ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 200–205
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 J. Coroničová Hurajová, T. Madaras

ITAT

Revising the Newman-Girvan algorithm

Jana Coroničová Hurajová[1]* Tomáš Madaras[2]

[1] The Faculty of Business Economics with seat in Košice,
The University of Economics in Bratislava,
Tajovského 13, 041 30 Košice, Slovakia
jana.coronicova.hurajova@euke.sk

[2] The Faculty of Sciences,
P.J. Šafárik University in Košice,
Jesenná 5, 040 01 Košice, Slovakia,
tomas.madaras@upjs.sk
WWW home page: umv.science.upjs.sk

Abstract: One of the common approaches for the community detection in complex networks is the Girvan-Newman algorithm [5] which is based on repeated deletion of edges having the maximum edge betweenness centrality. Although widely used, it may result in different dendrograms of hierarchies of communities if there are several edges eligible for removal (see [6]) thus leaving an ambiguous formation of network subgroups. We will present possible ways to overcome these issues using, instead of edge betweenness computation for single edges, the group edge betweenness for subsets of edges to be subjects of removal.

1 Introduction

One of fundamental analyses performed in the exploration of complex networks concerns the detection of their community structure, which means to find, within a graph representing the network, certain clusters of vertices which are, at one side, sparsely interconnected and, on the other side, they have dense in-cluster links by many edges. The vertices within a cluster show a kind of similarities and form functionally compact units. As there is no general definition of cluster, there are many ways to obtain collections of network communities; a comprehensive overview of contemporary state-of-art in this area can be found in [3].

Among the approaches that determine the graph communities by breaking it into smaller parts, an important role plays the Girvan-Newman algorithm described first in [5]. It is based on successive deletion of edges which have the maximum *edge betweenness centrality* which is the quantity measuring the frequency of appearance of an edge on geodesic paths in a graph. Formally, it is defined as the sum $B(e) = \sum_{u,v \in V(G)} \frac{\sigma_{u,v}(e)}{\sigma_{u,v}}$ where $\sigma_{u,v}$ is the number of shortest $u-v$-paths and $\sigma_{u,v}(e)$ is the number of shortest $u-v$-paths which contain the edge e. Interpreting the edge betweenness as an amount of information flow being propagated through a link between actors of a complex network (and assuming that the information exchange takes place mainly on shortest paths), one may argue that distinct communities within a network are mutually connected (and, hence, communicating) with relatively few edges whose edge betweenness is higher than of those ones between the actors of the same community. When deleting those edges, the network tends to simplify, eventually breaking into smaller subnetworks (note, however, that after each deletion, edge betweenness centralities of the resulting network shall be recalculated again). Thus, we obtain a sequence of graphs starting from the original one and ending with an edgeless graph, along with the sequence of partitions the vertex set (the initial partition is the whole set, the final one consists of isolated vertices); when two consecutive graphs differ in their connected components, we record the splitting (refining) of the partition of the predecessing graph. In this way, the sequence of partitions forms a dendrogram showing the hierarchy of communities within the graph (the choice of the appropriate level describing, in the best way, the community structure of the graph, is a matter of external decision and does not follow from algorithm).

Despite the elegancy of Girvan and Newman approach and the popularity of their algorithm, an attention recently turns to other methods, mainly due to the fact that they are quicker (the Girvan-Newman algorithm has, in general, the complexity $O(m^2 \cdot n)$, thus can be effectively used on graphs up to $n \sim 10000$, see [3]). Furthermore, it seems that many implementations of community detection algorithms which are based on recursive edge deletion do not make difference when equivalent edges (for example, with the same edge betweenness) are considered for deletion. This issue was adressed in [6] where it was demonstrated how the random deletion of different edges with the same maximum edge betweenness centrality results in different hierarchies of partitions, when used on the wheel graph W_6. A possible obvious suggestion to remove all such edges at once (as discussed, for example, in [1] in the connection with possible speeding up the

*Research supported by the project for young teachers, researchers and PhD. students No. I-16-104-00

original Girvan-Newman algorithm) would, however, individualize all the vertices (thus producing no reasonable hierarchy) – even at the very beginning of the process – of edge transitive graphs, and, more generally, of so called edge betweenness-uniform graphs (that is, the graphs whose edges have the same value of edge betweenness centrality). Such graphs are not so rare: in [4], it was shown that each strongly regular graph (that is, an n-vertex k-regular graph with the property that any pair of its adjacent vertices has λ common neighbours, and any pair of its nonadjacent vertices has μ common neighbours, for certain n,k,λ,μ) is edge betweenness-uniform; since it is also known that, for particular n,k,λ,μ, the number of nonisomorphic strongly regular graphs is at least exponential in terms of number of vertices (see [2]). We tested all edge betweenness-uniform graphs on 3–10 vertices (their list was published first in [6]) for communities using the procedures `FindGraphCommunities[...,Method -> "Centrality"]` (to obtain the list of sets of vertices forming communities) and `CommunityGraphPlot[...]` (to visualize communities within a graph) of Wolfram Mathematica, or using the procedure `IGCommunitiesEdgeBetweenness[...]` from the Wolfram Mathematica third-party package IgraphM (see [7]). The results for graphs on 3–9 vertices are shown in Figure 1 (the brown clusters correspond to graph communities based on partitioning by `FindGraphCommunities[...,Method -> "Centrality"]` procedure, the yellow clusters to the ones based on `IGCommunitiesEdgeBetweenness` procedure); one can see that, on some graphs, the community structure is different although the underlying algorithm should be the same (the most remarkable difference can be observed on the blue-highlighted 9-vertex graph of Figure 2 obtained from two 6-cycles by identifying the corresponding vertices of their maximum independent sets). Also, for many of these graphs it seems that they have no community structure (as both built-in and IGraphM community finding procedure aggregate all vertices into a single cluster); hence, when being processed by algorithm of [1], one would have only two possibilities for communities: either the whole vertex set or the partition consisting of singletons (and the more reasonable choice would be the single community). Nevertheless, our results show that there are also edge betweenness-uniform graphs for which both procedures (as well as other community detection methods, like modularity maximization) return non-trivial community structure which, however, cannot be obtained by the algorithm of [1].

2 The revised Newman-Girvan algorithm

In order to overcome – at least, on theoretical basis – the problem to decide which edge has to be removed if there are several ones with the same maximal edge betweenness, we will utilize the concept of *group edge betweenness* which is defined, for a subset A of edge set of a graph G, as the sum $B(A)=\sum_{u,v\in V(G)}\frac{\sigma_{u,v}(A)}{\sigma_{u,v}}$ where $\sigma_{u,v}(A)$ is the number of shortest $u-v$-paths which contain at least one edge from A. Note that if $|A|=1$ then we obtain the standard edge betweenness as in [5]. The revised Newman-Girvan algorithm on a graph G then proceeds as follows: starting with $G_0=G$, a sequence $\{G_i\}_{i=0}^{k}$ (where G_k is edgeless graph) is constructed in such a way that, if there appears, during the computation of edge betweennesses of G_i, a set M_i of $m_i\geq 2$ edges all of them having the maximum betweenness among the edges of G_i, then determine the smallest ℓ_i such that there is unique subset $\widehat{E_i}\subseteq M_i$ of ℓ_i edges with the property that the group edge betweenness centrality of $\widehat{E_i}$ is the maximal among all subsets of M_i consisting of ℓ_i edges (note that $\ell_i\leq m_i$, thus it is well defined). The graph G_{i+1} is then obtained from G_i by removing all edges from $\widehat{E_i}$; if G_{i+1} has more components than G_i, the vertex sets of its components forms the new level in hierarchy of partitions of $V(G)$. The pseudocode for this process is given in Algorithm 1; the used notation follows the common standards of graph theory, the particular specialized symbols are $b_0(G)$ (the zeroth Betti number of G, that is, the number of its connected components), $\langle V_i\rangle$ (the subgraph of G induced by the set $V_i\subseteq V(G)$) and $G\setminus E_i$ (the subgraph of G obtained by deleting all edges of E_i).

We have implemented the key elements of the algorithm in Wolfram Mathematica 10 along with the algorithm for edge group betweenness calculation. Since the latter algorithm is – according to our knowledge – not yet known to have effective implementation, we used the straightforward approach which determines, for each pair u,v of vertices of a graph G, all shortest $u-v$-paths (in Wolfram language, this can be done by calling procedure `FindPath[G, u, v, GraphDistance[G, u, v], All]` and then checks how many of them passes through an edge of the given edge group. Our implementation of the edge group betweenness algorithm – when being called on a single edge – is also useful as an alternative for built-in Wolfram Mathematica procedure `EdgeBetweennessCentrality[..]` which returns numerical approximations (although with high precision) of edge betweenness centralities whereas our version returns exact values in the form of fractions.

Let us note that our approach may lead, in particular cases, to much worse performance of the corresponding algorithm when compared with the original Newman-Girvan algorithm; this is caused mainly by large number of subsets of edges with the same maximum edge betweenness which have to be checked to select the unique one with the maximum group edge betweenness. This is, however, the trade-off for getting rid of uncertainties in edge removal.

To show the difference of behaviour of our algorithm in comparison with the original one or the one of [1], consider the graph of Figure 3. It contains

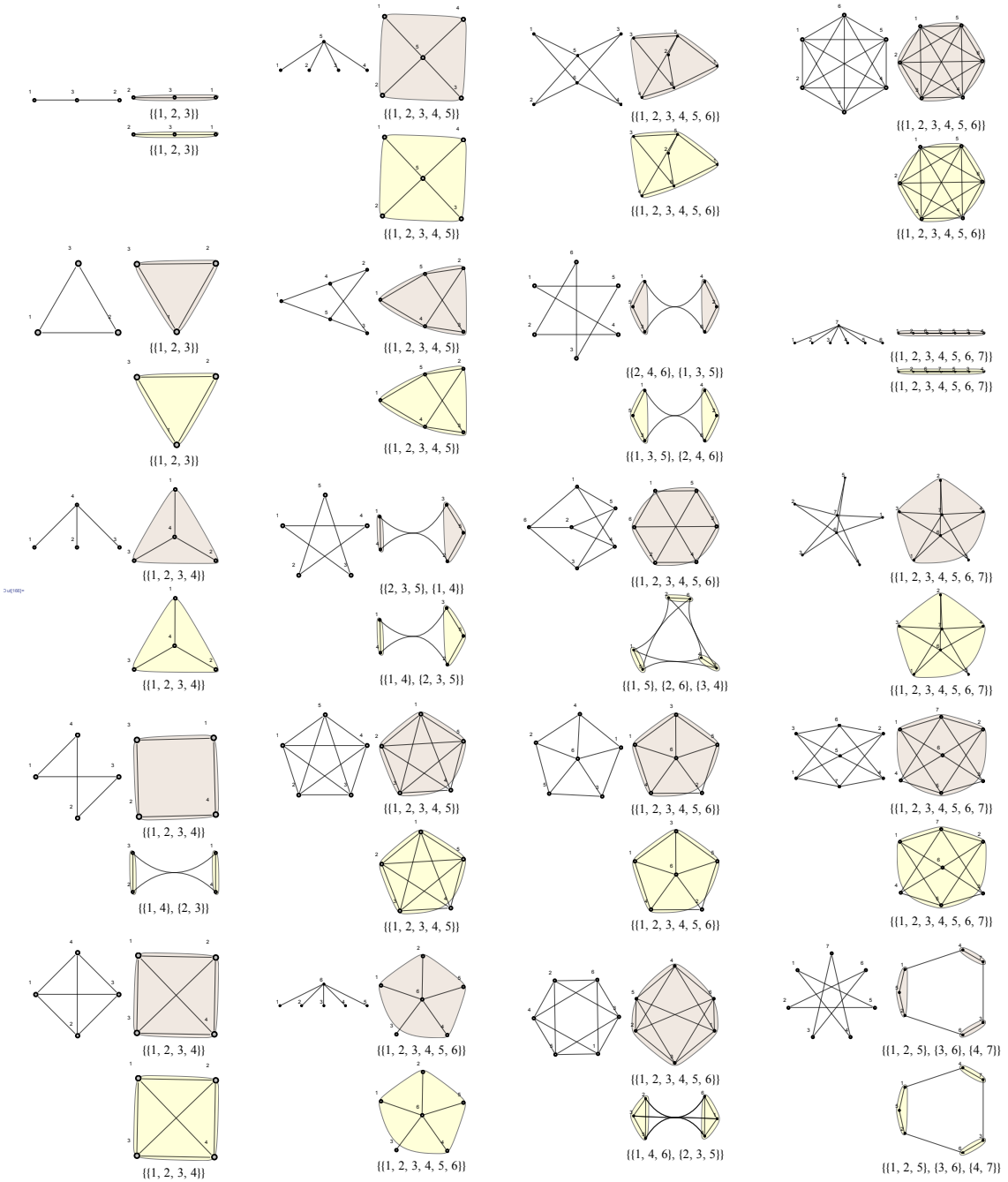

Figure 1: The communities in edge betweenness-uniform graphs detected on basis of edge betweenness

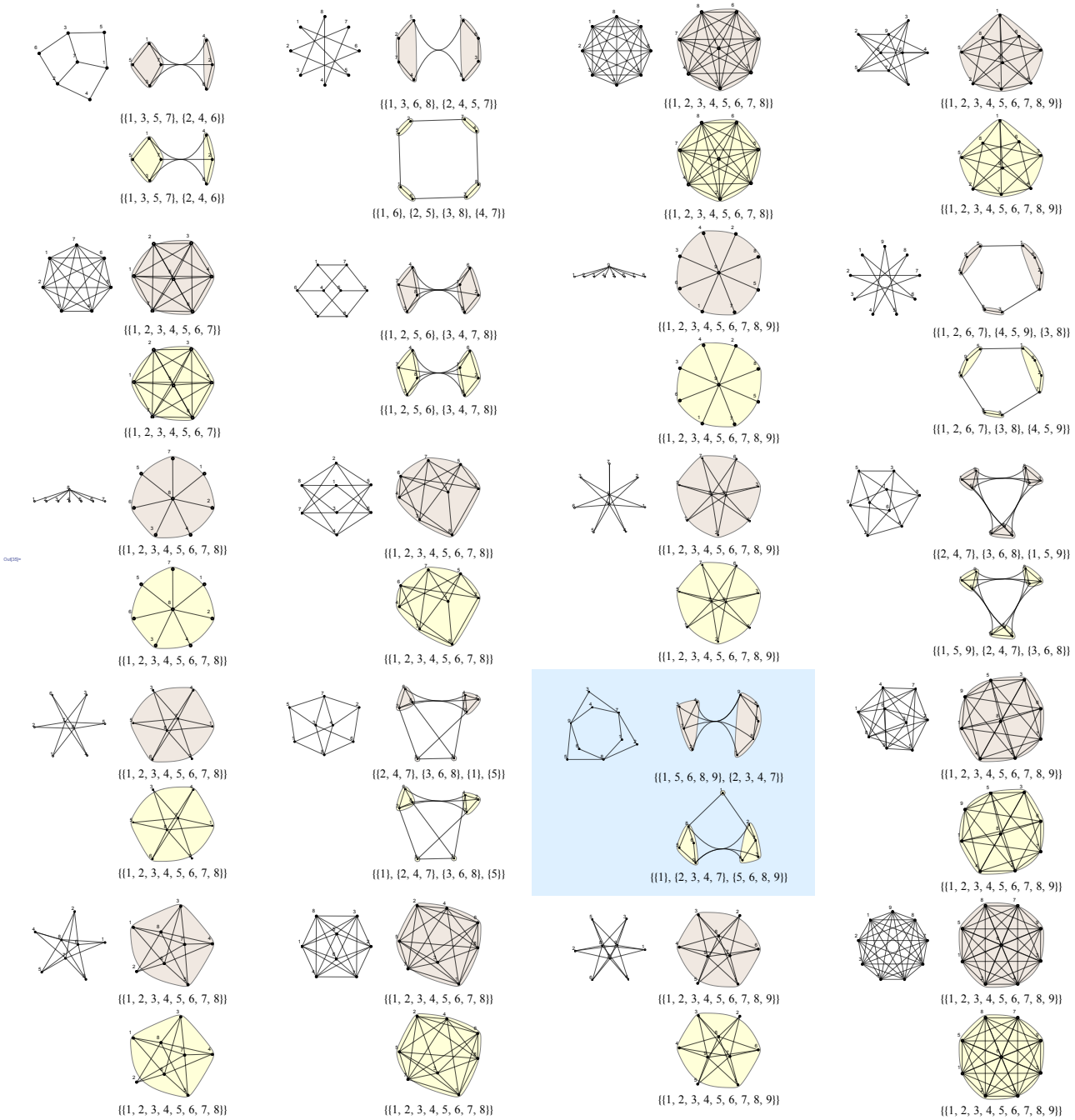

Figure 2: The communities in edge betweenness-uniform graphs detected on basis of edge betweenness (continued)

```
RevisedNewmanGirvan(G)
Data: a graph G
Result: the hierarchy of nested partitions of V(G)
G_0 := G ;
P_0 := {V(G)} ;
i := 0 ;
c := 0 ;
while E(G_i) ≠ ∅ do
    S_i := {B(e) : e ∈ E(G_i)} ;
    mx := max S_i ;
    M_i := {e ∈ E(G_i) : B(e) = mx} ;
    Ê_i := M_i ;
    ℓ_i := 1 ;
    while |Ê_i| > 1 do
        ℓ_i := ℓ_i + 1;
        U_{ℓ_i} := {B(A) : A ⊂ M_i ∧ |A| = ℓ_i} ;
        gmx := max U_{ℓ_i} ;
        Ê_i := {A ⊂ M_i : |A| = ℓ_i ∧ B(A) = gmx} ;
    end
    G_{i+1} := G_i \ Ê_i ;
    if b_0(G_{i+1}) > b_0(G_i) then
        c := c + 1 ;
        P_c := {V_1, ..., V_{r_c} :
        ⟨V_1⟩, ..., ⟨V_{r_c}⟩ are connected components of G_{i+1}}
        ;
    end
    i := i + 1 ;
end
return {P_i : i = 0, ..., c}
```

Algorithm 1: The group edge-centrality based Newman-Girvan algorithm for community detection.

five edges with the maximum edge betweenness, namely $\{8,11\},\{6,9\},\{3,10\},\{3,9\}$ and $\{2,11\}$. Now, these five edges form ten 2-element subsets with group edge betweenness centralities $\frac{52}{3},\frac{52}{3},\frac{52}{3},\frac{49}{3},\frac{52}{3},\frac{49}{3},\frac{52}{3},16,\frac{52}{3},\frac{52}{3}$, and ten 3-element subsets with group edge betweenness centralities $26,25,25,\frac{74}{3},25,25,\frac{71}{3},26,25,\frac{74}{3}$; we see that, among these subsets, the uniqueness with respect to the maximum group edge betweenness is not preserved. But, for five 4-element subsets of $\{\{8,11\},\{6,9\},\{3,10\},\{3,9\},\{2,11\}\}$, the group edge betweenness centralities are $\frac{97}{3},\frac{101}{3},\frac{98}{3},\frac{97}{3},\frac{97}{3}$, thus there is unique 4-element subset – the set $\{\{8,11\},\{6,9\},\{3,10\},\{2,11\}\}$ – reaching the maximum value $\frac{101}{3}$. Hence, the edges of this 4-element set are removed, and the sequence of single edge removals continues with $\{1,12\},\{6,7\},\{4,9\}$ after which there are detected two edges ($\{4,7\}$ and $\{1,7\}$) with the same highest edge betweenness. After they are removed, the edge removal continues with $\{2,12\}$ and then with $\{2,5\}$ where the graph splits, for the first time, into two components with vertex sets $\{1,2,4,6,10,11\}$ and $\{3,5,7,8,9,12\}$.

On the other hand, when the algorithm of [1] is used on the same graph, first, five edges $\{8,11\},\{6,9\},\{3,10\},\{3,9\},\{2,11\}$ are removed at once, followed by sequential removals of $\{3,12\},\{7,8\}$ and $\{8,12\}$ where the graph splits into two components, one of them being the single edge $\{3,8\}$. Therefore, we see that the hierarchies of nested partitions produced by our algorithm and the one of [1] differ already at the highest level. In addition, a particular run of the original Newman-Girvan algorithm (using random selection of an edge from the set of several edges with the same maximum edge betweenness) on this graph may produce yet another hierarchy: if the edge $\{3,9\}$ is removed first, then the edges $\{3,12\},\{3,8\}$ and $\{3,10\}$ are removed sequentially, thereby separating the single vertex 3 from the rest of the graph.

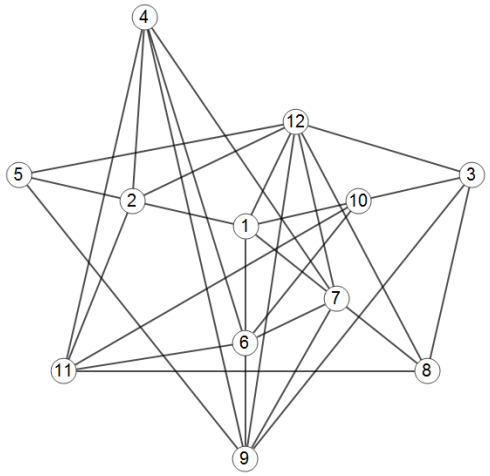

Figure 3: The example of a graph where revised Newman-Girvan algorithm behaves differently than the original one or the one of [1]

Note also that the existence of unique set of edges which have to be removed depends heavily also on the edge automorphism group Aut^* of a graph. It is easy to see that, for any edge automorphism φ of a graph G and any $A \subset E(G)$, $B(A) = B(\varphi(A))$ holds; consequently, if $Aut^*(G)$ is nontrivial and the edge automorphisms do not fix A, then there are several different subsets of edges with the same group edge betweenness as A. Thus the uniqueness of the edge subset of particular size with the maximal group edge betweenness cannot be guaranteed for graphs possessing a lot of symmetries. Some particularly bad examples occur among edge betweenness uniform graphs – in the wheel W_6, among all sets of edges of cardinality $i \leq 9$, there are always at least two distinct sets whose group edge betweenness is maximal among all i-sets, hence, the unique maximum group edge betweenness set coincides with the whole edge set of W_6, and the revised Newman-Girvan algorithm breaks the vertex set of W_6 into six singletons.

Unfortunately, similar issues may appear also in real networks. We illustrate this on the example of the network of Zachary karate club [8] shown at Figure 4.

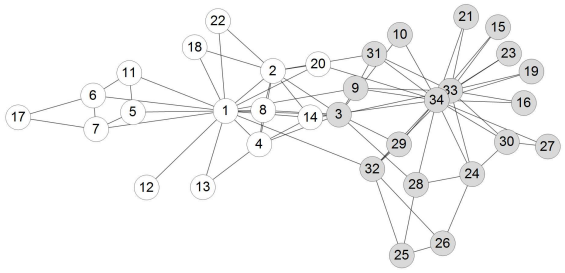

Figure 4: The Zachary karate club network

The standard Newman-Girvan algorithm removes the edges with the maximum edge betweenness in the order $\{32,1\},\{3,1\},\{9,1\},\{34,14\},\{34,20\},\{33,3\},\{31,2\}$, $\{3,2\},\{4,3\}$ after which two edges with the same maximum edge betweenness are detected, namely $\{14,3\}$ and $\{8,3\}$. After their simultaneous removal, the graph splits into two components and the sequence of removed edges continues with $\{34,10\},\{34,28\},\{10,3\}$ after which again two maximum betweenness edges, $\{7,1\}$ and $\{6,1\}$, are detected; their removal yields another pair of edges with the same maximum edge betweenness, namely $\{1,5\}$ and $\{1,11\}$. The sequence of single edge removals continues with edges $\{34,32\},\{33,32\},\{34,29\}$, $\{26,24\},\{28,24\},\{9,3\}$ followed by simultaneous removal of $\{34,27\}$ and $\{1,12\}$, then by single removals of $\{30,27\},\{13,1\},\{13,4\}$. Now here appears the situation when the graph contains even 10 edges of maximum edge betweenness, namely $\{34,23\},\{34,21\},\{34,19\},\{34,16\},\{34,15\}$, $\{33,23\},\{33,21\},\{33,19\},\{33,16\}$ and $\{33,15\}$ (which are all contained in the same star-like connected component). However, the computation of group edge betweenness for subsets of this edge set reveals that the whole set has to be removed as there are always many proper subsets of smaller sizes having the same maximum group edge betweenness (this is most likely also caused by symmetries of that particular connected component). Hence, for the network of Zachary karate club, the revised algorithm just confirms that the order of edge removal as obtained by the version of Newman-Girvan algorithm from [1] is probably optimal; nevertheless, it would be interesting to find an example of real network where the revised algorithm would lead to different sequence of removed edges, or even a different hierarchy of graph communities.

An area where the revised Newman-Girvan algorithm would apply concerns the looking for "null models", that is, the graphs without community structure (see [3], pages 90–91). Based on the above considerations, we propose to take, for such graphs, the ones which are edge betweenness-uniform, have trivial edge automorphism group and, moreover, for each i which is less than the number of edges, there are always at least two sets consisting of i edges such that their group edge betweenness is the maximal among all i-subsets. For these graphs, the hierarchy of communities produced by our revised algorithm collapses into singletons although their trivial automorphism group should prevent easy replication of subsets of edges with high group edge betweenness. At the moment, no infinite family of such graphs is known; nevertheless, we believe that the candidate graphs might be found among strongly regular graphs, where are known examples with trivial vertex automorphism group, and, possibly, also ones having trivial edge automorphism group.

References

[1] Despalatović, L., Vojković, T., Vukičevič: Community structure in networks: Girvan-Newman algorithm improvement. MIPRO, Opatija, Croatia, May 26-30, 2014, 997–1002.

[2] Fon-der-Flaass, D.G.: New prolific constructions of strongly regular graphs. Adv. Geom. **2** (2002) 301—306

[3] Fortunato, S.: Community detection in graphs. arXiv:0906.0612 [physics.soc-ph]

[4] Gago, S., Coroničová Hurajová, J., Madaras, T.: Betweenness centrality in graphs. In: Quantitative graph theory: mathematical foundations and applications (Dehmer, M. and Emmert-Streib, F., eds.), CRC Press, 2015

[5] Girvan, M., Newman, M. E. J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA **99** (2002) 7821-–7826

[6] Coroničová Hurajová, J., Madaras, T.: The edge betweenness centrality – theory and applications. Journal of innovations and applied statistics **5 (1)** (2015) 20–29

[7] https://github.com/szhorvat/IGraphM

[8] Zachary, W. W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research **33 (4)** (1977) 452—473