Community Detection Based on Girvan Newman Algorithm and Link Analysis of Social Media

K. Sathiyakumari(✉) and M.S. Vijaya

PSGR Krishnammal College for Women, Avinashi Road, Peelamedu, Coimbatore 641004, Tamilnadu, India
{sathiyakumari,msvijaya}@psgrkc.com

Abstract. Social networks have acquired much attention recently, largely due to the success of online social networking sites and media sharing sites. In such networks, rigorous and complex interactions occur among numerous one-of-a-kind entities, main to massive statistics networks with notable enterprise capacity. Community detection is an unsupervised learning task that determines the community groups based on common interests, occupation, modules and their hierarchical organization, using the information encoded in the graph topology. Finding communities from the social network is a difficult task because of its topology and overlapping of different communities. In this research, the Girvan-Newman algorithm based on Edge-Betweenness Modularity and Link Analysis (EBMLA) is used for detecting communities in networks with node attributes. The twitter data of the well-known cricket player is used right here and community of friends and fans is analyzed based on three exclusive centrality measures together with a degree, betweenness, and closeness centrality. Also, the strength of extracted communities is evaluated based on modularity score using proposed method and the experiment results confirmed that the cricket player's network is dense.

Keywords: Edge-Betweenness · Modularity · Degree · Closeness · Community detection · Social network

1 Introduction

The developing use of the internet has brought about the development of networked interplay environments inclusive of social networks. Social networks are graph structures whose nodes represent people, organizations or other entities, and whose edges represent relationship, interaction, collaboration, or influence between entities. The rims in the network connecting the entities might also have a path indicating the drift from one entity to the other; and a power denoting how plenty, how often, or how critical the connection is. Researchers are increasingly interested in addressing a wide range of challenges exist in these social network systems.

In recent years, social community research has been completed the use of large amount of data collected from on-line interactions and from explicit courting hyperlinks in online social community systems including facebook, Twitter, LinkedIn,

S. Subramanian et al. (Eds.): CSI 2016, CCIS 679, pp. 223–234, 2016.
DOI: 10.1007/978-981-10-3274-5_18

Flickr, instant Messenger, and so on. Twitter is pretty rated as a new shape of media and utilized in numerous fields, consisting of corporate advertising and marketing, education, broadcasting and etc. Structural characteristics of such social networks can be explored using socio metrics to understand the structure of the network, the properties of links, the roles of entities, information flows, evolution of networks, clusters/communities in a network, nodes in a cluster, center node of the cluster/network, and nodes on the periphery etc. To discover functionally associated objects from network groups [1, 2] allow us to observe interaction modules, lacking characteristic values and are expecting unobserved connections among nodes [3]. The nodes have many relationships among themselves in communities to percentage common homes or attributes. The identifying community is a trouble of clustering nodes into small communities and a node may be belonging to a couple of communities straight away in a community structure.

Two exclusive assets of facts are used to carry out the clustering mission, first is ready nodes and its attributes and the second one is ready the connection between nodes. The attributes of nodes in community structure are known properties of users like network profile, author publication, publication histories which helps to determines similar nodes and community module to which the node belongs. The connection between nodes provides information about friendships, authors collaborate, followers, and topic interactions.

A few clustering algorithms [4, 5] employ node attributes but ignores the relationships among nodes. However, the network detection algorithms use corporations of nodes which can be densely linked [6, 7] but ignore the node attributes. By means of the use of those two sources of data, the positive algorithm fails to describe the critical shape in a community. For example, attributes may also inform about which community node with few hyperlinks belonging to and it is difficult to determine from community shape alone. On the contrary, the community offers detail about nodes belongs to the equal community even someone of the nodes has no attributes values. Node attributes can balance the network shape which ends up in an extra correct detection of communities. Thus community detection becomes challenging task when taking into account of both nodes attributes and network topology.

The proposed method overcomes the above hassle by identifying groups based totally on the node and its attributes with the aid of implementing Girvan-Newman set of rules.

2 Related Work

A community is a densely linked subset of nodes that is sparsely connected to the remaining network. Social networks are a combination of vital heterogeneities in complex networks, which includes collaboration networks and interaction networks. Online social networking applications are used to represent and model the social ties among people. Finding groups inside an arbitrary community may be a computationally hard task. Various research dealings in recent past years have been conducted on the topic of community detection and some of the important research works are mention below.

Chen and Yuan have referred to that counting all feasible shortest paths in the calculation of the brink betweenness can make unbalanced partitions, with groups of very distinctive length, and proposed to rely on handiest non-redundant paths, i.e. paths whose endpoints are all special from every different. The resulting betweenness confirmed higher consequences than general facet betweenness for blended clusters at the benchmark graphs of Girvan and Newman. Holme et al. have used a changed model of the algorithm wherein vertices, instead of edges, have been removed. A centrality measure for the vertices, proportional to their web page betweenness that became inversely proportional to their in-degree turned into selected to perceive boundary vertices, which have been then iteratively removed with all their edges. Only the in-degree of a vertex becomes used as it indicated the number of substrates to a metabolic reaction concerning that vertex [8].

One of the most popular algorithms changed into provided via Newman and Girvan (denoted GN) [9, 10] which turned into a divisive hierarchical clustering set of rules. Edge removal divided network to groups, the rims to take away had been chosen through the usage of betweenness measure. The concept changed into that if companies are related by a few edges between them, then all the paths between vertices in a single group to vertices in different companies blanketed these edges. Paths give ratings to edges betweenness, with the aid of accounting all the paths passing via each aspect and removing the threshold with the maximal rating, hyperlinks inside the community had been broken. This system was repeated and turned into divided into smaller paths until a stop criterion is reached, this criterion become modularity. A hybrid model of this method in [11] and a faster version primarily based on the equal strategy in [12] become proposed.

Approaches to network detection based totally on the genetic set of rules have been available in [13–15]. A genetic method proposed by using [16] applied an algorithm that used a health characteristic which recognized businesses of vertices within the network that have dense intra connections and sparser inter connections.

In [17, 18] authors proposed a genetic set of rules that uses Newman and Girvan fitness feature for measuring network modularity. Characters become covered of N genes that N changed into the node range. The i^{th} gene corresponds to a j^{th} node, and its fee becomes the identifier of node i. Authors used a non-fashionable one-manner crossover in which, given two people A and B, a network identifier j was selected randomly, and the identifier j of nodes $j_1,\ldots,j_h$ of A become transferred to the identical nodes of B.

On this research, the Girvan-Newman set of rules based totally on part-Betweenness Modularity and link analysis (EBMLA) is applied for discovering groups in networks with node attributes. The twitter data of the famous cricket participant is taken for have a look at and network of friends and followers is analyzed based totally on 3 different centrality measures along with the degree, betweenness, closeness centrality, and modularity score.

3 Girvan-Newman Algorithm

3.1 Girvan-Newman Algorithm Based on Edge-Betweenness Modularity and Link Analysis

The Girvan and Newman is a general community finding algorithm. It performs natural divisions among the vertices without requiring the researcher to specify the numbers of communities are present, or placing limitations on their sizes, and without showing the pathologies evident in the hierarchical clustering methods. Girvan and Newman [19] have proposed an algorithm which has three definitive features (1) edges are gradually removed from a network (2) the edges to be removed are chosen by computing betweenness scores (3) the betweenness scores are recomputed for removal of each edge.

As a degree of traffic flows Girvan and Newman use part betweenness a generalization to the edge of the renowned vertex betweenness of Freeman [20, 21]. The betweenness of an edge is defined because of the quantity of shortest paths among vertex pairs. This quantity can be calculated for all edges in the time complexity of O (mn) on a graph with m edges and n vertices [22, 23].

Newman and Girvan [23] define a degree called modularity, which is a numerical index that shows proper separation among nodes. For a separation with g organizations, define as g × g matrix e. Whose thing e_{ij} is the fraction of edges in the authentic network that connects vertices in institution i to those in institution j. Then the modularity is described as

$$Q = \sum_{i} e_{ii} - \sum_{ijk} e_{ij} e_{ki} = \mathrm{Tr}\ e - \left\| e^2 \right\|,$$

wherein shows the sum of all elements of x, Q is the fraction of all edges that lie within groups minus the predictable value of the same amount in a graph in which the vertices have the same tiers however edges are positioned at random with outlook upon the communities. The Q = 0 indicates that community shape isn't any more potent than could be expected by using random chance and values other than 0 represent deviations from randomness. Restricted peaks inside the modularity for the duration of the progress of the community shape set of rules imply correct divisions of the community.

3.2 Girvan-Newman Partitioning Algorithm

Successively Deleting Edges of High-Betweenness

Step 1: discover the threshold or multiple edges with the best betweenness, if there may be a tie in betweenness then eliminate the rims from the graph. This technique may additionally spill the graph into numerous components; it makes first level partition of the graph.
Step 2: Recalculate all betweenness values and then remove the edges/edge with high betweenness value. Again split the first level region into several components such that there are nested within larger regions of the graph.
Step 3: Repeat steps (1) and (2) until edges continue to be within graph.

Computing Betweenness Values
For each node A:

Step 1: Do Breath First seek to start at node A
Step 2: remember the quantity of shortest paths from A to every different node
Step 3: determine the quantity of flow from A to all other nodes.

3.3 Centrality Measures and Modularity Scores

Centrality measures are used to discover the node's relative significance inside groups by using summarizing structural relation with different nodes. The three simple centrality measure targeted on this work are a degree, closeness, and betweenness.

Degree. The degree centrality represents the wide variety of connections a selected node has. In a directed graph, wherein the route of the node is applicable, there may be a differentiation between the in-degree and out-degree; the quantity of hyperlinks a specific node receives is in-diploma, and the range of links a selected node sends is out-degree. The sum of in-degree and the out-degree offers the degree measure. The following method gives degree and normalized degree centrality ratings.

Degree centrality

$$\mathrm{C_D(v)} = deg(\mathrm{v})$$

Normalized degree centrality

$$\mathrm{C_D(v)} = deg(\mathrm{v})/\mathrm{g} - 1$$

where g is the size of the group.

Closeness. The closeness measure represents imply of the geodesic distances among a particular node with other nodes related to it. It is a measure of ways long a message will take to unfold during the network from a specific node n sequentially. It also describes the speed of the message within social systems. Closeness is primarily based on the period of the average shortest direction among a vertex and all of the vertices in the graph. The subsequent formulation is used to calculate closeness centrality.

Closeness Centrality

$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j)\right]^{-1}$$

Normalized Closeness Centrality

$$C_C^{'}(n_i) = (C_C(n_i))(g-1)$$

Betweenness. The betweenness measure quantifies the quantity of times a node acts as a bridge alongside the shortest direction among other nodes. It's miles a measure for quantifying the manipulate of the node at the conversation between nodes in a social network. It also represents how a long way a message can reach inside a network from a specific node 'n' and additionally describes the span of the message within social systems. Nodes that arise on many shortest paths among other nodes have higher betweenness than those that do not. This is vertices that have an excessive possibility to occur on a randomly chosen shortest direction among two randomly chosen vertices have an excessive betweenness. The following formulas are used to determine betweenness centrality measure.

Betweenness Centrality

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i)/g_{jk}$$

where g_{jk} = the number of geodesics connecting *jk*, and $g_{jk}(ni)$ = the number that actor i is on.

Normalized betweenness centrality measure

$$C_B^{'}(n_i) = C_B(n_i)/[(g-1)(g-2)/2]$$

Modularity. The modularity Q is proposed via Newman and Girvan [23] as a degree of the nice of a selected division of a network, and is defined as follows:

$$\text{Q} = (\text{range of edges inside communities}) - (\text{predicted wide variety of such edges})$$

The modularity Q measures the fraction of the edges within the community that join vertices of the same type, i.e., inside-community edges, minus the expected value of the same quantity in a community with the equal network department however with random connections among the vertices. If the variety of inside community edges is not any higher than random, Q = zero. A price of Q this is near 1, which is the maximum, indicates strong community shape. Q usually falls inside the range from 0.3 to 0.7 and excessive values are rare.

4 Experiments and Results

The proposed framework includes four phases: twitter facts, directed network, Girvan-Newman algorithm, and modularity score. Every phase is described in following sections and the architecture of the proposed system is shown in Fig. 1.

Real time twitter statistics was extracted from twitter API 1.1 the use of R 3.3.1 tool. A directed network is created using Twitter buddies/followers listing as the graph. In this directed community, three centrality measures degree, closeness, and betweenness are used for the stage of evaluation of network Girvan-Newman algorithm

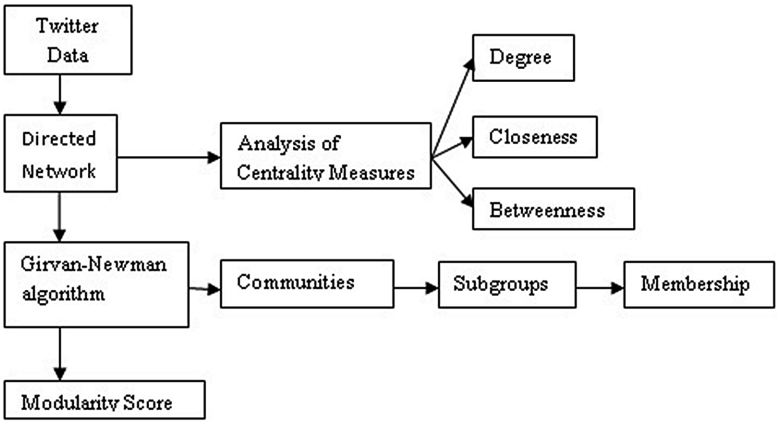

Fig. 1. Community detection framework

is used to detect communities and subgroups. The size of subgroups is found using Girvan-Newman algorithm of this network. The algorithm also detects modularity score of the community.

Girvan-Newman algorithm is implemented for community detection based on edge betweenness. Analysis of the social network is carried out using various centrality measures such as degree, closeness, and betweenness. These centrality measures are evaluated with various properties like minimum and maximum values of in-degree, out-degree, in-closeness, out-closeness, and betweenness. The real-time data is collected using the twitter application programming interface 1.1 for this research work. Nine thousands records of friends and followers list of the famous cricket player have been crawl from his twitter account. The data is collected at run time from twitter network using R3.3.1, a statistical tool.

Figure 2 shows the cricket player's initial community network and Fig. 3 depicts the relationship types of community network such as friends and followers both friends and followers of the initial network. This network has 7095 edges and 6831 vertices.

Degree, closeness, and betweenness are the three centrality measures that are evaluated for the above network using R script. The number of connections that a

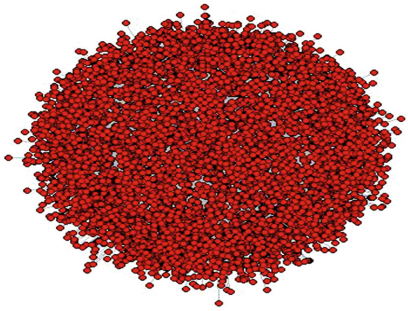

Fig. 2. Cricket player's initial network

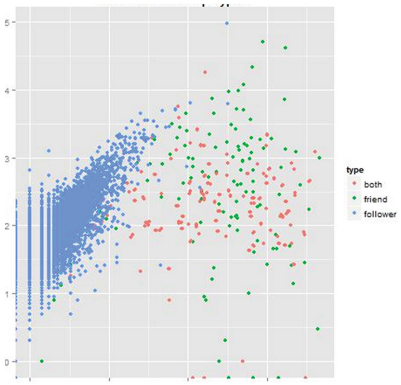

Fig. 3. Friends and followers network

particular node makes is called the degree centrality. The Twitter network is a directed graph and a node encompasses both in-degree and out-degree. The number of arcs from a node to other nodes is out-degree and it is 95 for this network. The in-degree is the number of arcs coming into a node from other nodes and it is 7000 on the same network. The total degree centrality measure is 7095. The histogram representation of in-degree, out-degree and total degree measurement for the cricket player's network is shown in Figs. 4 and 5. The minimum and maximum values of in-degree and out-degree measures are given in Table 1.

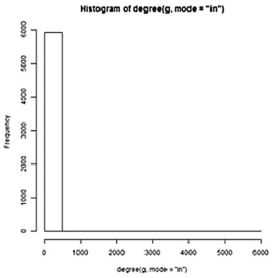

Fig. 4. In-degree of given network

Similarly, the closeness centrality measure is evaluated for the same directed graph. The closeness measure represents the shortest path between nodes connected with it. The out-closeness is 0.0000000217 for this network. The in-closeness is 0.00000161 for the same network. The total closeness centrality is 0.0000000217. Figures 6 and 7 displays the histogram representation of in-closeness and out-closeness of this network. The minimum and maximum values of in-closeness and out-closeness measures are given in Table 1.

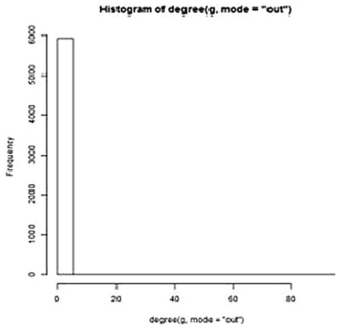

Fig. 5. Out-degree of given network

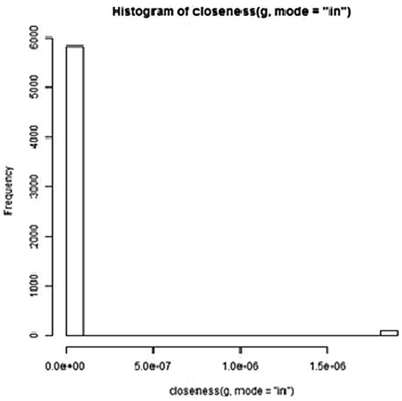

Fig. 6. In-closeness of given network

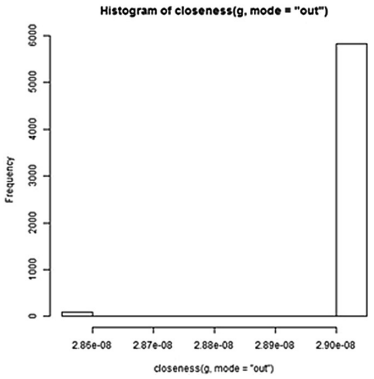

Fig. 7. Out-closeness of given network

The minimum and maximum values of betweenness measures are computed for this cricket player's network in a similar manner and the values are given in Table 1.

Table 1. Evaluation measures for community detection

Measures / Limitations	Community Detection Measures						
	Degree			Closeness			Betweenness
	In	Out	Total degree	In	Out	Total Closeness	
Min	1	1	1	0.0000000214	0.0000000214	0.0000000214	1
Max	7000	95	7095	0.00000161	0.0000000217	0.0000000217	640295
	Modularity Score is : 0.91						

A community consists of a closely connected group of vertices, with only meager connections to other groups. Girvan-Newman Algorithm is used here to find different communities from cricket player's twitter network based on edge betweenness measure. The modularity score for this network is obtained as 0.91. Thirty-nine different communities are extracted for this network based on edge betweenness modularity measure and demonstrated in different colors in Fig. 8. These 39 communities are clustered based on followers, friends, and both followers and friends in the network. The distribution of nodes in various communities is showed in Fig. 9. The membership of size of Community 1 is 69, community 2 has the highest size with 166 memberships. Communities 3 and 4 have the membership sizes 42 and 39 respectively. Communities 5 and 7 have the same membership size 37 and so on.

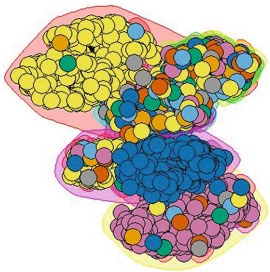

Fig. 8. Community detection using edge betweenness algorithm (Color figure online)

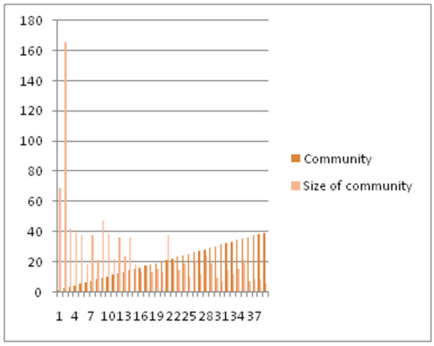

Fig. 9. Community size

5 Discussion and Findings

In this research work, the out-degree is 95 and in-degree are 7000 for the cricket player's network. An entity or node is an active player, hub when it has high degree centrality and obtains an advantaged position in the network. Since closeness centrality is low, the node has slow interaction to other entities in a network. The node has a better influence over the other nodes in the network and is in a powerful position because the betweenness centrality is 468 for this network. The modularity score obtained through Edge-Betweenness and Link algorithm is 0.91; it is proved that the cricket player's friends and followers network are highly dense. Also, the Girvan-Newman algorithm has detected 39 different communities from the cricket player's network and found that out of 39 communities, 5 communities are dense.

6 Conclusion and Future Work

This work elucidates the application of Girvan-Newman algorithm based on Edge-Betweenness and Link Analysis for detecting communities from networks with node attributes and its properties. The real time twitter directed network of a cricket player is used to carry out social network analysis with various centrality measures. The modularity value 0.91 of the tested network by EBMLA confirms that the network is dense and the algorithm is efficient in finding different communities. As a scope for further work more network properties can be used for social network analysis and more interpretation can be drawn. Also, other community detection algorithms can be adopted for detecting communities and finding significant node attributes for each community.

References

1. Coscia, M., Rossetti, G., Giannotti, F., Pedreschi, D.: DEMON: a local-first discovery method for overlapping communities. In: KDD 2012 (2012)
2. Girvan, M., Newman, M.: Community structure in social and biological networks. PNAS **99**, 7821–7826 (2002)
3. Yang, J., Leskovec, J.: Overlapping community detection at scale: a non-negative factorization approach. In: WSDM 2013 (2013)
4. Johnson, S.: Hierarchical clustering schemes. Psychometrika **32**, 241–254 (1967)
5. Fortunato, S.: Community detection in graphs. Phys. Rep. **486**, 75–174 (2010)
6. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and comparative study. ACM Comput. Surv. **45**, 43 (2013)
7. Fortunato, S.: Community detection in graphs. J. Phys. Rep. **486**(3–5), 75–147 (2010)
8. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. J. Phys. Rev. E **69**(2), 026113 (2004)
9. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. In: Proceedings of the National Academy of Science, USA, pp. 7821–7826 (2002)
10. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. J. Phys. Rev. E **69**(6), 066133 (2004)
11. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. J. Phys. Rev. E **70**(6), 066111 (2004)
12. Nandini, R.U., Reka, A., Soundar, K.: Near linear time algorithm to detect community structures in large-scale networks. J. Phys. Rev. E **76**(3), 036106 (2007)
13. Guardiola, X., Guimera, R., Arenas, A., Guilera, A.D., Antonio, L.: Macro- and micro-structure of trust networks, pp. 1–5 (2002). arXiv:cond-mat
14. Narasimhamurthy, A., Greene, D., Hurley, N., Cunningham, P.: Scaling community finding algorithms to work for large networks through problem decomposition. In: 19th Irish Conference on Artificial Intelligence and Cognitive Science (AICS 2008), Cork, Ireland (2008)
15. Pizzuti, C.: Community detection in social networks with genetic algorithms. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation, pp. 1137–1138 (2008)
16. Tasgin, M., Bingol, H.: Communities detection in complex networks using genetic algorithm. In: Proceeding of the European Conference on Complex Systems (ECSS), UK, pp. 1–6 (2006)
17. Tasgin, M., Herdagdelen, A., Bingol, H.: Communities detection in complex networks using genetic algorithms. Community detection in complex networks using genetic algorithms. arXiv preprint arXiv:0711.0491, pp. 1–6 (2007)
18. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. **99**, 7821–7826 (2002)
19. Freeman, L.C.: A set of measures of centrality based upon betweenness. Sociometry **40**, 35–41 (1977)
20. Anthonisse, J.M.: The rush in a directed graph. Technical report BN9/71, Stichting Mathematicsh Centrum, Amsterdam (1971)
21. Newman, M.E.J.: Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality. Phys. Rev. **E64**, 016132 (2001)
22. Brandes, U.: A faster algorithm for betweenness centrality. J. Math. Sociol. **25**, 163–177 (2001)
23. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks (2003). Preprint arXiv:cond-mat/0308217