituting links is minimal. There are many different kinds of algorithms for identifying clusters in general networks, but a general definition and algorithm for the detection of clusters in trees has not been established. In our current study, we first define clusters in trees and propose a tree agglomerative hierarchical clustering (TAHC) method for the detection of clusters in MSTs. By detecting clusters in both artificial trees and the MSTs of two weighted social networks, we demonstrate that the clustering of the MSTs reveals the underlying clusters of the original weighted graphs.

I. INTRODUCTION

In the past decade, complex network theory has been widely used in different disciplines, such as social, technological, and biological systems.[1–3] Clusters, also called the community structure or modules, are important features of complex networks.[4–6] Qualitatively, clusters can be defined as groups of nodes within a graph such that there is a higher density of links within clusters than between them.

In contrast to general graphs, trees are maximally sparse connected graphs.[7] To date, there are few studies that present methods for the detection of clusters or community structure in trees.[4] However, trees, especially MSTs play an important role in the investigation of the dynamical and topological properties of complex networks.[8] The MST has been identified as an important transport backbone of the original weighted graph.[9] Several studies have demonstrated that under certain conditions the transportation in weighted graphs is dominated by their MSTs.[10–12] In addition, for the comparison of empirical brain networks, the MST has been shown to be a sensitive and practical tool to overcome problems caused by differences in network density (the number of links) or average link weight.[13,14] The importance of both the MST and community structure in weighted graphs has motivated us to investigate whether the MST contains information about clusters in the underlying weighted network and whether these clusters can be detected in minimum spanning trees.

Many different kinds of clustering algorithms have been developed.[4,5] Among them, hierarchical clustering methods play an important role in linking the well-known scale-free and small-world network models, and also in predicting missing links.[15–19] There are two types of hierarchical clustering methods: divisive and agglomerative. Divisive algorithms start with all connections in the network and iteratively remove links, which divides the graph progressively into smaller and smaller disconnected sub-graphs identified as the clusters. The divisive approaches differ in how links are removed. The best-known divisive method, the Girvan-Newman (GN) algorithm, is based on the removal of links with high betweenness (In an unweighted graph, the shortest path length (hopcount) is the minimum number of links that must be traversed to move from one node to another.[20] The betweenness of a link is the number of shortest paths between all possible pairs of nodes in the graph that traverse that link). The GN algorithm is able to detect known clusters in both computer-generated and real-world graphs.[21]

[a)]Author to whom correspondence should be addressed. Electronic mail: m.yu@vumc.nl

However, in trees with many low degree nodes connected by links with high betweenness, the GN algorithm will produce too many isolated nodes.

In contrast, agglomerative algorithms start from an empty graph with the same number of nodes as the original network, but without links. Links are then iteratively added in order of decreasing similarity[22] until all the nodes are merged into one cluster. The so-called Louvain method is a popular agglomerative method.[23] The Louvain method is a greedy optimization algorithm, based on optimizing the modularity of a partition (the modularity of a partition measures the density of links inside clusters compared to links between clusters)[6,20,24] of a graph. In the Louvain method, the similarity between two nodes is quantitatively measured by the gain of modularity that would take place by merging the two nodes. This method will iteratively merge pairs of nodes until a maximum of modularity is attained. This greedy algorithm can deal with very large networks (networks with millions of nodes and billion of links) with high computational efficiency. However, several studies have shown that the optimization of modularity has a resolution limit: relatively small, but very dense clusters may be undetectable in the presence of relatively large clusters.[25,26] These studies imply that the resolution limit of modularity depends on the comparison between the total number L of links in the entire graph and the number l_s of links inside clusters. Usually, the resolution problem[25] will occur in clusters with a about $\sqrt{2L}$ number of internal links or smaller ($l_s \leq \sqrt{2L}$). Trees, being maximally sparse connected graphs, will consist mainly of sparse clusters. Hence, applying the Louvain algorithm or other optimization algorithms to trees will suffer from the resolution limit.[27] Moreover, trees and tree-like graphs can possess unexpectedly high modularity, so the Louvain method, as one of the modularity maximization algorithms, might behave unexpectedly in trees.[27] Recently, another agglomerative algorithm, called spanning tree separation (STS) clustering was introduced to identify clusters in general graphs.[28] The STS is based on calculating the number of spanning trees running through all the links in general graphs. However, the STS algorithm is not applicable to trees, since the number of spanning trees running through all the links would always equal to 1. In order to deal with the limitations of existing algorithms, we introduce here a new hierarchical clustering method based on an agglomerative algorithm that is able to detect clusters in trees. Moreover, rather than using modularity or any other criterion to stop the agglomerative process, we present the entire hierarchical clustering in the form of a dendrogram.

The paper is organized as follows. In Sec. II, we first describe how clusters can be defined for trees and hypothesize that trees consist of two fundamental motifs with characteristic clustering structures. We then detail the shortcomings of the existing divisive and agglomerative methods for tree clustering, with the GN algorithm and Louvain method as examples, respectively. To avoid these shortcomings, we propose a new agglomerative method and describe its implementation. In Sec. III, we demonstrate the performance of the method using two artificial trees with known clustering structures, as well as for the MSTs of two well-known weighted social networks containing clustering structure. We finish with a discussion and conclusions in Sec. IV.

II. FINDING CLUSTERS IN TREES

A. The definition of clusters in trees

In this paper, we present a hierarchical clustering method to discover clusters in MSTs. Our study concentrates on undirected (non-rooted) MSTs with unweighted links. Since MSTs possess all the properties of trees, we will start by discussing how to define clusters in trees.

To date, there is no standard definition of for clusters in a tree. Trees do not accord with the conventional definition of clusters, which are groups of nodes within a graph such that there is a higher density of links within clusters than between them. There are two extreme configurations for trees: a star and a path (Figs. 1(a) and 1(b)). A star consists of one central node (hub: nodes with higher degree than others) and several peripheral nodes (leaves, degree 1 nodes) connected to the central one with only one link. Trees can be constructed by connecting several stars of possibly different size. According to the definition of clusters as internally dense and externally sparse sub-graphs, each star within one general tree naturally forms one cluster, because in each star most leaves only connect to their own hub, except for a small number of leaves connecting the rest of the graph (Fig. 1(c)). In a path, however, except for the leaves at either end, all nodes are connected to their two neighbors (Fig. 1(b)). A path has no density fluctuations (except at the extremes), so none of the sub-graphs in a path satisfies the cluster definition. Therefore, clustering detection in trees consisting of paths will have to rely on their topological structure. For instance, in the Cayley tree (a Cayley tree is a regular tree in which every node i is linked to k neighbours ($k = 3$ in Fig. 1(d)), except for leaf nodes on the boundary[29]), every

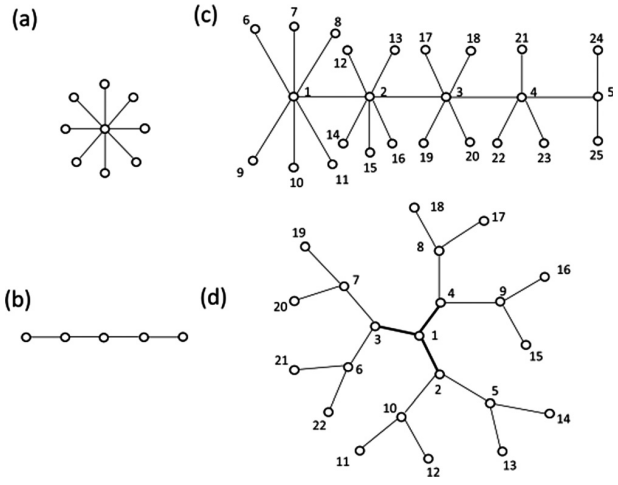

FIG. 1. Examples of two fundamental motifs in general trees. (a) star-motif with one central node and (in this case) eight leaves; (b) line-motif with (in this case) five nodes; (c) five connected star-motifs; (d) three connected line-motifs: Cayley tree. Note that in (c), nodes 1, 2, 3, 4, 5 are five hubs; in (d), the three line-motifs connecting to the central node 1 by three bold links correspond to three clusters, respectively.

node has the same degree except for the leaves on the boundary (Fig. 1(d)). In this case, we consider that the Cayley tree consists of one central node (node 1) and three "paths" (In the Cayley tree (Fig. 1(d)), the nodes in the three "paths," except for the leaves at either end, are connected to three neighbors. Here, we consider that the nodes belonging to the same line-motif in the Cayley tree form a "path," because the "path" also have no density fluctuations (except at the extremes). In the Cayley tree, node 1 plays an important role as the concentrator to integrate and separate the information from and to the three line-motifs. Given the symmetrical structure of the three "paths," we regard the three "paths" in Fig. 1(d) as three clusters and node 1 as one single cluster in the Cayley tree.

In our study, we define stars and paths as star-motifs and line-motifs in general trees, respectively. We hypothesize that any general tree that consists of some combination of the two motifs will possess corresponding clustering structures.

B. Shortcomings of divisive methods in the case of trees

Our proposed approach is based on agglomerative hierarchical clustering, since divisive methods have shortcomings for tree clustering, as will be illustrated here for the GN algorithm.

If two approximately equally large clusters in a graph are loosely interconnected by a few links, these intercluster links will have higher link betweenness than links within each cluster. The GN algorithm aims at finding and removing links with high betweenness. The link betweenness must be recalculated following the removal of each link, because the link betweenness for remaining links will no longer be correct for the remaining graph[24] (note that this step is redundant for trees, since for these acyclic graphs there is exactly one path connecting two clusters). By removing these links consecutively, the GN algorithm can identify the underlying clustering structure. However, as mentioned in the Introduction section, the GN algorithm will produce too many isolated nodes in trees consisting of low degree nodes. Trees are maximally sparse graphs, which could contain many of the line-motifs mentioned above. Fig. 2(a) shows a tree that consists of two star-motifs connected by one line-motif between them. In Fig. 2(a), node 5 is a concentrator playing the same role as node 1 in the Cayley tree (Fig. 1(d)), which can be regarded as one single cluster. The links belonging to the line-motif have higher betweenness. According to the GN algorithm, these links with higher betweenness will be removed step by step. As a result, there will be two star-motifs with several isolated nodes left. In fact, Fig. 2(a) has symmetrical structure, but the GN algorithm will ignore it.

C. Shortcomings of existing agglomerative methods in the case of trees

Although the agglomerative methods can deal with the disadvantages of divisive methods, the existing agglomerative methods still have limitations for tree clustering. The Louvain method is a typical agglomerative method, which has been successfully used to identify hierarchical clustering

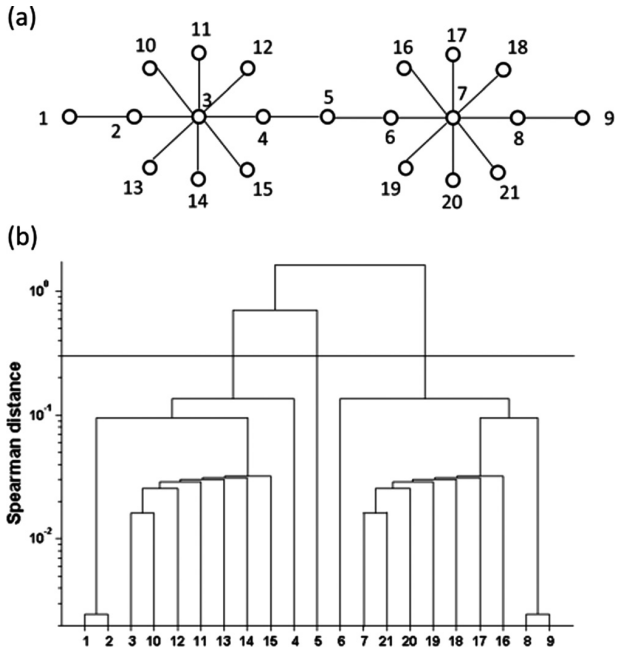

FIG. 2. A tree and its corresponding dendrogram of the hierarchical clustering structure. (a) A tree consisting of two star-motifs connected by one line-motif in the middle. (b) The dendrogram of the hierarchical clustering result for this tree. The numbers along the horizontal axis correspond to the 21 nodes in (a) and the upside-down U-shaped lines denote the links between the 21 nodes; the height of the U-shaped lines is the Spearman distance between two nodes, indicating that two nodes merged at that hierarchy have identical similarities; from bottom to top, nodes will be joined together by the U-shaped lines until all nodes are merged into a single cluster; the hierarchical clustering result can be obtained by the cross-section of the dendrogram at any hierarchy indicated by a solid line. From top to bottom, we define that the first hierarchy of the dendrogram starts with the highest U-shaped line, so the solid line in (b) is positioned between the second and the third hierarchy.

in huge real-world networks. In this section, we concentrate on presenting the limitations of the Louvain method.

Due to the resolution limit caused by modularity maximization, the Louvain method usually fails to detect small clusters when the following condition is met for the number of links within the cluster, $l_s \leq \sqrt{2L}$.[25] Fig. 1(c) shows a tree constructed by connecting five star-motifs. The five star-motifs correspond to five clusters. In this artificial tree, the total number of links is $L = 24$, and the number of links inside each cluster from left to right are $l_{s_i} = \{6, 5, 4, 3, 2\}$ ($i \in \{1, 2, 3, 4, 5\}$ corresponds to the five clusters from left to right, respectively). In this case $\sqrt{2L} \approx 6.93$, and therefore, $l_{s_i} < \sqrt{2L}$ ($i \in \{1, 2, 3, 4, 5\}$) for all the five clusters. Hence, the Louvain method applied to these kinds of trees will unavoidably be hindered by the resolution limit.

D. New agglomerative hierarchical clustering based on geodesic distance matrix

In this paper, we propose a new TAHC method to identify clusters in trees. In graph theory,[6] a graph can be expressed by its adjacency matrix $\mathbf{A}$, whose entries a_{ij} take the value 1 if there is a link between node i to node j in the tree and 0 otherwise. Based on the adjacency matrix $\mathbf{A}$, traditional agglomerative hierarchical clustering methods tend to

find only the core nodes of clusters, but not the peripheral nodes.[24] Therefore, for tree clustering, rather than directly analyzing A, we utilize the geodesic distance matrix C as input for the clustering algorithm.

The geodesic distance matrix C is a weighted matrix in which the entries are the geodesic distances between all possible pairs of nodes in a graph. Here we confine to a graph that is a tree. The geodesic distance between two nodes in a tree is equal to the number of links in the shortest path (there is only one path linking two nodes in a tree) between the two nodes (and therefore equal to 1 for directly linked nodes).

Agglomerative hierarchical clustering starts with an empty graph of N nodes with no links between them. In the agglomerative process, similarities between node pairs are calculated, hence a similarity measure is required. In the symmetric geodesic distance matrix C, each node corresponds to a row and a column vector. Based on **C**, we calculate vector similarities between all row pairs of the **C**. Here, we employ the commonly used Spearman's rank correlation $r_{Spearman}$.[30] The vector similarity is then simply the so-called Spearman distance, defined as $d_{Spearman} = 1 - r_{Spearman}$. This measure is computed for every node pair and used as the input matrix for hierarchical clustering. After obtaining the similarities between every node pair, links are added to the node pairs in order of decreasing similarity, starting with the nodes pairs with highest similarity. When two nodes are merged into one cluster, we use the average-linkage clustering (When there is more than one node in one cluster, the distance between this cluster and other clusters can be calculated in different ways. In the average-link clustering, the distance between two clusters is equal to the average distance from any nodes of one cluster to any node of the other cluster) to define the distance between the new cluster and other nodes. Based on average-link clustering, node pairs will be merged into corresponding clusters in order of decreasing similarity. In the final step, the agglomerative method will merge all nodes into one single cluster. The merging during the different stages of the algorithm can be represented in the form of a dendrogram (see Fig. 2(b)).

To summarize our TAHC algorithm:

(1) Calculate the geodesic distances between all possible pairs of nodes of the given graph (which is a tree) and use the geodesic distance matrix **C** as input to the agglomerative hierarchical clustering algorithm.
(2) Assign each node (the row vector of **C**) to one cluster.
(3) Define the $d_{Spearman}$ as vector similarity distance between all row pairs of **C**.
(4) Find the most similar pair of clusters and merge them into a single cluster.
(5) Calculate similarities between the new cluster and each of the old clusters based on average-linkage clustering.
(6) Repeating steps 4 and 5 until all nodes are merged into a single cluster.

E. Time complexity

For a MST with N nodes and $N-1$ links, the geodesic distance matrix C can be obtained by Dijkstra's algorithm in time $O((2N-1)log(N))$. The average-linkage clustering needs $O(N^2)$ time steps. If $N \to \infty$, $O((2N\text{-}1)log(N)) = O(Nlog(N))$, and $O(Nlog(N) + N^2) = O(N^2)$.[31] Thus the overall run time of the TAHC algorithm scales as $O(N^2)$ on a MST.

F. Comparing the clusters in the original weighted graph and the MST

To evaluate the performance of the TAHC method, we adopted normalized mutual information (NMI)[32] to quantify the similarity between the underlying "real" clusters and the clusters detected by the TAHC method. NMI is based on the confusion matrix **N**, in which the entries N_{ij} are the number of nodes in the "real" cluster i that appear in the "detected" cluster j. The measure of similarity between different clustering results is defined as follows:

$$NMI(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij}\log\left(\frac{N_{ij}N}{N_{i\cdot}N_{\cdot j}}\right)}{\sum_{i=1}^{C_A} N_{i\cdot}\log\left(\frac{N_{i\cdot}}{N}\right) + \sum_{j=1}^{C_B} N_{\cdot j}\log\left(\frac{N_{\cdot j}}{N}\right)},$$

where C_A and C_B are the number of "real" clusters and "detected" clusters, respectively. The sum over row i of matrix **N** is denoted $N_{i\cdot}$ and $N_{\cdot j}$ is the sum over column j of matrix **N**. NMI takes the maximum value of 1 when the "detected" clusters are equal to the "real" clusters. NMI equals zero when the "detected" clusters are completely independent of the "real" partitions.

III. APPLICATIONS

A. Artificial trees consisting of star-motifs and line-motifs

To test the performance of the proposed algorithm, we first applied it to two artificial trees.

Fig. 3 shows the hierarchical clustering result for the artificial tree shown in Fig. 1(c), which consists of five star-motifs, where five hubs connect to each other in order of decreasing number of leaves. Five clusters were obtained by the TAHC algorithm, which is in line with the definition of clustering for star-motifs in Sec. II. However, the Louvain method was able to detect only four clusters (see Fig. 3), as it incorrectly merged the two star-motifs that have few leaf nodes due to the limited resolution of this approach.

For the Cayley tree, each node has the same degree (all equal to 3), except for leaf nodes (see Fig. 1(d)). Using our definition for line-motifs in trees (Sec. II), this Cayley tree can be considered as containing one central node (node 1) connecting three line-motifs. Given that the three line-motifs have equivalent structures, which all link to node 1 by one node, there is no unequivocal way to decide to which cluster node 1 should be assigned, so node1 should be one single cluster. Nodes that are part of the line-motifs are symmetric in the lower hierarchy of the dendrogram (see Fig. 4) due to the symmetrical structures of the Cayley tree. The TAHC method can detect four expected clusters (Fig. 4), but the Louvain method fails (see Table I).

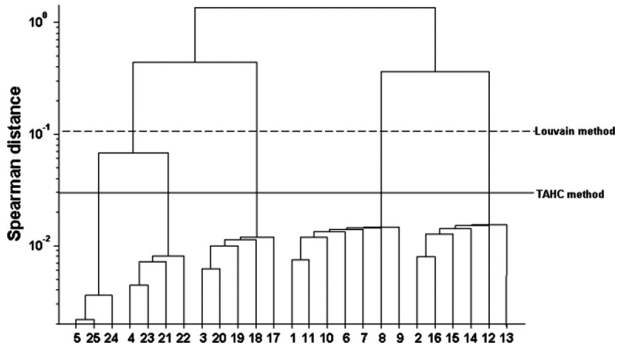

FIG. 3. Hierarchical clustering results for the artificial tree with five connected star-motifs in Fig. 1(b). The dotted line between the third and fourth hierarchy represents the results of Louvian method, and the solid line between the fourth and fifth hierarchy represents the results of the TAHC method. Note that the TAHC method finds the correct clusters, whereas the Louvain method erroneously merges two clusters.

Note that: L and l_s denote the total number of links in a graph and the number of links inside clusters, respectively. Usually, the resolution problem will occur when $l_s \leq \sqrt{2L}$.[25]

B. MSTs of the Zachary's karate club network and the *Les Misérables* network

Next, we tested the TAHC approach to the MSTs of two weighted social networks (Zachary's karate club network and *Les Misérables* network) that have been characterized previously in terms of clustering structure. Here, the MST was obtained by Kruskal's algorithm.[33] The two social networks belong to a widely used benchmark to test the performance of clustering detection algorithms, including the GN algorithm.[21,24] Although the original versions of the two social networks are weighted networks, most studies have only considered their unweighted version for simplicity, thereby missing the important underlying information reflected by the link weights. Here, we investigated the hierarchical clustering of the MSTs of the two *weighted* social networks. The key question here is whether the MSTs can reveal the clusters that are present in their original weighted networks.

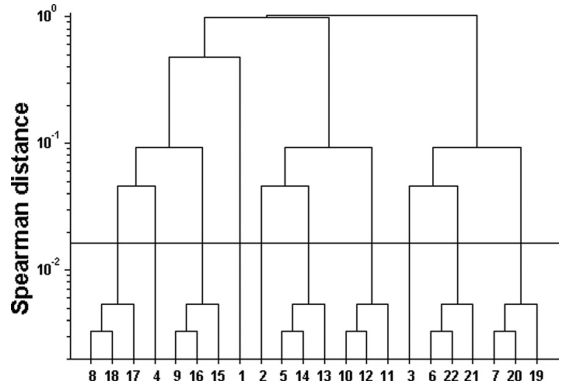

FIG. 4. Hierarchical clustering results of the Cayley tree using the TAHC method. The solid line between the second and third hierarchy provides the correct clustering for the Cayley tree, namely, the single central node (node 1), and three symmetric clusters.

TABLE I. Clustering results in Cayley tree of TAHC compared with Louvain Method.

Cluster	Louvain	TAHC
1	1,2,5,13,14	2,5,10,11,12,13,14
2	3,6,7,19,20,21,22	3,6,7,19,20,21,22
3	4,8,9,15,16,17,18	4,8,9,15,16,17,18
4	10,11,12	1
$l_s = 6 < \sqrt{2L} \approx 6.48$		

The Zachary's karate club network[34] consists of 34 members of the karate club at a U.S. university in 1970. Their mutual relationships were investigated over a period of 2 yr. The club split into two groups after a dispute between a teacher (node 1 in Fig. 5) and the administrator (node 33 in Fig. 5) of the club.

Fig. 5 illustrates that the weighted Zachary's karate club network consists of two groups Fig. 6(a) shows the MST of the weighted network. The two biggest star-motifs in the MST consist of the two largest hub nodes 1 and 34, respectively.

Fig. 6(b) shows the hierarchical clustering for the MST using the TAHC method, which showed that the MST can be divided into two clusters. This result successfully corresponds to the actual division in the original weighted network (Fig. 5), except for node 29 (NMI = 0.8372). This node is a leaf of node 3 in the MST, both of which will therefore always be assigned to the same cluster. Thus, for the weighted Zachary's karate club network, these results demonstrate that its MST can establish the known clustering with a high accuracy.

The co-appearance network of major characters in the novel *Les Misérables*, authored by Victor Hugo, as compiled by Knuth,[35] was also analyzed. In this network, the 77 interacting characters are represented by 77 nodes, and two nodes are linked if the two characters appear together in one or more chapters of the book. The weights of the links are the number of the co-appearances.

Fig. 7 shows the MST of the weighted *Les Misérables* network, which consists of six relatively larger star-motifs. These star-motifs contain six relatively higher degree nodes (six characters): the higher degree node Valjean (the leading

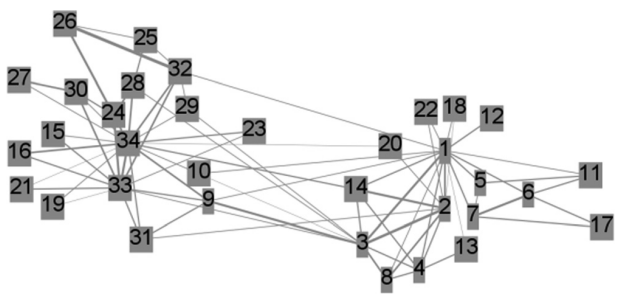

FIG. 5. Weighted Zachary's karate club network consists of the two known clusters. Link thickness represents the strength of the relationships between any two of the 34 members (i.e., the weights). The teacher and the administrator of the club are represented by nodes 1 and 34, respectively.

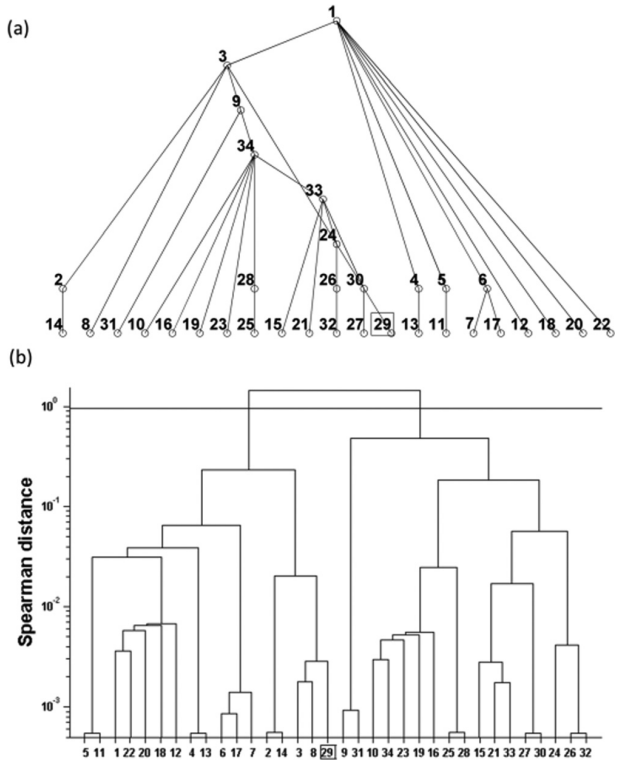

FIG. 6. The MST and corresponding hierarchical cluster of the Zachary's karate club network. (a) The MST of the weighted Zachary's karate club network. (b) The hierarchical clustering result of the MST. Note that the labels of the 34 nodes in the MST and the dendrogram correspond to the 34 labels in Fig. 5. The solid line between the first and the second hierarchy gives the clustering result obtained with the TAHC method, which correctly identified the two clusters in the network. Only the cluster assignment of Node 29 (squared) was inconsistent with that of the underlying weighted network (see Fig. 5).

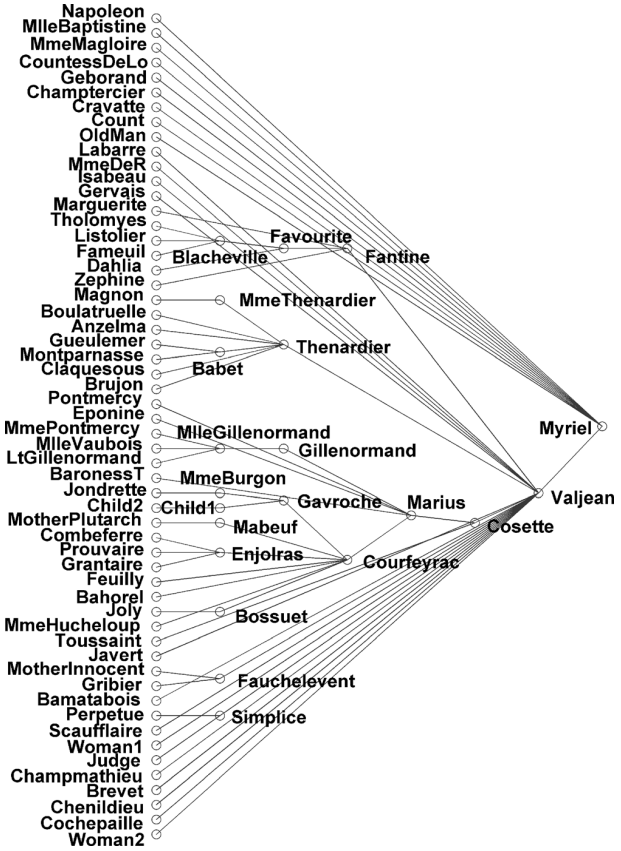

FIG. 7. The MST of the weighted *Les Misérables* network. Each node is labelled with the name of each character in the book.

hero: degree = 20); node Myriel (the Bishop, who saved Valjean: degree = 10); node Courfeyrac (one of members of the *friends of the ABC society*: degree = 8); node Thenardier (the leading villain: degree = 7); node Marius (the young hero: degree = 6); node Fantine (the leading heroine: degree = 4). In this MST, there is one node Cosette (the young heroine who is the daughter of Fantine and also the wife of Marius), linking Valjean and Marius (the husband of Cosette), playing an important role in the book.

The TAHC method successfully detects seven clusters consisting of the six large star-motifs and one important node Cosette with its leave Toussaint (who was appointed by Valjean to protect Cosette) (Fig. 8). The detailed description of the relationships between the characters in each cluster can be found in Appendix A.

For comparison, we also directly applied Louvain method to the weighted *Les Misérables* network (Fig. 9). The Louvain method detected six clusters in the original weighted *Les Misérables* network, which were similar to the seven clusters found in the MST by the TAHC method (NMI = 0.8237).

Hence, the TAHC algorithm can effectively detect the underlying hierarchical clustering embedded in the MST of the weighted *Les Misérables* network.

IV. CONCLUSION

In this article, we first defined the clusters in general trees in terms of two basic motifs (stars and paths) and then presented a novel TAHC method to investigate these types of clusters in MSTs. This is the first clustering method that can be successfully applied to trees, i.e., maximally sparse connected graphs. The TAHC algorithm used the geodesic distance matrix C to compute the similarity between two nodes in the tree. This similarity was defined as the Spearman distance between row pairs in the geodesic distance matrix C. We tested the effect of using a simpler metric for the hierarchical clustering, i.e., we replaced the Spearman distances between row pairs in the geodesic distance matrix C by the geodesic distances between pairs of nodes. This gave similar results, yet for the real social networks this simpler metric was less sensitive to the intricate details of the underlying clusters (see Appendix B).

We have shown that the TAHC method can detect the underlying known clusters for two artificial trees that consist of these two fundamental motifs. The TAHC method presents better results on the artificial trees compared to the Louvain method, which suffers from a resolution limit. We have also demonstrated the utility and reliability of the TAHC method by applying it to the MST of two weighted social networks. In 2002, Girvan and Newman first applied the GN method to the unweighted Zachary's karate club network and detected the two known clusters.[21] However, only

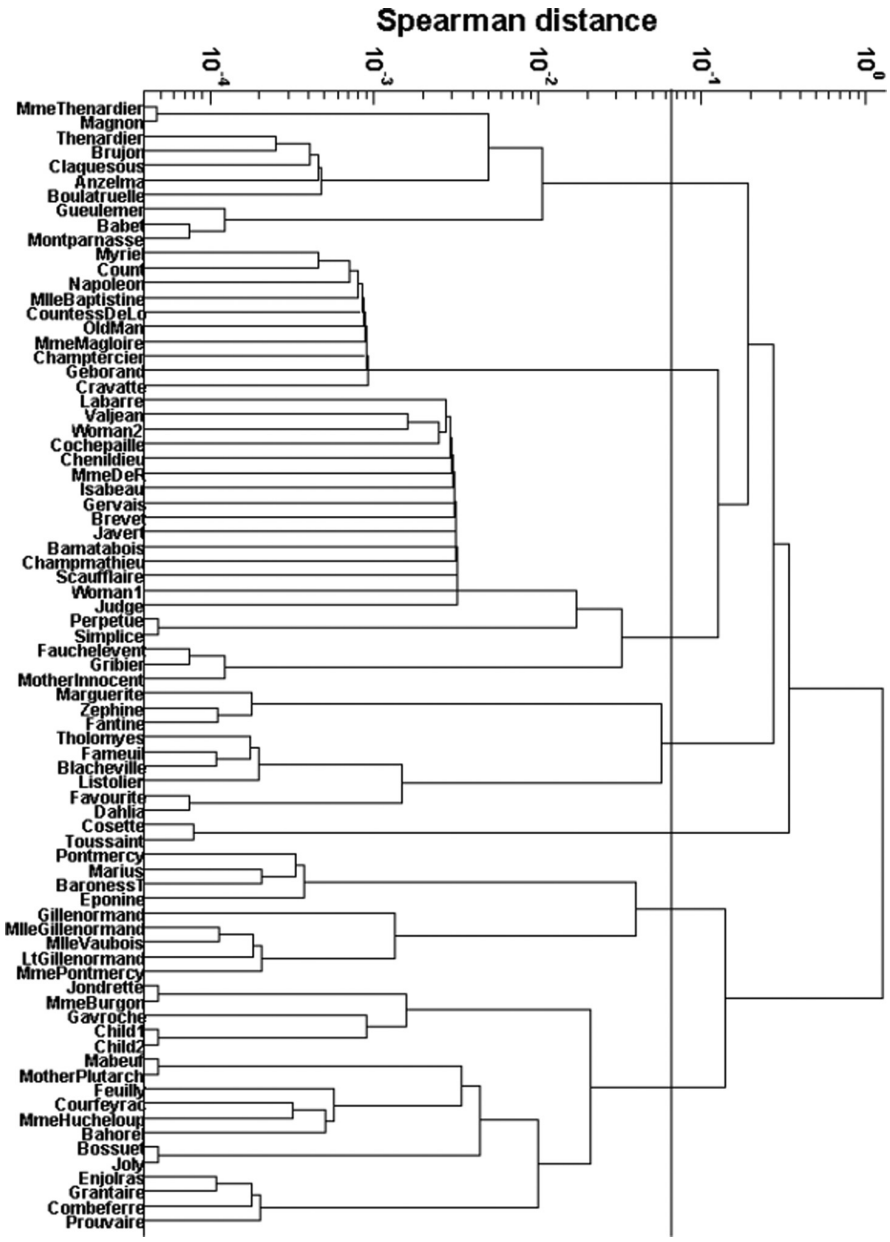

FIG. 8. The hierarchical cluster of the MST of the weighted *Les Misérables* network. Note that the label of each node corresponds to the label in Fig. 7. The solid line between the sixth and seventh hierarchy gives the clustering result obtained with the TAHC method, which identified the seven clusters in the network.

considering the unweighted network misses important information embedded in link weights. In 2008, Arenas and colleagues successfully detected the two clusters in the original weighted Zachary's network. In our study, the TAHC method extracts the known clusters in the MST of weighted Zachary's karate club network with a high degree of success (only one node was classified incorrectly). Similarly, many studies have analyzed the unweighted *Les Misérables* network, but the original weighted *Les Misérables* network has not yet been studied.[24,36,37] Application of the TAHC method to the MST of the weighted *Les Misérables* network revealed clusters that largely overlapped with the clusters we found for the original weighted network using the Louvain method. Our clustering results of the MST of two weighted social networks indicate that the MSTs can reveal most of the known clusters of their original weighted networks.

There were some small differences between the clusters of the original weighted networks and the clustering as obtained using the TAHC method applied to the MSTs. There could be several reasons for this: first, the TAHC method may not be sensitive enough to fully detect the underlying clustering structure in MSTs, especially for MSTs containing many line-motifs. Second, there may be overlapping clusters in the two employed social networks,[38] which have not been considered in this article. Third, by constructing the MST some information about the clustering of the underlying weighted network may have been removed. Finally, the previously reported clustering for the underlying weighted networks was used here as gold standard, yet this gold standard may not be the perfect.

Future studies will focus on finding more sensitive clustering method for MSTs and on the collection of weighted

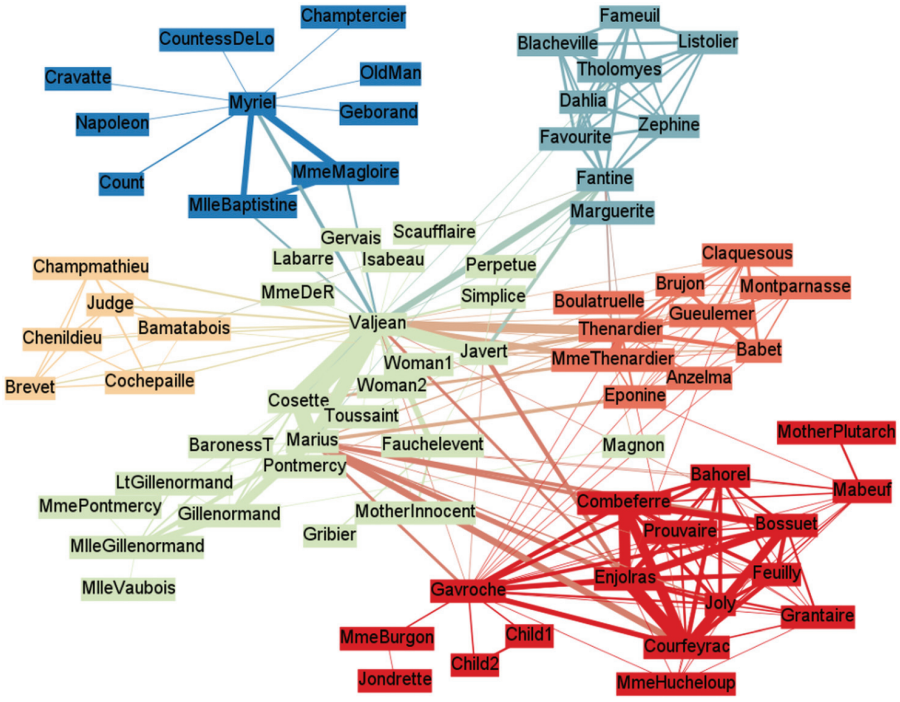

FIG. 9. The six clusters, represented in different colors, of the original weighted *Les Misérables* network identified by Louvain method. Link thickness indicates its weights, which represent the number of the co-appearances of corresponding characters in this book.

real-world networks with well-documented clusters for analysis. Moreover, we have not presented an objective metric (for example, the modularity[24]) to evaluate the hierarchical clustering results, because trees are likely to suffer from the resolution limit when using modularity.[25,27] In our study, we presented the entire hierarchical dendrograms and obtained clusters by comparing with previously reported clustering results of the original weighted networks. However, in practice, the clusters are not known in advance. Future studies should therefore develop an objective heuristic to set the threshold for the dendrograms. Some studies aimed to find relationships between MSTs and the single-linkage and average-linkage cluster analysis of the original graph.[39,40] In our study, we applied the average-linkage cluster analysis to the geodesic distance matrix of the MSTs rather than the original graph. We represented the results of the TAHC algorithm in the form of dendrograms, in which the nodes of the MSTs are depicted at the bottom. The roots of the dendrograms represent corresponding hierarchies, at which nodes are merged into a cluster. The hierarchical clustering results can be obtained by the cross-section of the dendrograms at any hierarchy or root. However, there may be also some relationships between finding clusters in the MSTs and finding an appropriate root in corresponding dendrograms of the MSTs. We believe that this is an interesting direction for future research.

We have demonstrated that MSTs contain information about the clustering in the underlying weighted networks, and that these clusters can be detected successfully in the MST using an agglomerative hierarchical clustering approach. We envisage that the TAHC method will be useful in the identification of clustering in various MSTs, as obtained from a range of complex networks, including social networks, genetic control networks, as well as functional and structural brain networks.

ACKNOWLEDGMENTS

This work has been supported by the China Scholarship Council (CSC).

APPENDIX A: THE RELATIONSHIPS BETWEEN THE CHARACTERS IN EACH CLUSTER OF THE MST OF THE WEIGHTED *LES MISÉRABLES* NETWORK

From top to bottom in Fig. 8, the first cluster includes the leading villain, Thenardier and persons related to him: MmeThenardier (his wife); Anzelma (his daughter); Magnon (the servant of Marius); six bandits (Brujon, Claquesous, Boulatruelle, Gueulemer, Babel, and Montparnasse).

The second cluster includes the Bishop Myriel; MlleBaptistine and MmeMagloire (two servants of Myriel); and some persons have important relations with Myriel during his life: Napoleon; CountessDeLo; Geborand; Champtercier; Cravatte; Count; OldMan.

The third cluster consists of the leading hero Valjean, his enemy and also life-saver Javert (the policeman who had been hunting Valjean all the time, but released Valjean at last), two sisters (Perpetue and Simplice) helped Fantine (the leading heroine) appointed by Valjean, two women (two servants of Valjean, who was also appointed by Valjean to protect the young heroine, Cosette); six protagonists in the "Champmathieu affair": Champmathieu (a thief who was wrongly regarded as Valjean), three convicts accused Champmathieu (Cochepaille, Brevet, and Chenildieuthree), Bamatabois (a juror), a judge; and

several persons who did not treat Valjean well after he was released from jail: Labarre (an innkeeper), Gervais (a little boy), Isabeau (a baker), MmeDeR, Sacufflaire (a horse merchant); three persons met by Valjean in relatively later chapters: Fauchelevent (an aged notary), Gribier (a gravedigger), MotherInnocent (prioress of a convent).

The fourth cluster corresponds to the leading heroine, Fantine; Marguerite (one woman who helped Fantine); her three friends: Zephine, Favourite, and Dahlia; the four Parisian students (Tholomyes, Listolier, Fameuil, and Blacheville), who often contacted with Fantine.

The fifth cluster consists of the young heroine, Cosette (the wife of Marius) and Toussaint (who was appointed by Valjean to protect Cosette).

The sixth cluster consists of the young hero, Marius; Pontmery (one of the family members of Marius); Eponine and Baroness (two friends of Marius); the family members of Marius: Gillenormand, MlleGillenormand, LtGillenormand, MmePontmercy, and one friend of Marius, MlleVaubois.

The seventh cluster contains the eight members of the *friends of the ABC society* (Courfeyrac, Enjolras, Combeferre, Pouvarie, Feuily, Bahorel, Grantaire, and Joly); MmeHucheloup (the innkeeper); and their friends, Mabeuf (the church prefect) and MotherPlutarch (the mad of Mabeuf); Gavroche (a street urchin); Jondrette (the father of Gavroche); MmeBurgon (a landlady); two children.

APPENDIX B: A SIMPLER AGGLOMERATIVE HIERARCHICAL CLUSTERING METRIC

In the main manuscript, we used the geodesic distance matrix C to compute the similarity between two nodes in the tree. This similarity was defined as the Spearman distance between row pairs in the geodesic distance matrix C.

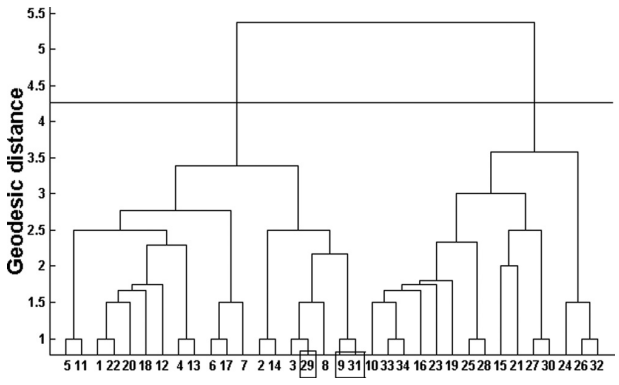

FIG. 10. The hierarchical cluster of the MST of the Zachary's karate club network. Note that the labels of 34 nodes in the MST and the dendrogram correspond to the 34 labels in Fig. 5. The solid line between the first and the second hierarchy provides the clusters when using the simple metric described in Appendix B. The cluster assignment of nodes 29, 9, and 31 (squared) was inconsistent with that of the underlying weighted network (see Fig. 5). Note that using a more intricate metric for node distance as input for the clustering algorithm results in more accurate clustering (Figure 6) than when using the simple metric.

Here, we evaluated the use of an alternative, simpler metric: that is, the agglomerative hierarchical clustering was based directly on the geodesic distances between the node pairs.

Here, we take the MST of the weighted Zachary's karate club network as an example to show the reduced sensitivity when using this simple metric. Fig. 10 shows the hierarchical clustering for the MST. Besides node 29, which was classified correctly using our original approach, nodes 9 and 31 are now also classified incorrectly. This demonstrates that using a simpler metric reduces the sensitivity as compared to the TAHC method. In fact, node 9 and its leave node 31 are located almost in between the two known clusters (see Figs. 5 and 6(a)), which is difficult to detect by simply considering the geodesic distance.

[1]M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404–409 (2001).
[2]R. Albert, H. Jeong, and A.-L. Barabási, Nature **401**, 130–131 (1999).
[3]C. J. Stam, Nat. Rev. Neurosci. **15**, 683–695 (2014).
[4]S. Fortunato, Phys. Rep. **486**, 75–174 (2010).
[5]M. E. J. Newman, Nat. Phys. **8**, 25–31 (2012).
[6]P. Van Mieghem, *Graph Spectra for Complex Networks* (Cambridge University Press, Cambridge, 2011).
[7]For undirected connected graphs, the graph density is defined as $D = \frac{2L}{N(N-1)}$ (N denotes the number of nodes, and L is the number of links.). Completed graphs have the maximum number of links $\frac{N(N-1)}{2}$, so the density of completed graphs is maximal ($D = 1$). Trees, as the connected graphs with N-1links, are the maximally sparse connected graphs ($D = \frac{2}{N}$).
[8]Z. Wu, L. A. Braunstein, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **96**, 148702 (2006).
[9]T. S. Jackson and N. Read, Phys. Rev. E **81**, 021130 (2010).
[10]P. Van Mieghem and S. Van Langen, Phys. Rev. E **71**, 056113 (2005).
[11]P. Van Mieghem and S. M. A. Magdalena, Phys. Rev. E **72**, 056138 (2005).
[12]H. Wang, J. M. Hernandez, and P. Van Mieghem, Phys. Rev. E **77**, 046105 (2008).
[13]B. C. Van Wijk, C. J. Stam, and A. Daffertshofer, PLoS One **5**(10), e13701 (2010).
[14]C. J. Stam, P. Tewarie, E. Van Dellen, E. C. W. Van Straaten, A. Hillebrand, and P. Van Mieghem, Int. J. Psychophysiol. **92**, 129–138 (2014).
[15]D. J. Watts and S. H. Strogatz, Nature **393**, 440–442 (1998).
[16]A.-L. Barabási and R. Albert, Science **286**, 509–512 (1999).
[17]E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2003).
[18]M. Scales-Pardo, R. Guimera, A. A. Moreira, and L. A. N. Amaral, Proc. Natl. Acad. Sci. U.S.A. **104**(39), 15224–15229 (2007).
[19]A. Clauset, C. Moore, and M. E. J. Newman, Nature **453**, 98–101 (2008).
[20]P. Van Mieghem, *Performance Analysis of Complex Networks and Systems* (Cambridge University Press, Cambridge, 2014).
[21]M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002).
[22]The similarity between two nodes of a graph quantitatively measures how similar (one of) their network properties are. In social network studies, *structural equivalence* has been used to quantitatively measure the degree of similarity between two nodes. Structural equivalence reflects how many identical entries two nodes have in their corresponding rows of the adjacency matrix, and can be computed as the Euclidean distance or Spearman's rank correlation between the rows.
[23]V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, J. Stat. Mech. **2008**, P10008 (2008).
[24]M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[25]S. Fortunato and M. Barthélem, Proc. Natl. Acad. Sci. U.S.A. **104**, 36 (2007).
[26]J. M. Kumpula, J. Saramaki, K. Kaski, and J. Kertesz, Eur. Phys. J. B **56**, 41 (2007).
[27]J. P. Bagrow, Phys. Rev. E **85**, 066118 (2012).
[28]J. Kim and T. Wilhelm, Phys. Rev. E **87**, 032816 (2013).
[29]H. A. Bethe, Proc. R. Soc. London, Ser. A **150**, 552–575 (1935).
[30]K. Pearson, Proc. R. Soc. **58**, 240–242 (1895).

[31]P. Van Mieghem, *Data Communciation Networking* (Techne Press, Delft, 2011).
[32]L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech.: Theory Exp. **9**, P09009 (2005).
[33]J. B. Kruskal, Proc. Am. Math. Soc. **7**, 48–50 (1956).
[34]W. W. Zachary, J. Anthropol. Res. **33**, 452–473 (1977), see http://www.jstor.org/stable/3629752.
[35]D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).
[36]J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).
[37]A. Bettinelli, P. Hansen, and L. Liberti, Phys. Rev. E **86**, 016107 (2012).
[38]G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, Nature **435**(7043), 814–818 (2005).
[39]J. C. Gower and G. J. S. Ross, J. R. Stat. Soc.: Ser. C **18**, 54–64 (1969), see http://www.jstor.org/stable/2346439.
[40]M. Tumminello, C. Coronnello, F. Lillo, S. Micciché, and R. N. Mantegna, Int. J. Bifurcation Chaos **17**, 2319–2329 (2007).