A Community Detection Technique for Research Collaboration Networks based on Frequent Collaborators Cores

Miloš Savić
svc@dmi.uns.ac.rs

Mirjana Ivanović
mira@dmi.uns.ac.rs

Bojana Dimić Surla
bdimic@uns.ac.rs

University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics
Trg Dositeja Obradovića 4, 21000 Novi Sad, Serbia

ABSTRACT

Community structure is one of prominent features of complex real-world networks. In this paper we propose a novel technique for detecting communities in research collaboration networks. The main idea of the algorithm is that research communities can be efficiently recovered from subgraphs encompassing frequent collaborators. Moreover, the algorithm can be used to cluster weighted undirected networks from other domains as well. An experimental evaluation of the algorithm was conducted on a co-authorship network representing collaborations between researchers employed at our Department. The results of the evaluation showed that the algorithm identifies strong and meaningful clusters corresponding to groups dealing with specific research topics. Moreover, we compared our method to seven other community detection techniques showing that it performs better or equally with respect to the quality of obtained community structures.

Keywords

Community detection; research collaboration networks; frequent collaborators

1. INTRODUCTION

Intuitively speaking, a community, cluster, cohesive group or module is a subset of nodes in a network that are more densely internally connected than with the rest of the network [6]. A network has a community structure if its nodes can be grouped into either non-overlapping or overlapping communities [3]. Community structure is a typical feature of social networks since people tend to form cohesive groups reflecting specific cultural characteristics, interests, opinions or behavior [12, 6]. The aim of community detection techniques is to identify communities in a network relying only on the structure of the network. Uncovering communities helps us to understand internal structure of complex networks at a higher level of abstraction, to identify cohesive subnetworks, and to obtain readable maps of extremely large networks by constructing their coarse-grained descriptions (networks of communities).

Scientific research is a driving force of economical and technological progress. Research findings are mostly collaborative efforts due to the complexity of contemporary research problems. The social structure of scientific collaboration can be understood by investigating co-authorship networks [11, 21]. Co-authorship networks are undirected weighted graphs where nodes represent researchers, links correspond to research collaborations, and link weights express strengths of research collaborations. The identification of communities in co-authorship networks reveals cohesive research groups having different research interest, as well as their interactions [9].

In this paper we propose a novel and simple community detection technique for co-authorship networks. The starting premise of the algorithm is that research communities are organized around frequent collaborators. The premise is quite reasonable since it is very unlikely that two frequent collaborators belong to different research communities. In contrary, inter-community links, links connecting members from different communities, would have very high weights which is contradictory to the intuitive understanding of clusters in co-authorship networks. In other words, we use link weight as a measure of how intra-communitarian link is: higher weights indicate intra-community links, while lower weights indicate inter-community links.

The rest of the paper is structured as follows. Related work is presented in Section 2. Section 3 describes our approach to community detection, while the experimental evaluation of the approach is presented in Section 4. Finally, in Section 5 we give conclusions and directions for future work.

2. RELATED WORK

The algorithms for community detection started intensively to develop after Newman and Girvan introduced a measure for the quality of a partition of a network into communities called *modularity* [16]. The underlying idea of the modularity measure is that a subnetwork can be considered as a community if the number of links inside the subnetwork is significantly higher than the expected number of links considering some null random graph model. In case of weighted networks modularity accumulates the differences between the total weight of links within a community and its mathematical expectation considering a random network with the same degree and link weight distributions [13].

Although used as a de facto standard, modularity has a weakness known as the resolution limit [7] – community detection techniques based on modularity maximization may fail to identify modules smaller than certain size, even in extreme cases when modules are cliques. Therefore, it is highly important to consider other notions of community when performing community detection relying on the modularity measure. Radicchi et al. [18] proposed definitions of strong and weak communities. Namely, a community in a weighted network is called Radicchi strong if for each node in the community the total weight of incident intra-community links is higher than the total weight of incident inter-community links.

A comprehensive overview of community detection techniques can be found in the article written by Santo Fortunato [6]. He classified existing community detection techniques into the following categories: traditional, divisive, modularity maximization and dynamic methods. Fortunato emphasized that the traditional graph partitioning methods (e.g. the Kernighan-Lin algorithm) and data clustering methods (e.g. k-means) are not widely used due to intrinsic limitations (e.g. the number of clusters has to be specified in advance) or demanding computational complexity.

The main characteristic of divisive methods is that they build clustering dendrograms by progressively removing links that are most likely to be inter-communitarian. The Girvan-Newman algorithm [9] relies on the edge-betweenness measure to detect inter-communitarian links. Other proposed indicators of inter-communitarian links are edge-clustering coefficient [18] and information centrality [8].

The optimization of the modularity measure is known to be a NP-complete problem [2]. Therefore, several researchers proposed various strategies for modularity maximization. The most used ones are greedy modularity maximization strategies [14, 4, 1]. Other approaches include spectral, extremal and simulated annealing maximization of the modularity measure [6].

The main philosophy of dynamic methods is that communities can be recovered by some process running on the network. For example, the Walktrap method [17] is based on the idea that relatively short random walks should be trapped into dense sub-networks due to the high density of intra-community links. Another widely used dynamic method for community detection is Label propagation [19]. The method is based on an iterative process in which each node adopts a label that most of its neighbors have starting from an initial configuration in which nodes have unique labels. As labels propagate through the network, densely connected sets of nodes form a consensus on their labels which determine the membership of nodes to communities.

It is quite common that authors when proposing a new community detection algorithm test it on a co-authorship network. For example, the Girvan-Newman algorithm [9] was experimentally evaluated on four different networks including the co-authorship network of scientists at the Santa Fe Institute. Other widely used community detection techniques, such as Greedy modularity maximization [14], Label propagation [19], Louvain [1] and Walktrap [17], were also experimentally evaluated by their authors on co-authorship networks. Savić et al. [22] investigated five different community detection techniques on a co-authorship network reflecting scientific collaboration in Serbian mathematical journals. The authors showed that the Louvain method gives the best performance considering the cohesiveness of identified communities.

3. COMMUNITY DETECTION BASED ON W-CORES

Through the rest of this Section we will assume that $G = (N, L)$ denotes an undirected weighted network, where N and L are the sets of nodes and links in G, respectively. The weights of links in G will be expressed by a function named $Weight$, $Weight : L \to R$, where R represents the set of real numbers.

The presence of frequent collaborators in co-authorship networks can be formalized through a more general notion of w-cores. A w-core $C_w = (N_w, L_w)$ in G is a maximal subgraph of G such that the weight of each link in C_w is higher or equal than w. W-cores of G can be easily identified by removing all links whose weight is less than w. Let S denote a graph derived from G after the application of previously mentioned link removal scheme. Then w-cores of G are actually non-trivial connected components (components encompassing more than one node) in S. Isolated nodes in S we call *w-isolated* nodes, while nodes belonging to w-cores are called *w-nodes*.

Our community detection algorithm consists of the two following steps:

- Step 1: Determine w-cores in the network. Assign community labels (integers) to w-nodes in the network such that two nodes from the same w-core have the same label and two nodes from different w-cores have different labels (see Algorithm 1).
- Step 2: Deterministically propagate community labels to w-isolated nodes (see Algorithm 2).

Community labels determine membership of nodes to communities, i.e. two nodes with the same community label belong to the same community. Obviously, the number of detected communities will be equal to the number of w-cores in the network since the initial assignment of community labels is completed after Step 1.

We will assign binary states to links and nodes in the network in order to avoid deletion of links in Step 1. Namely, a link will be considered as active if its weight is higher or equal than w. Similarly, a node is active if it is incident to at least one active link. The initial assignment of community labels can be done using simple graph traversal algorithms, such as breadth first search, but considering only active links and nodes as shown in our Algorithm 1.

The propagation of community labels in our community detection algorithm is based on the following principles:

- Safe label propagation: a label is propagated to an w-isolated node if its labeled neighbors have the same label.

Algorithm 1: Initial assignment of community labels

```
input  : G = (N, L) - undirected weighted network
         w – weight threshold
output: G with initial assignment of community labels
         A – set of labeled nodes

// initialization - mark all nodes as unlabeled
foreach n ∈ N do
    n.label := ∞

// determine active links (AL) and nodes (AN)
AL := empty set of links
AN := empty set of nodes
foreach (x, y) ∈ L do
    if Weight(x, y) ≥ w then
        AL := AL ∪ {(x, y)}
        AN := AN ∪ {x, y}

label = 0
foreach n ∈ AN do
    if n.label = ∞ then
        label := label + 1
        Q := empty queue of nodes
        Q.addLast(n)
        n.label := label
        while Q is not empty do
            c = Q.removeFirst()
            NS = G.getNeighbours(c)
            foreach x ∈ NS do
                if x ∈ AN ∧ (c, x) ∈ AL ∧ x.label = ∞ then
                    x.label := label
                    Q.addLast(x)
```

- If any safe propagation is not possible then a tie resolution strategy is employed for unlabeled neighbors of labeled nodes.
- The safe propagation step continues after tie nodes are resolved.

The process of label propagation starts with determining the set of unlabeled nodes whose neighbors are labeled. This set is denoted by C in Algorithm 2. Each node in C is examined in order to check whether it can adopt a label according to the safe label propagation principle. Otherwise, the node is added to the set of tie nodes T. If a label is safely propagated to a node from C then all of its unlabeled neighbors that are not ties are added to C in order to continue with the safe propagation of community labels. When safe propagation is not possible anymore then tie nodes are resolved.

Let LB be the set of labels of neighbors of a tie node t, $t \in T$. Let $s_{t,l}$ denotes the total strength of links incident to t whose end-points are labeled by l, i.e.

$$s_{t,l} = \sum_{x \in N} Weight(t, x), \;\; (t, x) \in L \wedge label.x = l.$$

Then t adopts label l, $l \in LB$, if

$$(\forall l' \in LB, l' \neq l) \; s_{t,l} > s_{t,l'}$$

In other words, a tie node adopts the label of a community to which it has the strongest connection. In case that there are more than one community with such property then one of them is selected randomly. Also, it is important to observe that ties nodes are resolved in decreasing order of weighted degree (the total sum of weights of links incident to a node) since tie nodes can be connected among themselves. Unlabeled neighbors of tie nodes are added to C and the whole process of safe label propagation and resolution of tie nodes is repeated until all nodes become labeled.

Algorithm 2: Propagation of community labels

```
input  : G = (N, L) - undirected weighted network with initial
         assignment of community labels
         A – set of labeled nodes
output: G with full assignment of community labels

// C – the set of unlabeled nodes whose neighbors
// are labeled
C := empty set of nodes
foreach n ∈ A do
    C := C ∪ G.getUnlabeledNeighbours(n)
end
while C ≠ ∅ do
    // T – the set of tie nodes
    T := ∅

    // safe propagation of labels
    while C ≠ ∅ do
        // D – the set of nodes labeled in the current iteration
        D := ∅
        foreach n ∈ C do
            if labeled neigbourhs of n have the same label l then
                n.label = l
                D := D ∪ n
            end
            else
                T := T ∪ n
            end
        end
        C = ∅
        foreach n ∈ D do
            UN := G.getUnlabeledNeighbours(n)
            foreach x ∈ UN do
                if x ∉ T then
                    C := C ∪ {x}
                end
            end
        end
    end

    // resolution of tie nodes
    sort T according to weighted degree
    foreach t ∈ T do
        NS := G.getLabeledNeighbours(t);
        w := empty array of weights;
        foreach n ∈ NS do
            w[n.label] := w[n.label] + Weight(t, n)
        end
        if w has an unique maximal element then
            t.label := i, where w[i] is the maximal element of w
        end
        else
            t.label := randomly selected index of maximal
              elements in w
        end
    end
    foreach t ∈ T do
        C := C ∪ G.getUnlabeledNeighbours(t)
    end
end
```

The community detection algorithm based on w-cores is actually a graph traversal algorithm where each node is visited exactly once to set its label. The neighborhood of each labeled node is explored either two times (for non-tie nodes) or three times (for tie nodes). Therefore, the time complexity of the algorithm is $O(n + l)$, where n and l are the number of nodes and links in the network, respectively. The algorithm also includes sorting of tie nodes. However, the number of tie nodes per one iteration in Algorithm 2 is significantly smaller than the total number of nodes, and consequently this operation does not affect the asymptotic time complexity of the algorithm. The algorithm has exactly one parameter – weight threshold that is used to determine w-cores. Therefore, a systematic application of the algorithm requires $O(d(n + l))$ time where d is the number of different link weights in the network. In the worst case, when all links have different weights, we have that $d = l$. In more realistic cases, when the network is sparse and the distribution of link weights is heavy-tailed, we have that $l \sim n$ and $d \sim \log l$ which results in $O(n \log n)$ time. Therefore, our method can be efficiently systematically applied to large-scale real-world networks.

4. EXPERIMENTAL EVALUATION

We conducted the experimental evaluation of our community detection method on a co-authorship network representing collaborations among researchers employed at our Department. This network was selected since we are familiar researchers contained in the network (all of the authors of this paper are also nodes in the network), and consequently in position to verify whether obtained communities are meaningful. The network was constructed from the data stored in the institutional CRIS (Current Research Information System) [10]. We used the normal weighting scheme for links, i.e. the weight of a link is the number of papers co-authored by researchers connected by the link.

The basic structural characteristics of the network are given in Table 1. In co-authorship networks, the degree of a node (researcher) is actually the number of co-authors, while weighted degree represents the number of multi-authored papers. The majority of links (51.54%) have weight that is less than 3 which implies that the researchers mostly publish one or two joint paper together. However, there are pairs of the researchers whose strength of collaboration is far from the average: there are two researchers in the network who have 220 papers authored together, while 14 pairs of researchers (7.21% of the total number) have more than 20 joint papers.

Table 1: Basic structural characteristic of the co-authorship network used in the experimental evaluation.

The number of nodes	81
The number of links	198
Average degree	4.889
Average weighted degree	38.074
Average link weight	7.789
Network diameter	8
Small-world coefficient	3.565
Clustering coefficient	0.517

We systematically applied the technique considering all possible values of link weights in order to identify the best partitioning that can be obtained. Figure 1 shows the change of the Girvan-Newman modularity measure (Q) for different weight thresholds (w) that are used to identify w-cores in the network. For $w > 60$ there is exactly one w-core. Consequently, only one community encompassing all nodes in the network will be detected for extremely high values of w. The maximal value of modularity, $Q_{\max} = 0.65834$, is attained when 11-cores constitute the base of communities. Usually a value of modularity higher than 0.3 is considered as a clear indication that the network possesses community organization according to the modularity based definition of community [7]. As it can be observed, the value of modularity is higher than 0.6 for a wide range of link weights which implies that the network consists of highly cohesive groups of researchers.

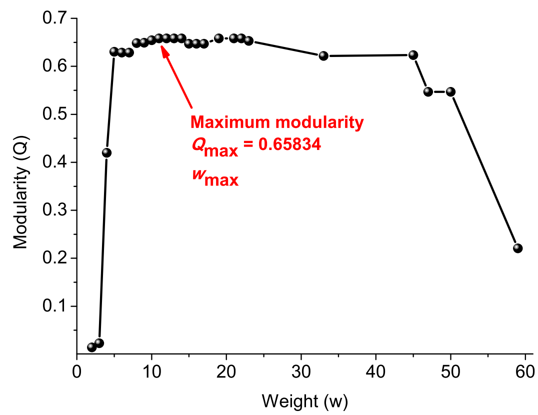

Figure 1: The value of modularity measure for different weight thresholds.

Figure 2 shows the visualization of the best partitioning according to the Newman-Girvan modularity measure. In total 7 communities are identified (labeled from C1 to C7 in Figure 2 and Table 2). The characteristics of detected communities are summarized in Table 2. Namely, for each community C we measured the following quantities:

- S - the size of C, i.e. the number of nodes that are contained in C.
- W^{in} - the total sum of weights of intra-community links incident to nodes in C,
- W^{out} - the total sum of weights of inter-community links incident to nodes in C,
- Con - conductance which is defined as $W^{out}/(W^{out} + W^{in})$. The value of conductance smaller than 0.5 implies that C is a Radicchi weak community.
- $max\left(\frac{w^{out}}{w^{in}}\right)$ - the maximal ratio between w^{out} and w^{in} for all nodes in C. w^{out} and w^{in} denote the total weight of inter-community and intra-community links, respectively, incident to a node in C. The value of $max\left(\frac{w^{out}}{w^{in}}\right)$ smaller than 1.0 implies that C is a Radicchi strong community.

As it can be observed from data presented in Table 2 for each identified community we have that $W^{in} \gg W^{out}$, $Con < 0.5$ and $max\left(\frac{w^{out}}{w^{in}}\right) < 1.0$. This means that all identified communities are Radicchi strong – each member of each community established stronger collaboration with members of its own community than with researchers belonging to other communities.

Table 2: Characteristics of identified communities.

ID	S	W^{in}	W^{out}	Con	$max\left(\frac{w^{out}}{w^{in}}\right)$
C1	15	171	26	0.13	0.66
C2	15	104	15	0.13	0.33
C3	4	32	17	0.35	0.52
C4	13	204	32	0.14	0.63
C5	14	131	28	0.18	0.5
C6	16	738	30	0.04	0.11
C7	4	84	8	0.09	0.35

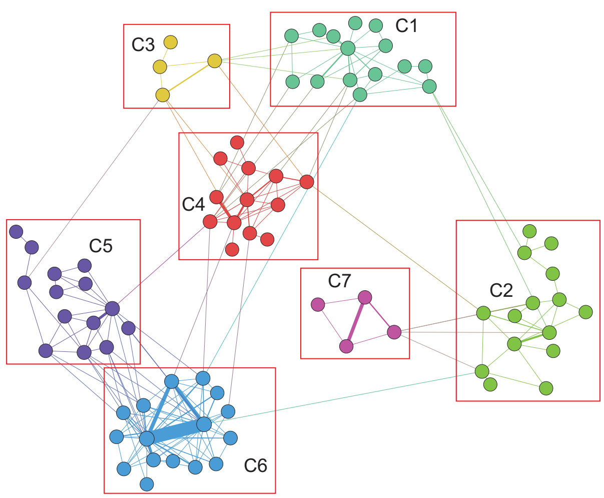

Figure 2: The visualization of the co-authorship network with the partitioning into communities for the maximum value of the modularity measure.

The largest community (C6) encompasses researchers that institutionally belong to the Chair of computer science whose research is focused on software engineering, data mining, agent-based computing and e-learning. Community C5 encompasses members the Chair of information technology and systems which is the another chair dealing with computing at our Department. The members of the Laboratory for the development of the information systems are also part of C5. The most central community in the network is C4. This community contains researchers institutionally organized into the Chair of numerical mathematics. Community C2 contains researchers who institutionally belong to two chairs with closely related research directions: the Chair of mathematical logic and discrete mathematics and the Chair of algebra and theoretical computer science. The similar situation can be observed for community C1 which aggregates researchers organized into the Chair of analysis, probability and differential equations and researchers from the Chair of functional analysis, geometry and topology. The two smallest communities consists of researchers dealing with applied mathematics: C3 corresponds to the Chair of applied analysis, while C7 corresponds to the Chair of applied algebra.

We also compared our community detection method to seven other techniques whose implementation is provided by the igraph R package [5]: the Girvan-Newman algorithm [9] (shortened as GN), Greedy modularity maximization [14, 4] (GMM), Walktrap [17] (WT), Label propagation [19] (LP), Spectral modularity optimization [15] (SMO), Infomap [20] (IM), and the Louvain method [1] (L). The results of the comparative analysis are summarized in Table 3 where our method is denoted as WC (short form of w-cores). As it can be seen all methods, except GN, identified community structures with highly close values of the Girvan-Newman modularity. The application of GN revealed a community structure whose modularity is significantly lower compared to other methods. However, only GMM, L and our method identified community structures where all communities satisfy the Radicchi strong criterion. Therefore, we can conclude that our method performs better than five other widely used community detection methods. On the other hand, community structures obtained using GMM, L and our method are highly similar: the normalized mutual information between the community structure identified using our method and community structures obtained using GMO and L are 0.941 and 0.966, respectively.

Table 3: The results of the comparative analysis. Q denotes the modularity measure, NC is the number of identified communities, while RSC is the number of Radicchi strong communities.

Method	Q	NC	RSC
GN	0.547	3	3
GMM	0.659	7	7
WT	0.658	9	8
LP	0.623	14	12
SMO	0.644	9	6
IM	0.652	11	10
L	0.659	6	6
WC	0.658	7	7

5. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a novel and simple community detection method for co-authorship networks. The baseline of our method is that frequent collaborators constitute cores of research communities. The presence of frequent collaborators in research collaboration network is formalized through a more general notion of w-cores which makes the method applicable for any weighted undirected network.

The method was experimentally evaluated on a co-authorship network consisting of mathematicians and computer scientists employed at our Department. The analysis of obtained community structure showed that the method identified meaningful and Radicchi strong communities. Moreover, we showed that it performs better than 5 other widely used community detection methods and gives similar results as the Louvian and Greedy modularity maximization methods. In our future work we plan to evaluate the method on other co-authorship networks, as well as on weighted undirected networks from other domains.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of this work by the Serbian Ministry of Education, Science and Technological Development through project *Intelligent Techniques and Their Integration into Wide-Spectrum Decision Support*, no. OI174023.

7. REFERENCES

[1] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[2] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Trans. on Knowl. and Data Eng.*, 20(2):172–188, Feb. 2008.

[3] Z. Chen, M. Jia, B. Yang, and X. Li. Detecting overlapping community in complex network based on node similarity. *Computer Science and Information Systems*, 12(2):843–855, 2015.

[4] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70(6):066111, 2004.

[5] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[6] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.

[7] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[8] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Phys. Rev. E*, 70(5):056104, Nov 2004.

[9] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[10] D. Ivanović, G. Milosavljević, B. Milosavljević, and D. Surla. A cerif-compatible research management system based on the marc 21 format. *Program: Electronic library and information systems*, 44(3):229–251, 2010.

[11] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[12] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2):167–256, 2003.

[13] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5):056131, 2004.

[14] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, 69(6):066133, Jun 2004.

[15] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[16] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[17] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2):191–218, 2006.

[18] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, 2004.

[19] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

[20] M. Rosvall and C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 105, pages 1118–1123, 2007.

[21] M. Savić, M. Ivanović, M. Radovanović, Z. Ognjanović, A. Pejović, and T. Jakšić Krüger. The structure and evolution of scientific collaboration in Serbian mathematical journals. *Scientometrics*, 101(3):1805–1830, 2014.

[22] M. Savić, M. Ivanović, M. Radovanović, Z. Ognjanović, A. Pejović, and T. Jakšić Krüger. Exploratory analysis of communities in co-authorship networks: A case study. In *ICT Innovations 2014*, volume 311 of *Advances in Intelligent Systems and Computing*, pages 55–64. Springer International Publishing, 2015.