Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 151 (2019) 45–52

Procedia Computer Science

www.elsevier.com/locate/procedia

The 10th International Conference on Ambient Systems, Networks and Technologies (ANT)
April 29 - May 2, 2019, Leuven, Belgium

# PercoMCV: A hybrid approach of community detection in social networks

Nathanaël Kasoro[a,b], Selain Kasereka[a,b,c,e,*], Elie Mayogha[b], Ho Tuong Vinh[c,d], Joël Kinganga[a]

[a] *University of Kinshasa, Mathematics and Computer Science Department, Artificial and Business Intelligence Lab, Kinshasa, DR Congo*
[b] *Université Libre des Pays des Grands Lacs, Faculty of Sciences and Technologies, Goma, DR Congo*
[c] *Institut Francophone International (IFI), Vietnam National University in Hanoi, Vietnam*
[d] *Sorbonne University, IRD, UMMISCO, F-93143, Bondy, France*
[e] *University of South Africa, College of Science, Engineering & Technology, Department of Mathematical Sciences, Florida, South Africa*

## Abstract

Knowledge extraction in social networks is a needful tool as it touches every aspect of our lives such as politic, socio-economic, scientific, etc. Community detection is one of the objectives of this specific tool used for knowledge extraction in social networks. Many algorithms of knowledge extraction from social networks have been developed these last years. However, many of them are not constant, effective and accurate when facing these social networks with many edges. In this paper, we propose a new approach of community detection in social networks with many links between communities. The proposed approach has two steps. In the first step, the algorithm attempts to determine all communities that the clique percolation algorithm may find. In the second step, the algorithm computes the Eigenvector Centrality method on the output of the first step in order to measure the influence of network nodes and reduce the rate of the unclassified nodes. To assess this new approach, we test it on different types of networks. Relevant communities that have been detected testifies effectiveness and performance of the approach over other community detection algorithms.

*Keywords:* Social networks; Knowledge extraction; Community detection; Clique percolation; Eigenvector Centrality; PercoMCV.

## 1. Introduction

The online social network is a term used to describe services based on web technologies; these services permit users to create public, semi-public or private profile within a domain such way that they can communicate with other

* Corresponding author. Tel.: +243-821-828-964.
*E-mail address:* selain.kasereka@unikin.ac.cd

10.1016/j.procs.2019.04.010

network users to which they are connected [4]. Mustafa H. Hajeer et al. describe in [9] a social network as a set of users who frequently interact and participate in some discussion. Since their creation in the 1990s, online social networks have undergone a major revolution by enabling the training and exchange of user-generated data [11]. The social network is a graph that uses nodes and links to represent social relations [1].

Since the last decade, the analysis of social networks has become an open question for academics, industries, and business [14] [16] [18] due to the emergence of the Internet which has become a useful tool in social life. Moreover, more and more, information is stored in unstructured formats (customer e-mail, call center notes, open survey responses, news, web forms, etc.) [8]. This flow of information is a problem for many organizations who would like to find a method to collect, study and exploit this information [6]. In addition, businesses specializing in custom social media marketing often face the problem of determining the number of groups that compose the population to properly target advertisements. While studying large and / or small world networks [5] [12] [19], which are complex data, such as social networks, the number of groups that we tend to obtain cannot be known in advance [3] [5] [8]. This difficulty led us to consider another more complex problem: *The community detection*.

Many community detection methods and algorithms have been proposed to break large graphs into sub-graphs [3]. These methods can be grouped into community detection methods for static communities and dynamic communities [5]. But generally these methods do not give acceptable results for some configurations of social networks. Some problems observed drew our attention, particularly high number of unclassified nodes, the failure to consider the principle of superposition of communities, the high complexity value of the algorithm and in terms of power, and so on. Several algorithms have limitation facing to the large graph. For networks with million of interconnected nodes, the search for maximal cliques becomes very expensive in term of calculation. Indeed, since the search for maximum clique is a NP-hard problem [17]. It should also be noted that techniques based on clique percolation have an uncomfortable tendency to ignore nodes that are not part of the clique, cycle [12]. It is therefore in order to tackle some of these shortcomings that this paper proposes a new method of community detection relying on the clique percolation algorithm.

This paper is divided into four sections. In the first one we define some key concepts related to community detection. The second one presents a brief literature review on community detection. The third one displays the proposed method and the fourth one focuses on the simulation and results obtained before discussing them.

## 2. Key concepts and some common community detection algorithms

This section introduces some basic essential notions on social networks before presenting the two known community detection algorithms namely CFinder and EAGLE algorithms.

### 2.1. Definitions

1. *A Graph:* is a set of entities and interactions. Entities are called nodes and interactions edges. The graph can be oriented that is to say that we can determine the origin and the end of a relation (interaction), otherwise it is undirected [22]. Note that in this article we work with undirected graphs.
2. *Social Network, World Network and Community:* by definition social network is a network where flows are from social interactions [14]. Generally it is represented by a graph whose nodes are the actors of the network and whose links illustrate the relationships between these actors. We define a world network as opposed to an artificial network. A world network is a network based on collected data which corresponds to a reality on the ground. But a generated networks do not correspond to any real-world data [5] [22]. A community can be defined as nodes strongly linked each other and weakly linked with the rest of the network. In other words, a community is made by a set of individuals who interact (social interaction) more often with each other than with others. It is therefore a set of nodes that share something: friends, colleagues, people with similar interests, web pages with the same content [5] [7] [10] [12] [14].
3. *Clique:* is a set of nodes, all connected to each other. A maximum clique is a clique of size $k$ not included in a clique of size $k + 1$ [14] [22].

## 2.2. *Some community detection algorithms*

Many community detection algorithms have been proposed these last years. This section deals with some community detection algorithms.

### 2.2.1. *CFinder algorithm*

CFinder is the most prominent algorithm using clique percolation method (CPM) as described by Palla G. et al [15]. It was designed to study the evolution of social groups. As a successful method, CFinder is used in genetics, social networks, and so on. This algorithm can be structured into three main steps [14]:

1. Compute the set of cliques of size $k$ (parameter of the algorithm) in the target graph $G$.
2. Build a clique graph where each clique is represented by a node. Two nodes are connected by a link if the two associated cliques share $k - 1$ nodes in the graph $G$;
3. The communities in the $G$ graph are then the connected components identified in the clique graph constructed in step 2.

An example of the execution of the CFinder algorithm is shown in Figure 1 below. A limitation of this algorithm is that it requires a parameterization: the value of $k$ (the size of the communities to be considered). Consider the graph $G$ as below:

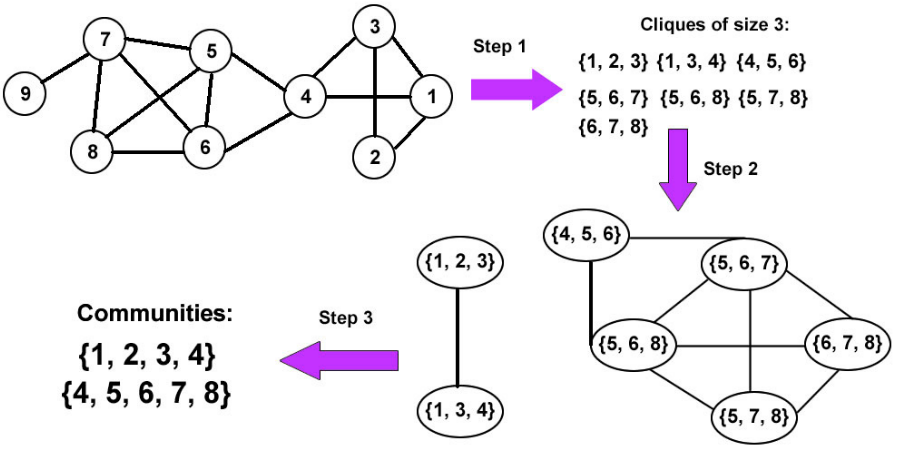

Fig. 1. An example of the clink percolation algorithm with $k = 3$ [12]

Being in principle NP-hard. The actual computational time of the CFinder algorithm, depends on the density of the network to a significantly higher degree than on the size of the network $N$.

### 2.2.2. *The EAGLE algorithm*

EAGLE is an agglomerative clustering algorithm that seeks to determine the community structure. It has the particularity of considering a set of cliques instead of a set of nodes. This algorithm uses a dendrogram of cliques. It starts by identifying all the maximum cliques that are the initial communities. Then, the communities with the highest similarity rate are merged, forming new communities, which in turn can be merged with similar communities. The optimal flat of the dendrogram is determined by using a modified version of the modularity, defined by equation 1 below [20].

$$EQ = \frac{1}{2m} \sum_{i} \sum_{v \in C_i, w \in C_i} \frac{1}{O_v O_w} \left[ A_{vw} - \frac{k_v k_w}{2m} \right] \tag{1}$$

Where $m$ is the total number of edges in the network, $O_v$ is the number of communities to which node $v$ belongs, $k_v$ is the degree of vertex $v$ and $A_{vw}$ is the element of adjacency matrix of the network. It should be noted that the estimated complexity of this algorithm is $O(n^2 + (h + n)s)$ for the first stage and $O(n^2 s)$ for the second one, where $s$ is the maximum number of cliques, $h$ the number of pairs of maximum cliques in neighbors and $n$ the number of vertices.

The presented two algorithms have the problem of parametrization of the value of $k$. The CFinder algorithm does not classify the node that does not belong to $k - clique$. This algorithm is too slow for networks with billions of nodes. We note that this algorithm is powerful to detect overlapping structures and is very flexible in execution on small (relative) networks. The EAGLE algorithm is evolving in term of research, thanks to its modified version of *Qc modularity*, this approach can simultaneously identify hierarchical structures and overlapping ones.

## 3. Materials and methods

### 3.1. Materials

Some materials were needed for the practical realization of this research. To test the proposed approach, we used data from Zachary et al. in [21]. These data concern a karate club. As simulation tools, we used Anaconda mainly designed to support programming languages like Python.

### 3.2. Methods

We adopt two specific steps to implement this approach. The first one uses the classic clique percolation method (CPM) to break the network into subnets. The second one aims to reduce the rate of unclassified nodes based on the computation of the eigen vector centrality.

- *First step:* The clique percolation method used here is inspired by the one proposed by Palla G. et al. [15]. Let us consider $G$ a given graph, in this approach, we consider as atomic community a clique of dimension $k$. And we use two stacks $A$ and $B$, the first stack $A$ will contain all the cliques of dimension $k$ which is in the graph $G$. The second stack $B$ must contain all the cliques that are adjacent. Thus, adjacent cliques with $k - 1$ nodes in common are part of the same community. The figure 2 below describes this algorithm.

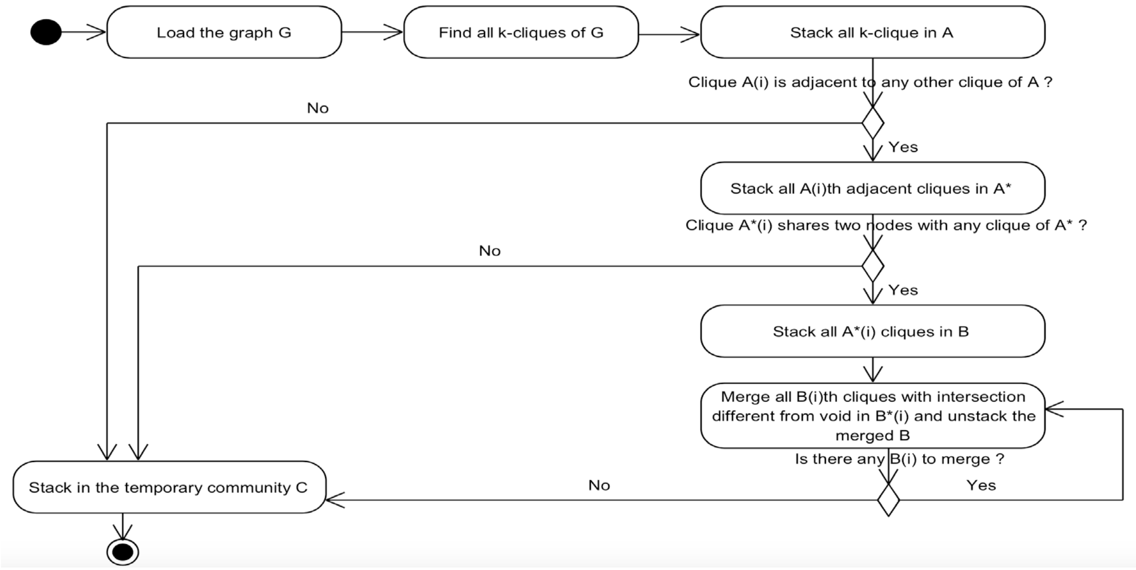
Load the graph G
Find all k-cliques of G
Stack all k-clique in A
Clique A(i) is adjacent to any other clique of A ?
No
Yes
Stack all A(i)th adjacent cliques in A*
Clique A*(i) shares two nodes with any clique of A* ?
No
Yes
Stack all A*(i) cliques in B
Merge all B(i)th cliques with intersection different from void in B*(i) and unstack the merged B
Is there any B(i) to merge ?
No
Yes
Stack in the temporary community C

Fig. 2. First-phase activity diagram of the PercoMVC algorithm

Since each clique can be visited only once, the complexity of the algorithm is therefore of the order $O(n)$ with $n$ the maximum clique number of dimensions $k$ of the initial graph.

- *Second step:* classify the nodes which do not belong to any clique. Thus, any node that is linked to the central node of a community is a member of the central node community. And we apply the measurement of centrality based on the Perron-Frobenius theorem [22] which is defined as follows:
  let $G$ be a graph, $A$ the adjacency matrix of $G$, $C_e$ is the eigenvector of centrality and $\lambda$ the eigenvalue such that:

$$C_e = \frac{1}{\lambda} A C_e \tag{2}$$

The complement to the clique percolation algorithm integrating this measure of centrality is presented in figure 3 below:

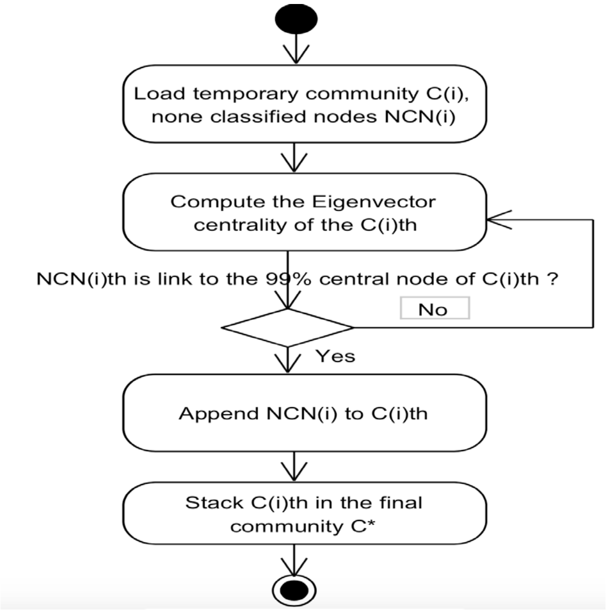
Load temporary community C(i), none classified nodes NCN(i)
Compute the Eigenvector centrality of the C(i)th
NCN(i)th is link to the 99% central node of C(i)th ?
No
Yes
Append NCN(i) to C(i)th
Stack C(i)th in the final community C*

Fig. 3. Second-phase activity diagram of the PercoMVC algorithm

Since each temporary community is visited only once, the complexity of the algorithm is therefore of the order $O(m)$ with $m$ the number of temporary communities. From where we can find the overall complexity of this new algorithm combining both by the sum of the complexities as shown in the equation 3:

$$Complexity = O(n) + O(m) \tag{3}$$

Since $n$ and $m$ are positive, and the number of cliques $n$ is greater than the one of temporary communities $m$, we have $Complexity = O(max(n, m))$. Thus, the complexity of $PercoMVC$ becomes (equation 4):

$$Complexity = O(n) \tag{4}$$

## 4. Case Study: Simulation and Evaluation

### 4.1. Simulation on an artificial network

The objective of this simulation is to test the second phase of the proposed approach. The graph initially consists of 12 nodes and 19 links and is connected. The first step found two communities. At the the second step, the node 12 that is not classified is added to the first as it is connected to the central node of this community. The evaluated graph is shown in Figure 4 below.

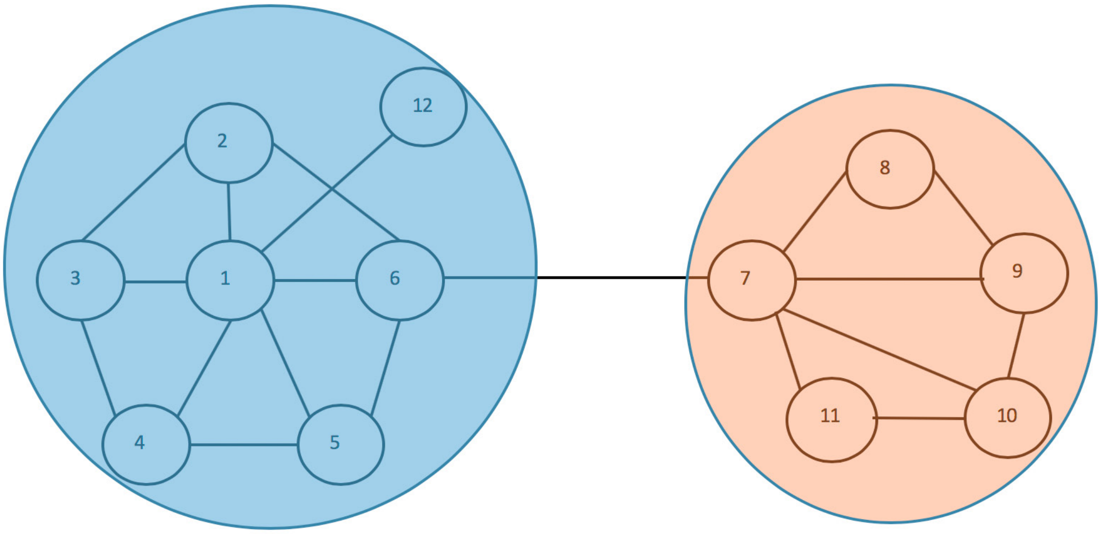

Fig. 4. Communities detected after running the PercoMVC algorithm on an artificial network

### 4.2. Simulation on a world network: the Zachary Karate Club

Zachary Karate Club is a network built from relationships between 34 members of a karate club at a university in the United States. It is a very popular network and is widely used by several algorithms [7] [13] [20] to test their performance since its community structure is known in advance. The objective of this second simulation is to test the first and second phase, and also to test the principle of overlapping communities. The Zachary Karate Club network has two communities, this information is known in advance. After applying the *PercoMCV* algorithm to this network, we obtained the result shown in figure 5.

To test the robustness of this new approach, we compared it to other known algorithms applied to the Zachary Karate Club network. The evaluation is presented in Table 1 below. The evaluation is based on the number of detected communities, the complexity of the algorithm and the value of omega-Index.

Table 1. Comparison of the new approach with some existing approaches on the Zachary Karate Club network

| Methods | Number of communities | Complexity | Omega-Index value |
|---|---|---|---|
| PercoMCV | 3 | $O(n)$ | 0,404 |
| EAGLE | 4 | $O(n^2 + (h + n)s)$ | 0.291 |
| Label Propagation | 3 | $O(n)$ | 0.372 |
| NEDIOUI M. A | 3 | $O(n)$ | 0,127 |
| CFinder | 3 | $O(n)$ | 0,095 |

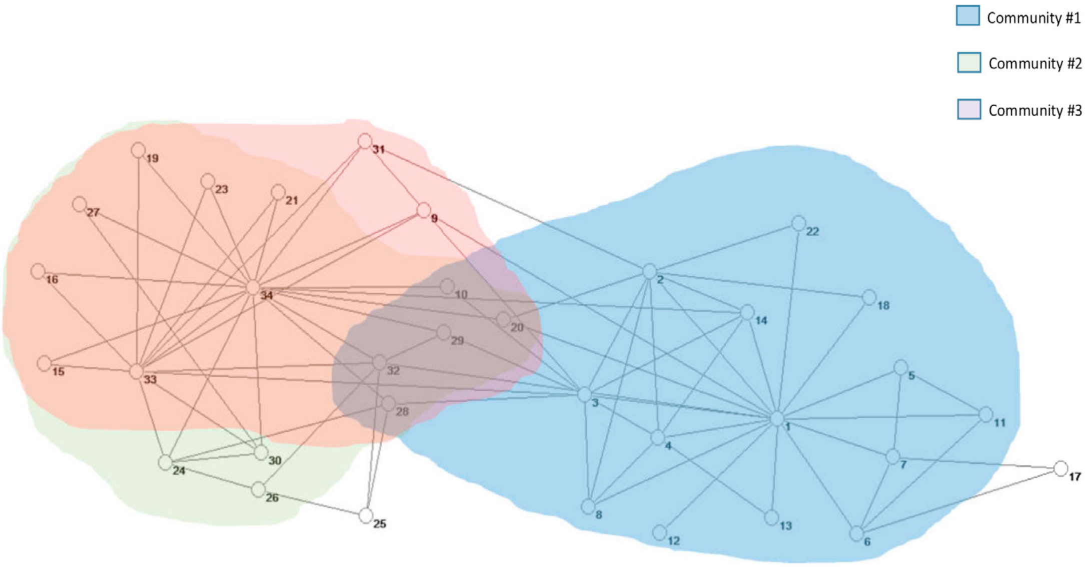

Fig. 5. Communities detected after running the PercoMVC algorithm on a world network: the Zachary Karate Club

## 5. Discussion of the results and concluding remarks

### 5.1. General discussion

From the table 1, the detection algorithm proposed, we find a complexity of $O(n)$ which is better compared to others like EAGLE. The robustness of the *PercoMCV* algorithm lies in its second phase which allows measurement of the influence of network nodes and classification of unclassified nodes. Our approach yields the same result as the Label propagation method, which is currently a reference among community detection techniques due to the linearity of its complexity. It should be noted that the results obtained in our research were evaluated on the basis of some parameters, including the number of communities, complexity, recovery and Omega-index metric. The complexity of our approach is of the order $O(n)$. Being linear and simple, our approach is more efficient than approaches with complexities of a power law. But the linearity of complexity reveals that this method is less powerful than those having a logarithmic complexity. And our approach admits the principle of overlapping communities. It allows to determine the nodes that can serve as a bridge between the communities. Using Zachary Karate Club network, we evaluate the quality of the communities detected by the proposed algorithm. We use Omega-index metric as a good measurement for evaluating detection algorithm with overlapping communities as in [2]. The results obtained here show that our approach with 0,404 of Omega-Index value is more efficient than other algorithms.

### 5.2. Concluding remarks

In this paper, our goal was to propose a new approach of community detection in social networks. This approach is one of the solutions to the problems of Clique Percolation approaches. The model has been implemented and tested on artificial and real networks. Based on the obtained results, the proposed approach responds favorably to the expectations of any community detection approach. This method benefits both the percolation method of cliques and the measurement of centrality from the Perron-Frobenius theorem. However, it has some drawbacks resulting from the clique percolation method and that only the measurement of centrality cannot dissipate. It should be noted that our research is destined to evolve, as it has been shown that community detection remains among the priorities of researchers nowadays because of the intensive use of the Internet, which generates a new form of targeted advertising and focus on several factors such as individual behavior, lifestyle, standard of living and so on. As part of the perspectives, a new measurement will be set up and it will significantly reduce the number of unclassified nodes. Taking into account dynamics in the analysis of social networks is also among the supplements to this approach to further improve this community detection method.

## Acknowledgements

The authors thank two anonymous reviewers whose comments have greatly improved this manuscript.

## References

[1] Borgatti, S.P., 2009. 2-mode concepts in social network analysis. Encyclopedia of complexity and system science 6, 8279–8291.

[2] Chakraborty, T., Dalmia, A., Mukherjee, A., Ganguly, N., 2017. Metrics for community analysis: A survey. ACM Computing Surveys (CSUR) 50, 54.

[3] Chen, J., Chen, L., Chen, Y., Zhao, M., Yu, S., Xuan, Q., Yang, X., 2018. Ga based q-attack on community detection. arXiv preprint arXiv:1811.00430 .

[4] Chen, Z., Kalashnikov, D.V., Mehrotra, S., 2009. Exploiting context analysis for combining multiple entity resolution systems, in: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, ACM. pp. 207–218.

[5] Creusefond, J., 2017. Caractériser et détecter les communautés dans les réseaux sociaux. Ph.D. thesis. Normandie Université.

[6] Efstathiades, H.A., 2013. Extract knowledge from social networks. URL: https://repository.tudelft.nl/islandora/object/uuid%3A4681f604-ef07-4ac8-86bf-42257e6c96ed. accessed: 2018-08-25.

[7] Gao, C., Ma, Z., 2018. Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. arXiv preprint arXiv:1811.06055 .

[8] Gregory, S., 2010. Finding overlapping communities in networks by label propagation. New Journal of Physics 12, 103018.

[9] Hajeer, M.H., Singh, A., Dasgupta, D., Sanyal, S., 2013. Clustering online social network communities using genetic algorithms. arXiv preprint arXiv:1312.2237 .

[10] Huang, Y., Zhan, J., Wang, N., Luo, C., Wang, L., Ren, R., 2018. Clustering residential electricity load curves via community detection in network. arXiv preprint arXiv:1811.10356 .

[11] Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! the challenges and opportunities of social media. Business horizons 53, 59–68.

[12] NEDIOUI, M.A., 2015. Fouille et apprentissage automatique dans les réseaux sociaux dynamique. Ph.D. thesis. Université Mohamed Khider-Biskra.

[13] Newman, M.E., Girvan, M., 2004. Finding and evaluating community structure in networks. Physical review E 69, 026113.

[14] Palla, G., Barabási, A.L., Vicsek, T., 2007. Quantifying social group evolution. Nature 446, 664.

[15] Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814.

[16] Raji, M., 2018. Refactoring software packages via community detection from stability point of view. arXiv preprint arXiv:1811.10171 .

[17] Reid, F., McDaid, A., Hurley, N., 2012. Percolation computation in complex networks, in: Proceedings of the 2012 international conference on advances in social networks analysis and mining (asonam 2012), IEEE Computer Society. pp. 274–281.

[18] Reihanian, A., Feizi-Derakhshi, M.R., Aghdasi, H.S., 2018. An enhanced multi-objective biogeography-based optimization algorithm for automatic detection of overlapping communities in a social network with node attributes. arXiv preprint arXiv:1811.02309 .

[19] Sallaberry, A., Zaidi, F., Melançon, G., 2013. Model for generating artificial social networks having community structures with small-world and scale-free properties. Social Network Analysis and Mining 3, 597–609.

[20] Shen, H., Cheng, X., Cai, K., Hu, M.B., 2009. Detect overlapping and hierarchical community structure in networks. Physica A: Statistical Mechanics and its Applications 388, 1706–1712.

[21] Zachary, W.W., 1977. An information flow model for conflict and fission in small groups. Journal of anthropological research 33, 452–473.

[22] Zafarani, R., Abbasi, M.A., Liu, H., 2014. Social media mining: an introduction. Cambridge University Press.