Social Network Analysis and Mining (2021) 11:91
https://doi.org/10.1007/s13278-021-00804-5

ORIGINAL ARTICLE

Evaluating the role of community detection in improving influence maximization heuristics

László Hajdu[1,2] · Miklós Krész[1,3] · András Bóta[4]

Received: 4 February 2021 / Revised: 22 August 2021 / Accepted: 15 September 2021

Abstract
Both community detection and influence maximization are well-researched fields of network science. Here, we investigate how several popular community detection algorithms can be used as part of a heuristic approach to influence maximization. The heuristic is based on the community value, a node-based metric defined on the outputs of overlapping community detection algorithms. This metric is used to select nodes as high influence candidates for expanding the set of influential nodes. Our aim in this paper is twofold. First, we evaluate the performance of eight frequently used overlapping community detection algorithms on this specific task to show how much improvement can be gained compared to the originally proposed method of Kempe et al. Second, selecting the community detection algorithm(s) with the best performance, we propose a variant of the influence maximization heuristic with significantly reduced runtime, at the cost of slightly reduced quality of the output. We use both artificial benchmarks and real-life networks to evaluate the performance of our approach.

Keywords Network science · Community detection · Influence maximization

1 Introduction

Networks provide a versatile modeling tool that can be applied in many situations. Networks were used to represent physical and virtual relationships between people (Borgatti et al. 2009; Serrat 2017), cities and geographical regions (Gardner et al. 2018; Bridgwater and Bóta 2021; Colizza et al. 2006), financial or technological connections between companies (Mantegna 1999; Bóta et al. 2015; Krész and Pluhár 2017), interactions between molecules (Bagler and Sinha 2007; Wuchty et al. 2003) and gene sequences (Balcan 2007; Diambra and Costa 2005), relationships between words (Gravino et al. 2012; Bóta and Kovács 2014) among many other examples (Costa 2011; Newman 2003). The field of network science identified several common characteristics of these networks and proposed a variety of research questions associated with them. In this paper, we aim to establish a connection between two of these: community detection and influence maximization.

A common property of complex networks is community structure. This concept is based on the observation, that the edge distribution of such networks are globally and locally inhomogeneous. Certain sets of nodes are densely connected, while the connection between the sets are sparse. This behavior corresponds to known phenomenon in other fields of science. For example in sociology, homophily denotes the observation, that people with similar interests or properties have a preference toward making connections with each other. Community detection aims to define and discover these communities, and a great variety of detection algorithms have been proposed so far (Fortunato 2010). The most important difference between existing approaches is whether they allow overlaps to exist between communities. The largest fraction of detection algorithms define

✉ András Bóta
andras.bota@ltu.se

László Hajdu
laszlo.hajdu@innorenew.eu

Miklós Krész
miklos.kresz@innorenew.eu

1 Innorenew CoE, Livade 6, SI-6310 Izola, Slovenia

2 Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška 8, SI-6000 Koper, Slovenia

3 Department of Applied Informatics, University of Szeged, Boldogasszony sgt 6, HU-6725 Szeged, Hungary

4 Department of Computer Science, Electrical and Space Engineering, Embedded Intelligent Systems Lab, Luleå University of Technology, SE-97187 Lulea, Sweden

Published online: 03 October 2021

communities as disjoint sets of nodes, while a smaller field allows overlaps. Since the heuristic examined in this work uses a property of overlapping community detection, we will only consider algorithms in the latter category. A more detailed description of the topic and the used detection methods can be found in Sect. 3.

The problem of influence maximization is closely related to the spreading of information on networks. The most common network-based models describing the diffusion of information among nodes are in the families of compartmental models (Pastor-Satorras et al. 2015), including the popular Independent Cascade model (Kempe et al. 2003), and in the family of threshold models (Granovetter 1978). Both these model families represent spreading processes in an iterative way, starting from an initially active set of nodes. Most models in this family are constructed to terminate in a finite number of iterations, activating a fraction of the nodes. The task of influence maximization is to find the set of initially infected nodes activating the largest fraction of inactive nodes, when the size of the initial set is limited. The problem was originally proposed by Kempe et al. in (2003) and was shown to be NP-Hard. In the same paper, the authors proposed a greedy heuristic providing a good guaranteed solution to the problem. Since then, a great number heuristics and approximations have been proposed to tackle this problem (Kempe et al. 2005; Chen et al. 2010; Jung et al. 2012; Liu et al. 2014; Srivastava et al. 2015; Kingi et al. 2020; Tang et al. 2018; Hajdu et al. 2018). The survey paper of Li et al. gives an excellent summary of the field (Li et al. 2018). In Hajdu et al. (2018), initial results for an approach linking the two fields (community detection and influence maximization) together were proposed. According to the taxonomy of Li et al. (2018), the algorithm of Hajdu et al. falls into the category of influence ranking proxy algorithms. This heuristic is based on the output of overlapping community detection methods. However, in Hajdu et al. (2018), the authors only investigate two less frequently used community detection methods and they only use a limited number of test graphs.

As our first contribution in this paper, we select eight popular, frequently used overlapping community detection algorithms in conjunction with the approach of Hajdu et al. to investigate whether one or multiple methods outperforms the original greedy heuristic of Kempe et al. (2003), and three centrality measures frequently used in the literature: degree and betweenness centrality and PageRank. It is important to emphasize that it is not our intention to provide a general comparison between the community detection methods, we only investigate how well they can be used for this specific task. We give a short description of the selected methods in Sect. 3. As our second contribution, we propose a simplified variant of the algorithm of Hajdu et al., reducing its computational complexity considerably, with a slight loss in the quality of the results.

In order to thoroughly investigate the performance of our approach, we generate 1080 benchmark graphs with different characteristics with the graph generator of Lancichinetti and Fortunato (2009a). Furthermore, we select three real-life networks frequently used in the influence maximization literature to serve as an additional benchmark. Our results show that the algorithm of Hajdu et al. in conjunction with the best overlapping detection method outperforms the greedy algorithm of Kempe et al. and the centrality measures on real-world networks.

2 Background

We begin by introducing the concepts forming the background of our work. We define the influence maximization problem and describe the most fundamental algorithm related to the problem: the algorithm of Kempe et al. (2003).

2.1 Influence maximization

Influence maximization is an optimization problem, where the objective is to maximize the number nodes activated (or infected) by a diffusion (or infection) model on the network, choosing k initially infected nodes. The problem is usually defined in the terms of the Independent Cascade Model (Kempe et al. 2003).

In Independent Cascade Model, let $G(V, E)$ be an undirected network where for all $(u, v) \in E$, there is a $0 < p(u, v) \leq 1$ probability. The nodes in the network can be in a susceptible, infected (activated) or in a removed state, corresponding to the states of the SIR compartmental model with an infectious period of one iteration. The spreading process starts from a set of initially infected nodes $A_0 \in V(G)$ and takes place in an iterative way in discrete steps. Let the set of active nodes in iteration t be denoted as A_t. In each iteration, infected nodes $v \in A_t$ try to infect their susceptible neighbors $u \in V(G) / \bigcup_{i=0\ldots t} A_i$ based on the edge probabilities. If the attempt is successful, u joins the set of infected nodes A_{t+1} in the following iteration. If more than one node is trying to infect u in the same iteration, the attempts are made independently of each other in an arbitrary order within the same iteration. The process terminates naturally at iteration t' when $A'_t = \emptyset$.

The influence maximization problem can be defined in the following way. As before, let $A_0 \in V(G)$ be the set of the initially infected nodes, let k be the cardinality of A_0 and let $\sigma(A_0)$ be the corresponding expected size of $\bigcup_{i=0\ldots t'} A_i$, containing all nodes infected during the process. The objective is to maximize the number of activated nodes on the network, when choosing k initial infectors. The original

problem was introduced by Kempe et al. (2003) where they also have proven the NP-hardness of the problem and introduced a greedy optimization method which provides at least 63% of the optimum. In the next subsection, we introduce the greedy algorithm. A great variety of heuristics have been proposed for influence maximization. The survey paper by Li and Fan gives a good overview of these (Li et al. 2018). This paper focuses on one such heuristic, the algorithm of Hajdu et al., which will be defined in the next section.

2.2 The greedy algorithm of Kempe et al.

The greedy algorithm of Kempe et al. (2003) is still widely used as a general algorithm for influence maximization. The method starts with an empty A_0 and iteratively increases the size of the set until it reaches k, maximizing $\sigma(A_0)$ with greedy decisions. Algorithm 1 shows the pseudocode of the greedy algorithm.

Algorithm 1 Greedy method

1: **Input:** Graph $G(V, E)$, k: desired size of the A_0
2: **Output:** A_0
3: $A_0 \leftarrow \emptyset$
4: **While** $|A_0| \leq k$
5: $\quad A_0 = A_0 \cup \ arg\ max_{v \in V(G) \setminus A_0)} \sigma(A_0 \cup \{v\})$

The method starts with an empty set of initially infected nodes, and in the first iteration selects the node from V that has the maximal expected value of infection ($\sigma(A_0)$). Let say the greedy algorithm has chosen the node v_1 so $k = 1$. In the second iteration, the method searches the second node from $v_2 \in V(G) \setminus A_0$ set. It follows the same logic iteratively until the size of A_0 reaches the k. It can be seen that in the first iteration it computes the value of σ $|V(G)|$ times while in the second iteration $|V(G)| - 1$ times. It can be seen that the algorithm itself is very computational intensive since it has to check the whole search space in every iteration. At the same time, the greedy method is one of the best and widely used algorithms for influence maximization. In their original paper, the authors have proven that it provides a solution with a guaranteed precision of 63% of the optimum. The following subsection introduces our methodology to reduce the search space of the algorithm with the help of community detection, improving the efficiency of the method and at the same time giving a benchmarking system for community detection algorithms.

3 Methods

In this section, we are going to give a short summary of each of the algorithms playing a part in our work as well a description of the benchmark networks we used. We will begin by introducing the algorithm of Hajdu et al. as it appeared in Hajdu et al. (2018), as well as the newly proposed simplified algorithm. We will then describe how we selected the overlapping community detection methods included in our analysis and give a brief summary of them. Finally, we will give a short description of the benchmark networks used in our paper.

3.1 The algorithm of Hajdu et al.

The idea of this heuristic is based on an assumption that the nodes can be ordered based on their position in the network. Using only the highly ranked nodes in the greedy algorithm can significantly reduce the search space and the running time. Let $V^*(G) \subset V(G)$ be a reduced node set, where the greedy algorithm chooses from $V^*(G) \setminus A_0$ in every iterations. Let $f(v) : v \rightarrow Z$ be a function that assigns an integer number to every node. The nodes are ordered based on their $f(v)$ value, and the nodes with the highest $f(v)$ scores can be included in the set $V^*(G)$.

The communities are dense, strongly connected subgraphs where the nodes have stronger connection with each other than with the other parts of the network. If the subgraphs are dense enough, infection or influence can spread between the nodes more easily. Nodes connecting different communities in multiple dense subgraphs play a critical role that can be used for influence maximization. Let $f_c(v) : v \rightarrow Z$ be a function that assigns to each node the number communities it belongs to. We denote this value as *community value*. Before we run the greedy algorithm, we order nodes of the network based on their f_c value. We select the top $X\%$ nodes from this ordering to become reduced $V^*(G)$ selection set, which will replace (and reduce) the search space of the original greedy algorithm, where X is an adjustable parameter of the method. Algorithm 2 shows the pseudocode of the Hajdu et al. algorithm.

Algorithm 2 Algorithm of Hajdu et al.

1: **Input:** Graph $G(V, E)$, $k = |A_0|$, $X = |V^*(G)|$, M overlapping community detection method
2: **Output:** A_0
3: Detect communities on $G(V, E)$ using M
4: Ordering the nodes in set V based on their $f_c(v)$ values
5: $V^*(G) \leftarrow$ top X nodes from the ordered list
6: $A_0 \leftarrow \emptyset$
7: **While** $|A_0| \leq k$
8: $A_0 = A_0 \cup \ arg\ max_{v \in V^*(G) \setminus A_0)} \sigma(A_0 \cup \{v\})$

The algorithm takes the $G(V, E)$ network, the desired size of the A_0 , the desired size of the reduced selection set $V^*(G)$ and an overlapping community detection method M as parameters. In step 3-5, the reduced $V^*(G)$ set is created based on X and the given community detection method. Step 6-8 is the optimization part where the greedy method chooses the next node from $V^*(G)$ in each iteration. The original community value was defined for weighted and unweighted situations, but here we only consider unweighted networks.

3.2 The simplified algorithm of Hajdu et al.

While reducing the size of the selection set reduces the runtime of the greedy algorithm, it still needs to compute the optimization steps 7-8 of Algorithm 2. In this paper, we show that the optimization steps can be completely omitted, while still achieving better performance than the original greedy method. Algorithm 3 shows the simplified heuristic. Instead of setting the size of a reduced selection set to a fraction of the nodes of the network, we simply select the top k highest ranking nodes according to f_c and return them as the output A_0.

Algorithm 3 Simplified algorithm of Hajdu et al.

1: **Input:** Graph $G(V, E)$, $k = |A_0|$, M overlapping community detection method
2: **Output:** A_0
3: Detect communities on $G(V, E)$ using M
4: Ordering the nodes in set V based on their $f_c(v)$ values
5: $V^*(G) \leftarrow$ top k nodes from the ordered list
6: $A_0 \leftarrow V^*(G)$

3.3 Selected overlapping community detection methods

The community value in the algorithm of Hajdu et el. is based on an output of an arbitrary overlapping community detection algorithm. In order to identify which detection algorithm matches the problem of influence maximization and the algorithm of Hajdu et al. best we selected eight methods among the most popular approaches. An important part of the selection criteria was that the method had to have a publicly available implementation.

- One of first overlapping community detection methods proposed was the Clique Percolation Method (CPM) of Palla et al. (2005). It is based on a percolation process defined between k-cliques. It is still frequently used and has a publicly available implementation online (CFinder http://www.cfinder.org/).
- The COPRA algorithm (Gregory 2010) proposed by S. Gregory, is a label propagation method for overlapping community detection based on the non-overlapping algorithm of Raghavan, Raghavan et al. (2007). An implementation of the method can be downloaded from COPRA https://gregory.org/research/networks/software/copra.html.
- The Greedy Clique Expansion method (Lee et al. 2010) uses the maximal cliques of the input network and expands them by greedily maximizing a local fitness function based on the function defined by Lancichinetti and Fortunato (2009b). The algorithm is available for download at GCE (https://sites.google.com/site/greedycliqueexpansion/)[1].
- The map equation forms the core of the well-known Infomap community detection method (Rosvall and Bergstrom 2008). It models the probability flow of random walks on a network as an information flow and seeks to compress the description of this flow. The method was extended for overlapping communities (Esquivel and Rosvall 2011) and is available for researchers on InfoMap (http://www.mapequation.org).
- The MOSES algorithm (McDaid and Hurley 2010) was proposed to detect highly overlapping community

[1] Unfortunately the method it is no longer available on this website.

structure. The algorithm uses a variant of Overlapping Stochastic Block Modeling to define a global objective function and employs a greedy maximization strategy to assign nodes to communities. It can be downloaded from MOSES (https://sites.google.com/site/aaronmcdaid/downloads).

- Another frequently used detection algorithm is OSLOM (Lancichinetti et al. 2011). The method provides great flexibility, it is able to detect directed, weighted, overlapping and hierarchical communities, and can also be used for the refinement of existing community structures. Like many other approaches, it is based on the local optimization of a unique fitness function. OSLOM can be downloaded from OSLOM http://www.oslom.org/.
- Stochastic block modeling, or more specifically, inferring the underlying stochastic block model behind the communities of an existing network is a popular and well-studied approach (Doreian et al. 2005; Peixoto 2015). Stochastic block models can be defined for overlapping community structure too. In our work, we have used the software framework available from InfoMap (http://www.mapequation.org).
- The SLPA method (Xie et al. 2011) is a label propagation approach closely related to COPRA. It employs a speaker–listener information propagation process, allowing nodes to switch between the roles of speaker and listener during the stochastic detection process. During this process, the nodes accumulate knowledge of repeatedly observed labels. The algorithm is available online from SLPA (https://github.com/sebastianliu/SLPA-community-detection).

3.4 Centralities

We employ the following centrality metrics to provide a benchmark for our approach beside the original greedy algorithm. In order to provide a fair comparison we will test the performance of the centrality metrics both in conjunction with the Hajdu algorithm, and independently, by selecting the top k nodes according to the centrality-based rankings, and computing $\sigma(A_0)$.

- Degree Centrality (Freeman 1979) of a node in a network is defined as the number of the edges that are incident upon the node so in an undirected network it is the degree of the actual node.
- Betweenness Centrality (Freeman 1977) can be defined both on nodes and edges. If we define the shortest path between each node pairs in the network, the betweeness centrality is the number of the cases when a shortest path passes through the node.
- PageRank estimates the importance of a node by counting the number and quality of the links that are incident upon the node (Brin and Page 1998). The original method is used by the Google to measure the importance of the website pages.

3.5 Test networks

We will use both artificial and real-life benchmarks to test our approach.

3.5.1 Artificial benchmark networks

Benchmark networks were created using the C implementation of the graph generator proposed by Andrea Lancichinetti and Santo Fortunato in Lancichinetti and Fortunato (2009a). The following parameters were used during the graph generation:

- N: 1000 (number of nodes)
- d: 25 (average degree)
- d_{max}: 33 (maximum degree)
- μ: $0.1, 0.2, \ldots, 0.6$ (mixing parameter)
- t_1: -2 (minus exponent for the degree sequence)
- t_2: 1.5 (minus exponent for the community size distribution)
- c_{min}: 10 (minimum for the community sizes)
- c_{max}: 50 (maximum for the community sizes)
- o_n: $0.1, 0.2, \ldots, 0.6$ (number of overlapping nodes)
- o_m: 2, 3, 4 (number of memberships of the overlapping nodes)

The parameters were chosen so the test networks contain a large variety of different community structures. Six different values were chosen for parameter o_n governing the fraction of overlapping nodes and three values for o_m setting the number of communities a node may belong to. Additionally, six different values were chosen for the mixing parameter μ influencing the amount of connections inside and between the communities. Since the graph generator is stochastic, 10 test networks where generated with each parameter configuration resulting in $6 * 3 * 6 * 10 = 1080$ test graphs.

3.5.2 Real-life networks

We use three real-life networks from the Stanford Large Network Dataset Collection (Leskovec and Krevl 2014) to provide additional benchmarks.

- CA-CondMat network (Leskovec et al. 2007) is a Condense Matter collaboration network that is containing scientific collaboration information between authors based on their common papers.
- CA-HepPH (Leskovec et al. 2007) is a High Energy Physics - Phenomenology collaboration network which

is also containing collaboration information between researchers.
– The Com-Dblp (Yang and Leskovec 2012) is a DBLP collaboration network contains author connections based on computer science bibliography information.

4 Results

We conduct the evaluation of the heuristics of Hajdu et al. in the following framework. We start by applying all community detection algorithms in Sect. 3.3 to both the real-life and artificial test networks described in Sect. 3.5. Then, we assign community values to the nodes of the test networks for all detection methods. Let $f_c^a(v_j)$ denote the community value of node $v_j \in V(G_j)$, where G_j is the j-th test network and a marks the algorithm. We rank the nodes according their community values for each algorithm and each test graph. We also define the set of influence maximization tasks, one for each test network G_j, by setting $k = 50$ that is we are looking for the 50 most influential nodes.

We run the original algorithm of Hajdu with the size of the reduced selection set to 20%, 10% and the simplified version on all test graphs, each time using community values obtained with one of the eight community detection algorithms. Due to performance issues with the original greedy algorithm of Kempe et al., and the number of test networks, we use different strategies for comparing our approach with benchmark methods. For the artificial networks we compare results with the original greedy algorithm and the three centrality metrics applied independently from the Hajdu algorithm by selecting the top k nodes and computing $\sigma(A_0)$. For the real-life networks, we only use the three centrality metrics as benchmarks, but we both use them in conjunction with the Hajdu algorithm by selecting the top 20% and 10% highest ranking nodes to create the reduced set, and independently as with the artificial networks.

4.1 Results on artificial networks

According to our results, all heuristics and centrality metrics provide similar performance on the artificially generated networks. Considering only the community-based heuristics, on test networks with moderately or highly overlapping community structure, the community values provided by the overlapping Infomap algorithm provide the best results, except when $\mu = 0.1$ and 0.2 with a low amount of overlaps. In these situations CPM, SLPA or SBM occasionally gives the best solution. Comparing the different variants of the Hajdu algorithm, we experienced a drop of performance for the 10% and the simplified variant. However, the decrease depends highly on the detection method used to compute the community values. The Infomap method only has a slight drop, further highlighting the robustness of the method. Furthermore, Infomap is a fast algorithm, so the performance overhead introduced by running it on the test networks to obtain the community values is small. Figure 1 shows $\sigma(A_0)$ for all algorithms with varying values for o_m and o_n, and $\mu = 0.3$. Values for test graphs with the same parameters (see 3.5) were averaged.

Overall however, the PageRank-based ranking clearly provides the best performance, giving the best solution on 840 test graphs when compared to the simplified algorithm, 765 graphs when compared with the Hajdu algorithm with a reduced set of 10% and 489 times with a reduced set of 20%. Simple degree centrality provides the best solution on a smaller fraction of the test networks, while the original greedy algorithm occasionally gives better results. It should be noted however that the Infomap-based community heuristic performs very close to these methods. Table 1 shows the number of test networks, where each heuristic and centrality metric has the best performance.

4.2 Results on real-life networks

Our results on real-life networks paint a completely different picture. While again all approaches perform close to each other, the community-based heuristics clearly provide the best performance, while centrality metrics fall into the second half of the methods, even when used in conjunction with the Hajdu algorithm to compute the reduced set. Figure 2 shows the results of the overlapping community detection methods as well as the centralities on the three real networks.

The SLPA algorithm provides the best community values for almost all test graphs and algorithm variants, and both COPRA and GCE performs consistently above the rest of the methods. In contrast with our observations on the artificial networks, the overlapping Infomap algorithm produces poor results here, along with MOSES and OSLOM. The behavior of CPM and SBM inference changes depending on the test graph. The former competes with SLPA in providing the best results for the high energy physics collaboration network, while the latter has good but not outstanding results on the DBLP collaboration network.

The different variants of the Hajdu algorithm also show a mixed pattern. On most of the test graphs and with most of the detection algorithms, the 20%, 10% and simplified variants have nearly identical performance. In a smaller number of scenarios, we see the same pattern as with the artificial benchmarks: the 20% variant performs best, while the 10% and simplified algorithms have decrease performance.

Overall, we can conclude that the community-based heuristics perform well on both artificial and real-life benchmark networks. On artificial networks, community values provided by the overlapping Infomap method provide the

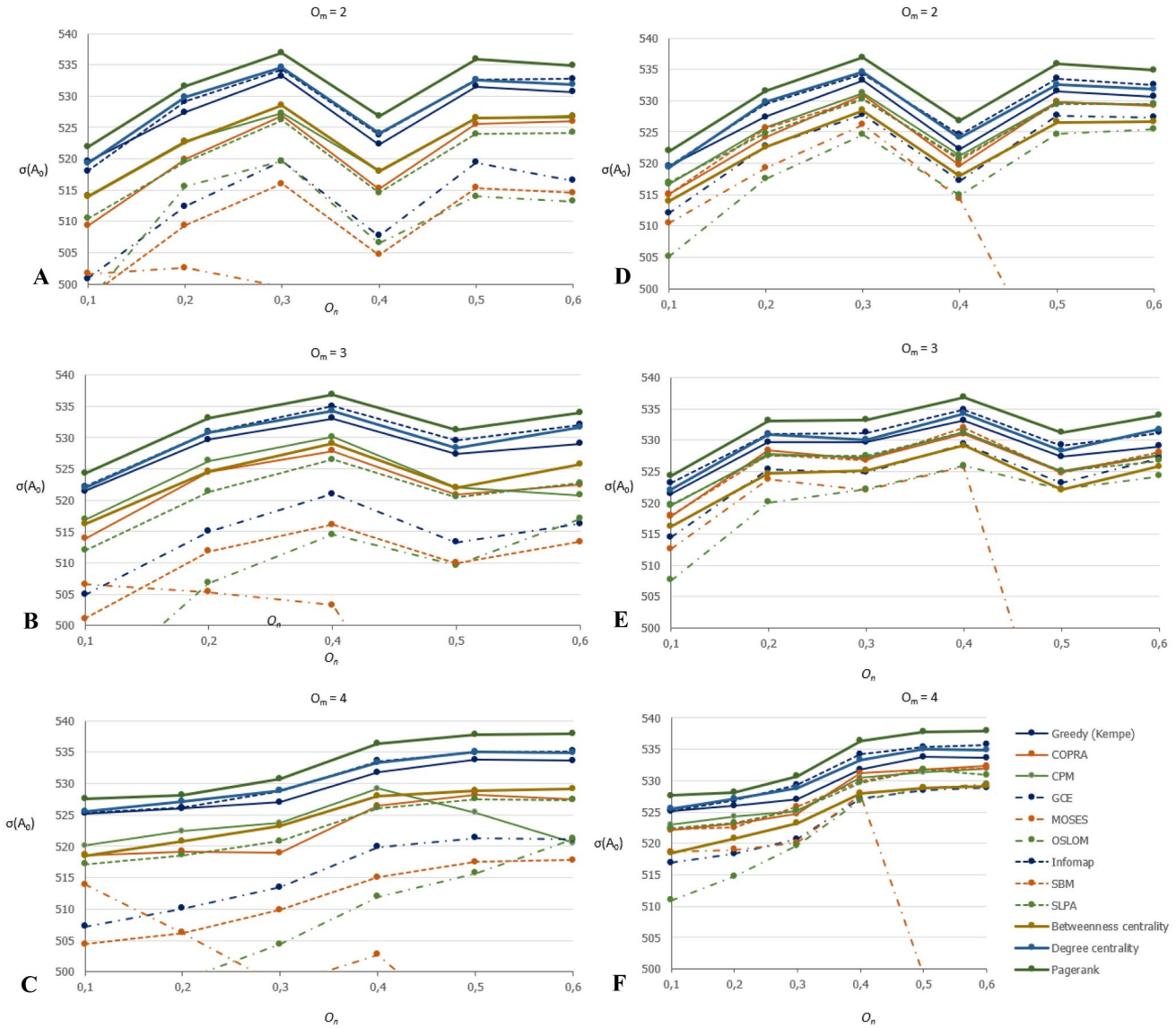

Fig. 1 Performance of the community detection algorithms in conjunction with the heuristic. Left side: simplified heuristic. Right side: original heuristic with the selection set reduced to 20%. $\sigma(A_0)$ is shown for all algorithms with varying values for o_m and o_n, and $\mu = 0.3$. The performance of the greedy algorithm is also shown for comparison

best results and reach $\sigma(A_0)$ values close to those provided by the PageRank algorithm. On real-life networks, community values provided by the SLPA detection algorithm outperform all other approaches.

5 Discussion

While the community-based heuristics proposed by this paper have good performance, there is a noticeable difference between them depending on what kind of community values are used and what kind of benchmark network are they used on.

On the artificial benchmark networks, our heuristic performs better on test graphs with a moderately or heavily overlapping community structure, indicated by greater values of parameters o_m and o_n. Since our algorithm is based on the overlaps between communities, this behavior is not surprising. The mixing parameter μ also has an effect on the quality of the results, with greater values improving performance. The community values provided by Infomap give the best results here, and the method is also robust, as it depends only minimally on the size of the reduced selection set of the heuristic algorithm. Infomap performs close to the PageRank centrality metric on these networks (and on the real-life benchmarks too). A likely explanation for this

Table 1 Number of test graphs (out of 1080) where a community detection method provided the best $\sigma(A_0)$ in conjunction with our heuristic for all overlapping community detection methods. Three variants of the algorithm are shown: the unmodified heuristic with the selection set reduced to 20%, to 10% and the simplified heuristic proposed in Sect. 3.2. We also show the number of times the greedy algorithm of Kempe et al. provided the best result

	20 % selection set	10 % selection set	Simplified heuristic
Greedy (Kempe)	35	28	97
Community heuristics	293	87	66
Betweenness centrality	6	0	0
Degree centrality	257	200	77
Pagerank	489	765	840

is, that both algorithms are based on the idea of random walks. Why these approaches perform so well on the Lancichinetti–Fortunato–Radicchi benchmarks is more difficult to explain. One possible factor is that these networks are randomly generated and even with the same parameter settings the resulting test graphs have slightly different structure. We averaged the $\sigma(A_0)$ values for test graphs generated with the same parameters when comparing the performance of each approach, implicitly providing an advantage to probabilistic methods, like those based on random walks (Infomap and PageRank).

While all of our real-life test networks are collaboration networks, their individual structure might be different, depending on the publication habits of scientific fields and their source. On these networks, our community based heuristics provide much better performance than the centrality metrics. The community detection methods consistently providing the most useful community values are based on label propagation, the SLPA and COPRA methods. A likely explanation is the similarity between the diffusion processes underlying the influence maximization problem and the independent cascade model, and the mechanics of label propagation. Collaboration networks in scientific fields, where most papers have three or more authors, are based on cliques. This explains why clique-based methods provide good performance on these networks, although it should be noted that the GCE algorithm outperforms the less flexible CPM algorithm on two test graphs, while the two provide near identical results on the third.

In order to better understand the relationship between results provided by the individual detection algorithms, we ranked the nodes of the three real-life networks according to the community values assigned by all detection algorithms and all three centrality metrics, and we computed the Kendall τ rank correlation coefficient between all pairs of rankings. The results show an interesting pattern. The centrality metrics show moderate correlation, which was to be expected on small-world networks. In general, rank correlation between

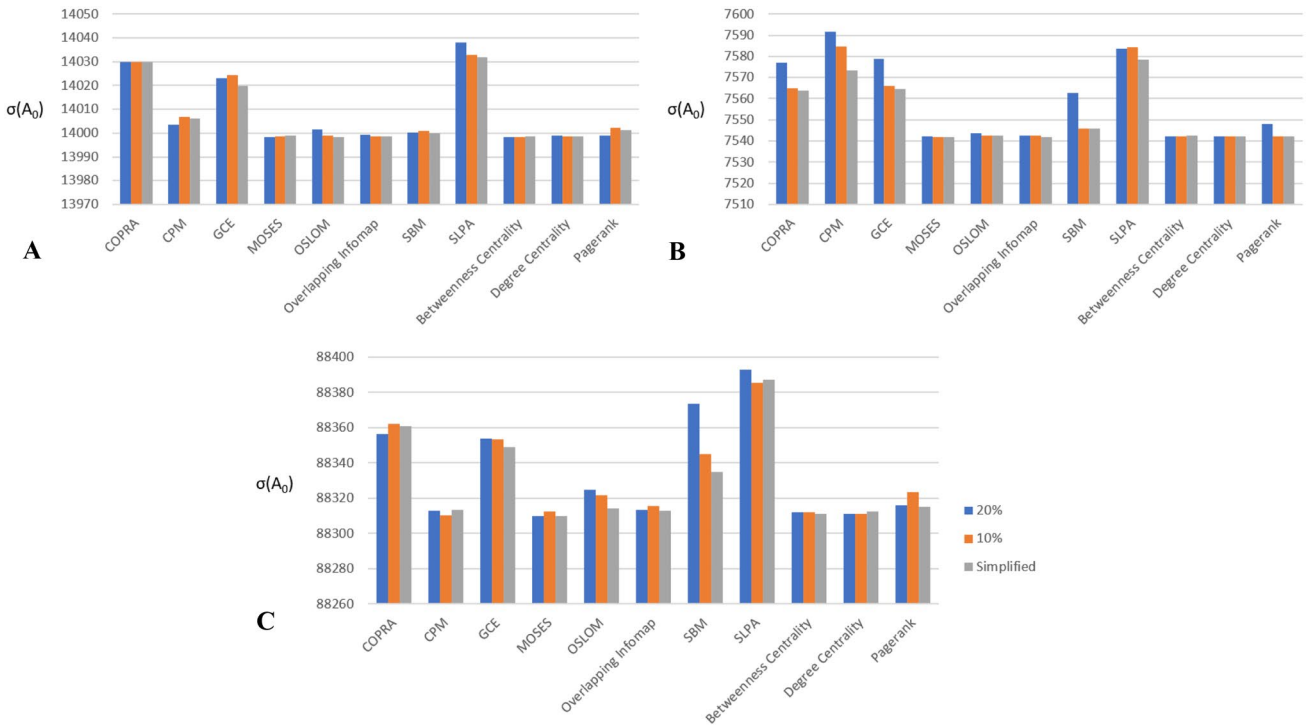

Fig. 2 Performance of all algorithms on real world networks. A: CA-CondMat (Leskovec et al. 2007) B: CA-HepPH (Leskovec et al. 2007) C: com-dblp (Yang and Leskovec 2012)

Table 2 Kendall τ (tau-b) rank correlation between community value and centrality based node rankings for all community detection methods and centralities. A:COPRA. B:CPM. C:GCE. D:MOSES. E:OSLOM. F:INFOMAP. G:SBM. H:SLPA. I:Betweenness. J:Degree. K:Pagerank

	A	B	C	D	E	F	G	H	I	J	K
A	1.00	0.04	−0.01	0.02	0.06	0.00	0.15	0.02	0.02	0.06	0.02
B	0.04	1.00	0.14	0.70	0.21	0.32	0.07	0.00	0.47	0.48	0.39
C	−0.01	0.14	1.00	0.10	0.07	0.14	0.03	0.13	0.09	−0.08	−0.18
D	0.02	0.70	0.10	1.00	0.36	0.49	0.12	−0.06	0.67	0.64	0.55
E	0.06	0.21	0.07	0.36	1.00	0.32	0.19	−0.04	0.34	0.28	0.28
F	0.00	0.32	0.14	0.49	0.32	1.00	0.10	−0.03	0.49	0.32	0.27
G	0.15	0.07	0.03	0.12	0.19	0.10	1.00	−0.01	0.13	0.10	0.11
H	0.02	0.00	0.13	−0.06	−0.04	−0.03	−0.01	1.00	−0.06	−0.06	−0.16
I	0.02	0.47	0.09	0.67	0.34	0.49	0.13	−0.06	1.00	0.51	0.54
J	0.06	0.48	−0.08	0.64	0.28	0.32	0.10	−0.06	0.51	1.00	0.63
K	0.02	0.39	−0.18	0.55	0.28	0.27	0.11	−0.16	0.54	0.63	1.00

Table 3 Overlap between the final 50 selected nodes for all pairs of community detection algorithms and centralities for the Com-Dblp network. A:COPRA. B:CPM. C:GCE. D:MOSES. E:OSLOM. F:INFOMAP. G:SBM. H:SLPA. I:Betweenness. J:Degree. K:Pagerank

	A	B	C	D	E	F	G	H	I	J	K
A	50	0	0	0	0	0	0	0	0	0	0
B	0	50	0	9	1	2	0	0	6	7	5
C	0	0	50	0	0	0	0	0	0	0	0
D	0	9	0	50	4	1	0	0	12	13	13
E	0	1	0	4	50	1	1	0	3	7	3
F	0	2	0	1	1	50	0	0	1	1	0
G	0	0	0	0	1	0	50	0	0	1	1
H	0	0	0	0	0	0	0	50	0	0	0
I	0	6	0	12	3	1	0	0	50	8	10
J	0	7	0	13	7	1	1	0	8	50	12
K	0	5	0	13	3	0	1	0	10	12	50

community detection methods is very low, underlining the uniqueness of each approach. The exceptions are CPM, MOSES and Infomap. CPM and MOSES have a moderate-high correlation value of 0.7, which is surprising since CPM is clique-based and MOSES is related to stochastic block modeling, two different concepts. Infomap also shows moderate correlation with MOSES (0.49), despite the differences between these methods. These methods also show moderate-high correlations with the centralities, especially MOSES and CPM, while Infomap shows moderate correlation (0.49) with betweenness centrality. Table 2 shows the pairwise rank correlations between all detection algorithms and centralities for the CA-HepPH network.

Furthermore, to better understand how community values contribute to finding the solution with the best $\sigma(A_0)$, we computed the overlap between the final 50 selected nodes for all pairs of community detection algorithms and centralities. According to our results, the amount of overlaps vary between the networks, but in general, there is very little similarity between nodes selected by the greedy algorithm using the community value based rankings and even the centralities. Furthermore, increasing the size of the reduced set decreases the sizes of the overlaps even more. Table 3 shows the number of overlaps between the final 50 selected nodes for all pairs of community detection algorithms and centralities for the Com-Dblp graph, with the reduced set size of 10%. We can see, that there is very little similarity between values selected from the community detection methods. Surprisingly, there is only a limited amount of overlap between the centrality metrics too, which otherwise correlate well on small-world networks. There is also only a limited amount of overlap between the centrality metrics and MOSES, CPM and Infomap, contradicting the pattern we have seen during the correlation analysis. This implies, that considering the applicability of the community value and centrality-based rankings in selecting the most influential nodes, there is great difference between the individual rankings. In this situation, selecting the right community detection method is critical.

6 Conclusions

In this paper, we examined the properties of a community-based heuristic for influence maximization, and proposed a simplified variation with a greatly reduced runtime. We

have used the output of eight well-known overlapping community detection methods in conjunction with the algorithm of Hajdu et al. on a large variety of test networks with different community structures. We compared the expected size of infected nodes obtainable with our heuristic with the original greedy method of Kempe et al. and three centrality metrics to serve as benchmarks. Our results show that on real-life networks the community values provided by the SLPA heuristic combined with the algorithm of Hajdu et al. provide better $\sigma(A_0)$ than all other methods and centralities, and thus the performance of our approach is comparable to state-of-the-art methods in the field. In our analysis of the results, we examined the relationship between the community values calculated by the different detection algorithms and the centrality metrics. We found that most of the detection methods provide community values dissimilar to each other and the centralities, and the "unique" detection methods perform better on real-life networks. Furthermore, we examined the overlap between the final 50 selected nodes for all pairs of community detection algorithms and centralities and found that the overlap is minimal. This implies significant differences between the community detection methods and metrics regarding their applicability to influence maximization, but also implies that influence maximization heuristics based on simply ranking nodes according to some metric can always be improved by combining them with the optimization step of the original greedy algorithm.

Acknowledgements László Hajdu and Miklós Krész gratefully acknowledge the European Commission for funding the InnoRenew CoE project (Grant Agreement 739574) under the Horizon2020 Widespread-Teaming program, and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund). They are also grateful for the support of the Slovenian Academy of Sciences and Arts (project title: 'Deployment and analysis of sensor networks in buildings'), and for the support of the Slovenian Research Agency (ARRS) through grants N1-0093, N2-0171 and J2-2504. The Authors would also like to acknowledge the work of Máté Vass, who was helping the research in testing and evaluation. He was supported by the 'Integrated program for training new generation of scientists in the fields of computer science', no EFOP-3.6.3- VEKOP-16-2017-0002 (supported by the European Union and co-funded by the European Social Fund).

Funding Open access funding provided by Lulea University of Technology.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bagler G, Sinha S (2007) Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics 23(14):1760–1767

Balcan D et al (2007) The information coded in the yeast response elements accounts for most of the topological properties of its transcriptional regulation network. PLoS One 2(6):e501

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Bridgwater A, Bóta A (2021) Identifying regions most likely to contribute to an epidemic outbreak in a human mobility network. In: Proceedings of the 2021 Swedish artificial intelligence society workshop (SAIS), pp. 1-4, IEEE

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine (PDF). Comput Netw ISDN Syst 30(1–7):107–117

Bóta A, Csernenszky A, Győrffy L, Kovács G, Krész M, Pluhár A (2015) Applications of the inverse infection problem on bank transaction networks. Cent Eur J Oper Res 23(2):345–356

Bóta A, Kovács L (2014) The community structure of word association graphs. In: Proceedings of the 9th international conference on applied informatics (Vol. 1, pp. pp-113). Eger, Hungary

CFinder, http://www.cfinder.org/

COPRA, https://gregory.org/research/networks/software/copra.html

Chen W, Wang C, Wang Y (2010) Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp 1029-1038)

Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proc Natl Acad Sci 103(7):2015–2020

Costa LDF et al (2011) Analyzing and modeling real-world phenomena with complex networks: a survey of applications. Adv Phys 60(3):329–412

Diambra L, Costa LDF (2005) Complex networks approach to gene expression driven phenotype imaging. Bioinformatics 21(20):3846–3851

Doreian P, Batagelj V, Ferligoj A (2005) Generalized blockmodeling, vol 25. Cambridge University Press, Cambridge

Esquivel AV, Rosvall M (2011) Compression of flow can reveal overlapping-module organization in networks. Phys Rev X 1(2):021025

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174

Freeman Linton (1977) A set of measures of centrality based upon betweenness. Sociometry 40(1):35–41

Freeman Linton C (1979) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

GCE, https://sites.google.com/site/greedycliqueexpansion/

Gardner LM, Bóta A, Gangavarapu K, Kraemer MU, Grubaugh ND (2018) Inferring the risk factors behind the geographical spread

and transmission of Zika in the Americas. PLoS Negl Trop Dis 12(1):e0006194
Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420–1443
Graph-tool https://graph-tool.skewed.de/
Gravino P, Servedio VD, Barrat A, Loreto V (2012) Complex structures and semantics in free word association. Adv Complex Syst 15(03n04):1250054
Gregory S (2010) Finding overlapping communities in networks by label propagation. New J Phys 12(10):103018
Hajdu L, Krész M, Bóta A (2018) Community based influence maximization in the Independent Cascade Model. In: 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 237-243) IEEE
InfoMap http://www.mapequation.org
Jung K, Heo W, Chen W (2012) Irie: Scalable and robust influence maximization in social networks. In: 2012 IEEE 12th international conference on data mining (pp 918-923)
Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence though a social network. *Proceedings of the 9th ACM SIGKDD international conference on knowledge discovery and data mining*, ACM (2003) 137–146
Kempe D, Kleinberg J, Tardos É (2005) Influential nodes in a diffusion model for social networks. In: international colloquium on automata, languages, and programming (pp 1127-1138). Springer, Berlin, Heidelberg
Kingi H, Wang LAD, Shafer T, Huynh M, Trinh M, Heuser A et al (2020) A numerical evaluation of the accuracy of influence maximization algorithms. Soc Netw Anal Min 10(1):1–10
Krész M, Pluhár A (2017) Economic network analysis based on infection models. In: Alhajj R, Rokne J (eds) Encyclopedia of social network analysis and mining. Springer, New York
Lancichinetti A, Fortunato S (2009) Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Phys Rev E 80(1):016118
Lancichinetti A, Fortunato S (2009) Community detection algorithms: a comparative analysis. Phys Rev E 80(5):056117
Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) Finding statistically significant communities in networks. PloS One 6(4):e18961
Lee C, Reid F, McDaid A, Hurley N (2010) Detecting highly overlapping community structure by greedy clique expansion. arXiv preprint arXiv:1002.1827
Leskovec J, Kleinberg J, Faloutsos C (2007) Graph Evolution: densification and Shrinking Diameters. ACM transactions on knowledge discovery from data (ACM TKDD) 1(1)
Leskovec J, Krevl A (2014) SNAP Datasets: stanford large network dataset collection, http://snap.stanford.edu/data
Li Y, Fan J, Wang Y, Tan KL (2018) Influence maximization on social graphs: a survey. IEEE Trans Knowl Data Eng 30(10):1852–1872
Liu Q, Xiang B, Chen E, Xiong H, Tang F, et al (2014) Influence maximization over large-scale social networks: a bounded linear approach. In: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, pp 171-180
MOSES https://sites.google.com/site/aaronmcdaid/downloads
Mantegna RN (1999) Hierarchical structure in financial markets. Eur Phys J B-Condens Matter Complex Syst 11(1):193–197
McDaid A, Hurley N (2010) Detecting highly overlapping communities with model-based overlapping seed expansion. In: 2010 international conference on advances in social networks analysis and mining (pp 112-119) IEEE
Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
OSLOM, http://www.oslom.org/
Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818
Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. Rev Mod Phys 87(3):925
Peixoto TP (2015) Model selection and hypothesis testing for large-scale network models with overlapping groups. Phys Rev X 5(1):011033
Raghavan UN, Albert R, Kumara S (2007) Near linear time algorithm to detect community structures in large-scale networks. Phys Rev E 76(3):036106
Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123
SLPA, https://github.com/sebastianliu/SLPA-community-detection
Serrat O (2017) Social network analysis. In Knowledge solutions (pp 39-43). Springer, Singapore
Srivastava A, Chelmis C, Prasanna VK (2015) The unified model of social influence and its application in influence maximization. Soc Netw Anal Min 5(1):1–15
Tang J, Tang X, Yuan J (2018) An efficient and effective hop-based approach for influence maximization in social networks. Soc Netw Anal Min 8(1):1–19
Wuchty S, Oltvai ZN, Barabási AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35(2):176–179
Xie J, Szymanski BK, Liu X (2011) Slpa: uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In: 2011 ieee 11th international conference on data mining workshops (pp. 344-349), IEEE
Yang J, Leskovec, J (2012) Defining and evaluating network communities based on ground-truth. ICDM