Computer Science Journal of Moldova, vol.**28**, no.1(82), 2020

Community Detection Based on Node Similarity without thresholds

Makhlouf Benazi Chaabane Lamiche

Abstract

To identify communities in social networks represented by a graph, we simply need to detect the edges that connect vertices of different communities and remove them, but the problem is what measure has to be used to identify these edges? and, how we use it? To tackle this problem, this paper proposes an efficient algorithm based on node similarity. This algorithm neither needs a predefined number of communities nor threshold to determine which edges to be deleted. The algorithm tries to add new edges for the most similar nodes to strengthen intra-community links and remove edges between the least similar nodes to weaken links between communities. In order to prove its efficiency, the algorithm was evaluated with synthetic and real-world networks.

Keywords: Social network, Community detection, node similarity, modularity, GN algorithm.

MSC 2010: 91C20.

1 Introduction

Today, with the emergence of new technologies of communication and the Internet especially web 2.0, social networks have grown exponentially in size and complexity. In order to understand the structure of these networks, analyze its characteristics and extract useful information and knowledge, several fields of research in social network analysis

have appeared. Some focus on the identification of the most influential individual (leadership), others on missing links, and the third – on network dynamics. In this paper we will focus on community detection.

The concept of community can be seen as a group of densely interconnected nodes compared to other nodes [1]. To detect the communities that constitute a network, researchers working in this field have proposed several approaches, see for instance [2], [3], recent surveys on the subject. These methods can roughly be grouped as methods emerging from graph theory known as partitioning algorithm such as spectral bisection algorithm [4], and methods from sociology known as hierarchical clustering algorithms [5], [6]. Researchers working in the field of sociology noticed that individuals belonging to the same community share some similarities, such as gender, age, common interests or professional activity etc. After measuring the similarity matrix, they incorporate the two nodes with the highest similarity together iteratively (agglomerative) such as Newman Fast Greedy FN [7] and the Concor algorithm of [8] or they iteratively remove the edge with the lowest similarity (divisive) such as Radicchi [9], Spectral [10] and Girvan-Newman (GN) algorithm [11]. We find also approximation algorithms which seek to maximize or minimize the value of a given quality function. Certainly, the most popular is the modularity [6]; such is in fast Newman algorithm (FN) [7], Genetic Algorithm [12], Simulated annealing [13], [14], PSO [15].

In this paper we propose a new algorithm for communities detection based on node similarity, where we use four different similarity measures to add new edges so that communities can get bit-by-bit closer to what we call a clique in graph theory, and to identify edges that will be removed so that communities can break away from each other.

The rest of this paper is organized as follows. We firstly present some basic concepts, so graph, similarity and Modularity are briefly introduced in Section 2. Then some research related to our algorithm is considered in Section 3. In Section 4 our approach is presented in detail. In order to show the effectiveness of our approach, in Section 5 we test our algorithm on different artificial and real-world networks with four different similarity measures, and make comparisons with Girvan-

Newman (GN) algorithm [11] and Fast Newman (FN) algorithm [7]. Finally, we conclude and present some perspectives in Section 6.

2 Preliminary Knowledge

2.1 Graph

We often have recourse to graphs to analyze a social network in order to identify local or global patterns, to locate influential entities, or to examine network dynamics. A graph G (undirected and unweighted) is an ordered pair $G(V, E)$, where V is the set of nodes or vertices that represent individuals, and $E \subset V \times V$ is the set of edges or links that represent interactions between individuals (friendship, collaboration, love etc.). The degree k_i of a vertex i is the number of edges connecting i to the rest of the graph. If A is the adjacency matrix of G, then $k_i = \sum_j A_{i,j}$.

2.2 Graph partitioning

The goal of graph partitioning is to divide G into k disjoint sub-graphs $G_i = (V_i, E_i)$, in which $\forall\ i \neq j\ :\ V_i \cap V_j = \Phi\ and\ \bigcup_1^k V_i = V$.

2.3 Modularity

There are an exponential number of diverse alternative partitions. Enumerating all these partitions is an NP-Complete problem [16]. Moreover, not all partitions of a graph are equally good. In order to choose the best partition Newman and Girvan [11] introduced a metric that computes the difference between the fraction of edges for a given partition of the original graph and a random graph having a similar degree distribution as the original. This metric is known as Modularity.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{i,j} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j), \tag{1}$$

where m is the number of edges. A denotes the adjacency matrix, where $A(i, j)$ is equal to 1 if there is an edge between nodes i and j, and 0

otherwise. k_i and k_j are the degrees of vertices i and j respectively, C_i and C_j are the communities of the vertices i and j respectively, and $\delta(C_i, C_j)$ is equal to 1 if vertices i and j belong to the same community, and 0 otherwise.

2.4 Structural similarity

A community represents a set of nodes that are more similar to each other, but dissimilar from the rest of the network [6]. But how similar are two nodes? To answer this question several methods have been proposed in the literature. Some of these methods measure the distance between two nodes, others – the local paths and the third count how many neighbors two nodes have in common. Here are some of those measures that have been tested with our algorithm.

In what follows $\Gamma(x)$ represents the set of neighbors of node x, $|\Gamma(x)|$ denotes its degree, n is the number of nodes, and $A(x, i)$ denotes an element of the adjacency matrix.
Jaccard Index: [17]

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}, \tag{2}$$

Cosine Index: [18]

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{|\Gamma(x)| \times |\Gamma(y)|}}, \tag{3}$$

Pearson Coefficient: [19]

$$S_{xy} = \frac{\sum ((A(x,i) - \bar{x}) * (A(y,i) - \bar{y}))}{\sqrt{\sum (A(x,i) - \bar{x})^2} * \sqrt{\sum (A(y,i) - \bar{y})^2}}, \tag{4}$$

where: $\bar{x} = \frac{|\Gamma(x)|}{n}$.
Hub Promoted Index: [20]

$$S_{xy} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(|\Gamma(x)|, |\Gamma(x)|)}. \tag{5}$$

3 Related Work

Our approach is closely related to Girvan Newman algorithm [11], which can be expressed as follows:

1. Calculate edge betweenness for every edge in the graph.
2. Remove the edge with highest edge betweenness.
3. Calculate edge betweenness for remaining edges.
4. Repeat steps 2-4 until all edges are removed.

Despite its simplicity, Girvan Newman Algorithm has several fundamental limitations. First, it uses betweenness centrality to identify which edge to delete. Betweenness centrality is a global measure that identifies edges with the highest number of the shortest paths which pass through them. The calculation of edge betweenness has time complexity $O(MN)$; therefore, the algorithm's time complexity is $O(M^2N)$. Here M is the number of edges and N is the number of nodes, unlike our approach, where we use a local similarity measure based on neighborhood. Removing a single edge at each iteration is another problem which will be remedied in our approach by removing multiple edges at every iteration. In order to increase the accuracy of our algorithm new edges are added at each iteration to strengthen communities. Finally, there is no implicit stopping criterion, either we give the number of communities or the algorithm will continue until all edges are removed, in the latter case the algorithm returns partition with the highest modularity.

Our approach is also related to algorithms that use node similarity like in [21], where the authors proposed an algorithm for detecting community. In their paper they introduce the similarity threshold ϵ which takes different values for different datasets. For example, with Zachary's karate club network the authors put $\epsilon = 0.9$ and with college football network they put $\epsilon = 0.75$. But with this strategy how to determine the value of this threshold? We have observed that this strategy is frequently used in the literature, as in [22] and [23], unlike our approach, where we do not use any parameter or threshold.

We find also approaches like DBSCAN [24], DENGRAPH [25], SCAN [26], DEEN [27], and SMP [28]. However not all these algorithms are free-parameter, they always depend overly on manually choosing thresholds of measure that have been used, a minimum cluster size or cluster number, which can be difficult to determine.

4 The Algorithm

We have developed an effective algorithm based on nodes similarity. The main idea in our algorithm is: 1) compute the similarity matrix at every iteration; 2) try to remove edges that have the least similarity on both sides; and 3) add new edges for nodes that have the most similarity on both sides for every node in the graph. In other terms, to delete a link (i, j), the node least similar to node i must be j and the node least similar to node j must be i, and to add a new link (i, j), the node most similar to node i must be j and the node most similar to node j must be i.

We can describe our algorithm as follows:

1. Sort nodes in descending order using node degree.
2. Calculate similarity matrix S (eq. 2, 3, 4 or 5).
3. Remove all edges that fit both of the following conditions:
   - For every two nodes i, j that have common edge, the node least similar to node i must be j, and the node least similar to node j is i.
   - The deletion does not generate an orphan node.
4. For every two nodes i, j add new edge if the node most similar to node i is j, and the node most similar to node j is i.
5. Repeat steps 2-4 until the set of deleted links is the same added or a maximum number of iterations is reached.
6. Compute community vectors using Hopcroft & Tarjan algorithm [29].

7. Return Community vector with the best modularity calculated using formula (1).

5 Experimental Evaluation

In order to measure the accuracy of our algorithm vis-à-vis ground truth and compare it to other algorithms we use NMI (Normalized Mutual Information) [30] as measure of partition (i.e., clustering) similarity. It takes values in the interval [0, 1].

The NMI of two partitions A and B of a graph is given as follows:

$$I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} M_{ij}\ log\left(\frac{M_{ij}n}{M_{i.}M_{.j}}\right)}{\sum_{i=1}^{C_A} M_{i.}\ log\left(\frac{M_{i.}}{n}\right) + \sum_{j=1}^{C_B} M_{.j}\ log\left(\frac{M_{.j}}{n}\right)}, \qquad (6)$$

where C_A and C_B denote the numbers of communities in partitions A and B respectively. The notation M_{ij} denotes the element of matrix (M) $C_B \times C_B$, representing the number of nodes in the i^{th} community of A that appear in the j^{th} community of B. The sum over row i of matrix M is denoted by $M_{i.}$, and that over column j – by $M_{.j}$; and n is the number of nodes. I(A, B) = 1, if A=B. I(A, B) = 0, if A and B are completely different.

In this section, we applied our algorithm to synthetic and real-world networks.

5.1 Synthetic Benchmark Networks

Girvan and Newman [31] proposed a benchmark network with four communities, with every community containing 32 vertices (for a total number of vertices $n = 128$), and fixed the average total degree of each node to 16: $k = Z_{in} + Z_{out} = 16$, where Z_{in} is the number of edges connecting a node with the others in its own community, and Z_{out} is the number of edges connecting a node with the rest of the network. We vary Z_{out} from 0 to 8. The values of the NMI measures are shown in Figure 1. As it can be seen from Figure 1, the results given by Pearson Coefficient and Hub Promoted Index are better than the other indices

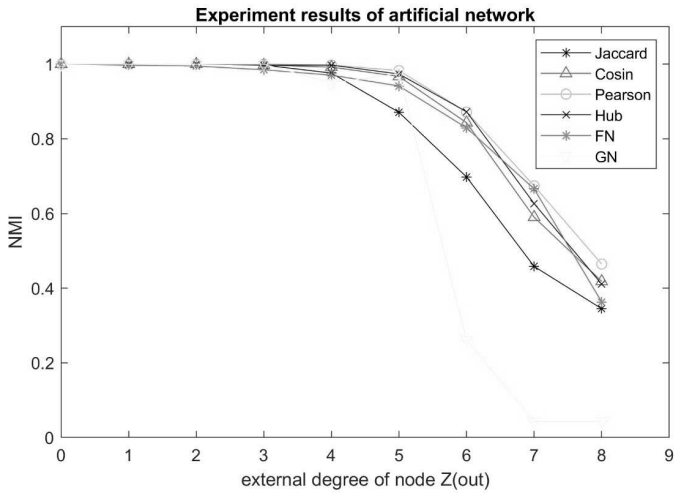

Figure 1. Experiment results of artificial network

and better than Fast Newman (FN) algorithm whatever the value of Z_{out}. Jaccard index obtains the worst performances. We observe also that GN fails to find the true partition when $Z_{out} > 5$.

5.2 Real-world Networks

Just like in other algorithms for detection community, we selected the following 4 well known real-world networks to further verify the performance of our algorithm: 1) the Zachary's Karate Club network; 2) Dolphins network; 3) the American College Football network; 4) Krebs' political books. The properties of these four real-world networks are listed in Table 1.

To verify the performance of our algorithm, we compared it with two algorithms considered as reference in the field of communities detection, the GN algorithm and FN algorithm. GN or Girvan and Newman [11] algorithm calculates the betweenness centrality of all edges (number of the shortest paths passing through an edge) and removes the edge with the biggest betweenness recursively. The second algo-

Networks	Ref.	C	N	M	Q
Zachary	[32]	2	34	78	0.37
Dolphins	[33]	2	62	159	0.38
Football	[31]	12	115	613	0.55
Polbooks	[10]	3	105	441	0.41

Table 1. Properties of real-world networks employed in the tests. C – number of communities, N – number of nodes, M – number of edges, Q – original modularity.

rithm is FN (Fast Newman) [7], a greedy algorithm that tries to maximize the modularity.

Since FN is a stochastic optimization algorithm, we perform the experiments 10 times on each network. The average value of Q and NMI are calculated. The results are shown in Table 2. The best results are shown in **bold**.

It can be seen from Table 2 and Figure 3 that our algorithm can correctly detect the community structure on Zachary's karate club (see Figure 2, where solid lines indicate edges that have been added and dashed lines indicate edges that have been deleted), except when we use Pearson as similarity measure although it can detect it from the first iteration. But because we have used the modularity maximization as criterion to choose the best partition, it will select other partition that has the best modularity. The same phenomenon is seen with Dolphins network, when we use hub as similarity measure, where the correct partition is found in the third iteration, but the maximum modularity is found in the 5_{th} iteration. Figure 4 shows the evolution of NMI and modularity for Dolphins network. On Football network, our approach performs well especially when we use Jaccard, Cosine or Hub as similarity measure, until 0.9306. Regarding the Polbooks network, the NMI reached its maximum 0.5836 when using Cosine index as similarity measure.

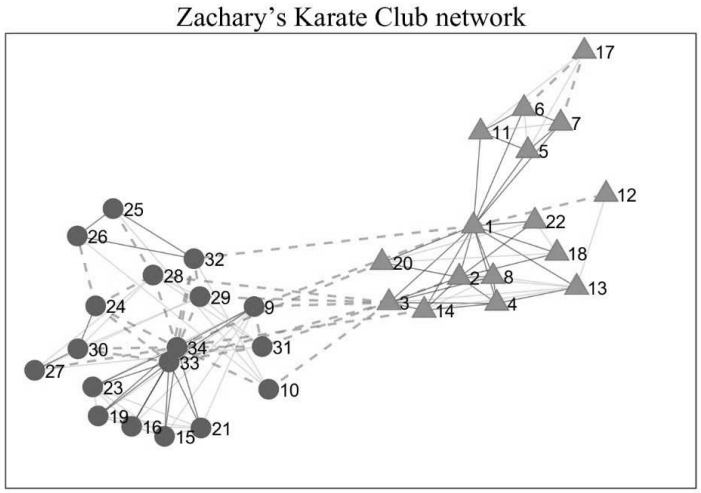

Figure 2. Result of Karate Club network obtained by our algorithm.

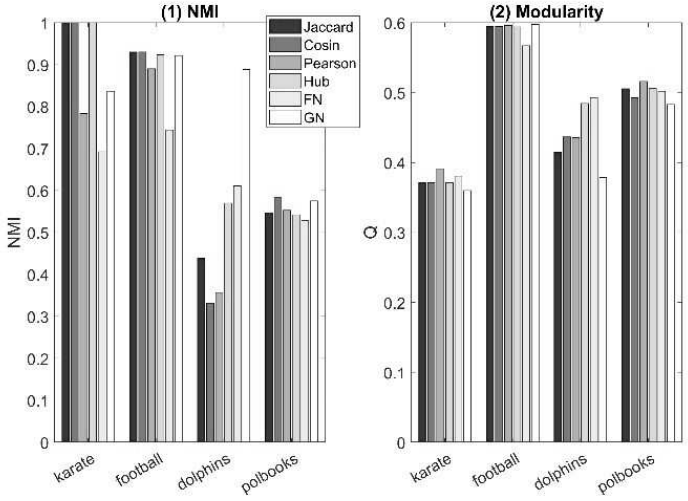

Figure 3. NMI (1) and modularity (2) of four real-world networks obtained by our algorithm and fast Newman.

algorithm	**Zachary**		**Dolphins**	
	Q	**NMI**	**Q**	**NMI**
Fast Newman	0.3807	0.6925	**0.4942**	**0.5919**
Girvan Newman	0.3600	0.8365	0.3787	0.8888
Jaccard	0.3715	**1**	0.4160	0.4200
Cosine	0.3715	**1**	0.4338	0.3260
Pearson	**0.3949**	0.7534	0.4417	0.3573
Hub	0.3715	**1**	0.4849	0.5719
algorithm	**Football**		**Polbooks**	
	Q	**NMI**	**Q**	**NMI**
Fast Newman	0.5698	0.7460	0.5019	0.5292
Girvan Newman	0.5976	0.9218	0.4831	0.5754
Jaccard	0.5946	**0.9306**	0.5057	0.5471
Cosine	0.5946	**0.9306**	0.4926	**0.5836**
Pearson	**0.6012**	0.9019	**0.5176**	0.5575
Hub	0.6007	0.9195	0.5040	0.5413

Table 2. NMI and modularity of four real-world networks obtained by our algorithm and fast Newman.

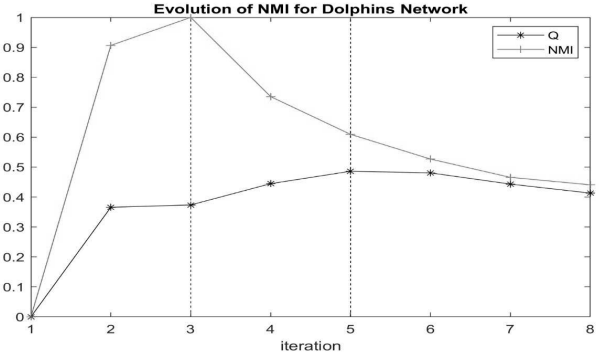

Figure 4. Evolution of NMI and modularity for Dolphins network.

6 Conclusions

Community detection is a very hard problem that has not yet been satisfactorily solved despite many methods have been proposed. In this paper, we have proposed a new algorithm to find high quality communities in social network based on node similarity that, we think, performs well. Experimental results show that the algorithm achieves better performance compared to FN and GN algorithms, especially on real world networks. Finally, it is worth to mention that our algorithm can be used with several similarity measures, which makes it more convenient and more flexible in real application, especially if we know that each application domain has its own measure of similarity. In future work, we will focus on optimizing the complexity of our algorithm to be able to apply it on larger networks.

References

[1] A. Clauset, M. E. Newman, and C. Moore, "Finding community structure in very large networks," *Physical review E*, vol. 70, no. 6, p. 066111, 2004.

[2] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics reports*, vol. 659, pp. 1–44, 2016.

[3] Q. Cai, L. Ma, M. Gong, and D. Tian, "A survey on network community detection based on evolutionary computation," *International Journal of Bio-Inspired Computation*, vol. 8, no. 2, pp. 84–98, 2016.

[4] E. R. Barnes, "An algorithm for partitioning the nodes of a graph," *SIAM Journal on Algebraic Discrete Methods*, vol. 3, no. 4, pp. 541–550, 1982.

[5] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[6] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[7] M. E. Newman, "Fast algorithm for detecting community structure in networks," *Physical review E*, vol. 69, no. 6, p. 066133, 2004.

[8] R. L. Breiger, S. A. Boorman, and P. Arabie, "An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling," *Journal of mathematical psychology*, vol. 12, no. 3, pp. 328–383, 1975.

[9] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," *Proceedings of the national academy of sciences*, vol. 101, no. 9, pp. 2658–2663, 2004.

[10] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[11] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical review E*, vol. 69, no. 2, p. 026113, 2004.

[12] C. Pizzuti, "Ga-net: A genetic algorithm for community detection in social networks," in *International conference on parallel problem solving from nature*. Springer, 2008, pp. 1081–1090.

[13] R. Guimera and L. A. N. Amaral, "Functional cartography of complex metabolic networks," *nature*, vol. 433, no. 7028, p. 895, 2005.

[14] Z. Masdarolomoor, R. Azmi, S. Aliakbary, and N. Riahi, "Finding community structure in complex networks using parallel approach," in *2011 IFIP 9th International Conference on Embedded and Ubiquitous Computing*. IEEE, 2011, pp. 474–479.

[15] C. Cao, Q. Ni, and Y. Zhai, "A novel community detection method based on discrete particle swarm optimization algorithms in complex networks," in *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2015, pp. 171–178.

[16] M. R. Garey, D. S. Johnson, and L. Stockmeyer, "Some simplified np-complete problems," in *Proceedings of the sixth annual ACM symposium on Theory of computing*. ACM, 1974, pp. 47–63.

[17] P. Jaccard, "Étude comparative de la distribution florale dans une portion des alpes et des jura," *Bull Soc Vaudoise Sci Nat*, vol. 37, pp. 547–579, 1901.

[18] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. mcgraw-hill, 1983.

[19] K. Pearson, "Vii. note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[20] E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A. Barabasi, "Community structure in social and biological networks," *Science*, vol. 297, p. 1553, 2002.

[21] Y. Pan, D.-H. Li, J.-G. Liu, and J.-Z. Liang, "Detecting community structure in complex networks via node similarity," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 14, pp. 2849–2857, 2010.

[22] T. Dang and E. Viennet, "Community detection based on structural and attribute similarities," in *International conference on digital society (icds)*, 2012, pp. 7–12.

[23] X. Wang, G. Liu, J. Li, and J. P. Nees, "Locating structural centers: A density-based clustering method for community detection," *PloS one*, vol. 12, no. 1, p. e0169355, 2017.

[24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[25] T. Falkowski, A. Barth, and M. Spiliopoulou, "Dengraph: A density-based community detection algorithm," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 2007, pp. 112–115.

[26] X. Xu, N. Yuruk, Z. Feng, and T. A. Schweiger, "Scan: a structural clustering algorithm for networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 824–833.

[27] P. Jancura, D. Mavroeidis, and E. Marchiori, "Deen: a simple and fast algorithm for network community detection," in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, 2011, pp. 150–163.

[28] K. Zhou, A. Martin, and Q. Pan, "A similarity-based community detection method with multiple prototype representation," *Physica A: Statistical Mechanics and its Applications*, vol. 438, pp. 519–531, 2015.

[29] J. Hopcroft and R. Tarjan, "Algorithm 447: efficient algorithms for graph manipulation," *Communications of the ACM*, vol. 16, no. 6, pp. 372–378, 1973.

[30] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, p. P09008, 2005.

[31] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[32] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

[33] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behavioral Ecology and Sociobiology*, vol. 54, no. 4, pp. 396–405, 2003.

Makhlouf Benazi, Chaabane Lamiche

Received September 15, 2019
Revised January 21, 2020

Makhlouf Benazi
Department of computer science,
Faculty of mathematics and computer science,
Mohamed Boudiaf University of Msila, Msila, 28000, Algeria
E–mail: Makhlouf.benazi@univ-msila.dz

Chaabane Lamiche
Department of computer science,
Faculty of mathematics and computer science,
Mohamed Boudiaf University of Msila, Msila, 28000, Algeria
E–mail: Chaabane.lamiche@univ-msila.dz