A hybrid Cat Optimization and K-median for Solving Community Detection Problem

Amany A. Naem[1], Lamiaa M. El Bakrawy[1] and Neveen I. Ghali[1]
[1]Al-Azhar University, Faculty of Science, Cairo, Egypt
Email: manoo_basom@yahoo.com, lamiaabak@yahoo.com, nev_ghali@yahoo.com

Abstract

Detecting the hidden community structure, which is conclusive to understanding the features of networks, is an important topic in Social Network Analysis (SNA). Community detection problem is the process of network clustering into similar clusters. K-median clustering is one of the popular techniques that has been applied in clustering as a substitute of K-means algorithm but it attempts to minimize the 1-norm distance between each point and its closest cluster center. The problem of clustering network can be formalized as an optimization problem where a qualitatively objective function that captures the intuition of a cluster as a set of nodes with better internal connectivity than external connectivity is selected to be optimized. Cat Swarm Optimization (CSO) is one of the latest population based optimization methods used for global optimization. In this paper, a hybrid cat optimization and K-median for solving the community detection problem is proposed and named as K-median Modularity CSO. Experimental results which are applied on real life networks show the ability of the hybrid cat optimization and K-median to detect successfully an optimized community structure based on putting the modularity as an objective function.

Keywords: Community Detection, Social Network, K-median Clustering, Cat Swarm Optimization, Modularity.

1 INTRODUCTION

A social network is a group of people associated with social relationships. Studying a social network denotes the study and knowledge of connections and dependencies between members and groups that belong to it. Facebook, LinkedIn, Instagram, and Bebo are popular examples of social networks where they are attracted to millions of users, these users log in, exchange messages or pictures, and generally interact with friends or cooperators [1].

Generally, communities are groups of nodes (members) that are connected heavily inside the group but connected sparely with the rest of the network. Community structure is the key feature for unmasking the global property in social networks, which is very consequential for studying social networks. Community detection in large networks is potentially very advantageous. Nodes belonging to a tight-knit community are more than likely to have other attributes in common. For example on the World Wide Web (WWW) a cluster can be looked at as information or as physical links and paths connecting to each other. In biochemical or neural networks, communities may be functional groups, and unplugging the network into such groups would simplify functional analysis considerably [2][3][4].

Many classic methods have been presented to detect community structures in social networks. They can be roughly codified into two categories. The first category uses heuristic strategies, such Girvan-Newman (GN) algorithm [5], Wu-Huberman (WH) algorithm [6], and Hyperlink Induced Topic Search (HITS) algorithm [7] etc. The second category selects optimization methods or approximation methods, such spectral method [8]. Afterward, Girvan and Newman [9] introduced a new technique called Modularity. As modularity optimization is typically NP-hard optimization, various NP-hard optimization techniques are used to maximize it. Others authors depended on Meta heuristic techniques such Pizzuti [10] proposed a genetic based approach to discover communities in social networks. Their algorithm was a simplex but effective fitness function able to identify densely connected groups of nodes with sparse connections between groups. Honghaoet et al. [11] suggested an ant colony optimization (ACO) based approach to discover communities. They demonstrated that ACO based approach results in a significant enhancement in modularity values as compared to existing heuristics in the literature. Masdarolomoor et al. [12] proposed a novel method for community detection in networks and used to simulated annealing to maximize the modularity. Their algorithm was evaluated by modularity metric and worked better in time and accuracy compared to similar methods. Song et al. [13] applied discrete Bat Algorithm to the community detection of showing networks and achieved good results. Their algorithm was powerful practical value. Barawy et al. [14] presented the idea of using the results of an optimization algorithms Particle Swarm Optimization (PSO) and Exponential Particle Swarm Optimization EPSO as input to the k-means clustering algorithm in order to have a well community detection for social network data. In this paper, a hybrid Cat Swarm Optimization and K-median for solving the community detection problem is proposed and named as K-median Modularity CSO. Where using K-median clustering to detect the community and CSO for optimizing the modularity. By setting the modularity as an objective function in order to have high value for modularity as a well community detection for social network data.

The remainder of the paper is organized as follows: Brief introduction on community detection problem, Cat Swarm Optimization algorithm, and K-median algorithm are introduced in Sections 2. The details of the proposed method are presented in Section 3. Section 4 shows our experimental results on datasets social networks . Finally, we supply conclusions in Section 5.

2 PRELIMINARIES

2.1 Community Detection Problem

Identifying network communities can be viewed as a problem of clustering a set of nodes into communities, but a node can belong multiple communities at once. Because nodes in communities involve common properties or attributes and they have many relationships among themselves. Community detection algorithms aim to find communities based on the network structure, to existing groups of nodes that are heavily connected. This problem can be modeled as an optimization problem where one usually needs to optimize the given fitness measure [14][15]. So, to evaluate the clustering performance, modularity metric is put into use as a measure of the quality of a particular clustering of a network of that

considers communities coupling (external relations among communities) and cohesiveness (internal relations within communities) [16].
Modularity function was proposed by Girvan and Newman in 2004 [9], which is one of the most famous community detection measures.
Suppose we have a network that includes n vertices, and let the number of edges between vertices i and j be A_{ij}, which will usually be 0 or 1, so the quantities A_{ij} are the elements of the so-called adjacency matrix. Concurrently, the expected number of edges between vertices i and j if edges are placed at random is $k_i * k_j/2m$, where k_i and k_j represent the degrees of the vertices and m is the total number of edges in the network. So the modularity can be formalized as equation (1) [16]:

$$Q = \frac{1}{2m}\sum_{i,j}(A_{ij} - \frac{k_i * k_j}{2m})\delta(H_i, H_j) \quad (1)$$

Q denotes the modularity of network. H_i and H_j are the identity of the community which the node i and j belong to in certain iteration respectively. If vertices i and j are in the same community, $\delta(H_i, H_j)$= 1, else 0.

2.2 Cat Swarm Optimization Algorithm

By close investigation on the behavior of cats in nature Chu et al. in [17] proposed a novel optimization algorithm. Where they first decided how many cats they would like to use in the iteration by defining a mixture ratio (MR) which dictates the joining of seeking mode with tracing mode, then they applied the behavior of cats into CSO to solve the problems. According to their findings, cats spend most of their time when they are awake on resting. While they are resting, they move their position neatly and slowly, sometimes they dont move at all. The CSO retains the best solution until it arrives the end of the iterations. So the CSO algorithm composes of two modes Seeking Mode and Tracing Mode [17][18]. An itemized descriptions of these modes are given down:

2.2.1 Seeking Mode

Seeking mode corresponds to the resting state of the cats. In this mode, they look around and seek for the next position to move. There are four necessary factors in this mode: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC), and self-position considering (SPC) [17][20].

- SMP determines the size of seeking memory for every cat, which marks the points sought by the cat.
- SRD is used to represent the mutation ratio for the selected dimensions.
- CDC shows how many dimensions will be change.
- SPC is a Boolean variable, which chooses whether existing position of cat will be with the candidates to move to or not.

The seeking mode is reminded as follows:

Step 1: Create L copies of the present position of cat_k, where L= SMP. If the value of SPC is true, let L= (SMP - 1), then keep the present position with the candidates.

Step 2: According to CDC, at random plus or minus SRD percents the present values and replace the old ones this apply on every copy.

Step 3: Compute the fitness values (FS) of each candidate point.

Step 4: If it happens that the fitness functions for all of the cats have exactly the same values, assign a similar probability to all of the candidates, else compute the selecting probability of each candidate point according to equation (2).

$$P_i = \frac{FS_i - FS_b}{FS_{max} - FS_{min}}, 0 < i < j \tag{2}$$

Where FS_i is the fitness of i^{th}. If the aim of the fitness function is to find the minimum solution, $FS_b = FS_{max}$, otherwise $FS_b = FS_{min}$.

Step 5: Randomly choose the point to move to from the candidate points, and replace the position of cat_k.

2.2.2 Tracing Mode

At tracing mode cat tries to trace goals. In this mode, the next move of each cat is identified based on based on the velocity of the cat and the best position found by members of cat swarm. This mode can be abstracted in three steps as follows [19][18]:
Step1: Update the velocities for every dimension ($v_{k,d}$) according to equation (3).

$$v_{k,d} = v_{k,d} + r_1 c_1 (x_{gbest,d} - x_{k,d}), d = 1, 2, ..., M. \tag{3}$$

$x_{gbest,d}$: Best position of cat, who has the best fitness value.
$x_{k,d}$: Position of cat_k.
$v_{k,d}$: Velocity of cat_k.
r_1: Random value in the range of [0, 1].
c_1: Constant.
M: Dimensional solution space.

Step2: Check if the velocities are within the limits of velocity. In case the new velocity is over range, set it to the limits.

Step3: Update the position of cat_k according to the following equation.

$$x_{k,d} = x_{k,d} + v_{k,d} \tag{4}$$

2.3 K-Medians Clustering Algorithm

The performance of k-medians clustering algorithm is homologous to k-means clustering algorithm, but they are different in the step for updating cluster center, the median of the same cluster becomes the new cluster center, rather than the average value. The co-ordinate of a median is single dimension between each point [21][22].
Given a set of points, randomly choose k points from the data to be the initial cluster

centers. Place the data into a cluster with the closest cluster center. The distance between point of data and cluster center is evaluated using equation (5).

$$d(x, c) = \|x - c\| \tag{5}$$

Where x point in data and c cluster center. The k-medians algorithm attempts to make k disjoint cluster that minimize the following equation.

$$U = \sum_{i=1}^{k} \sum_{x \in D} \|x - c_i\| \tag{6}$$

x = member of data D.
c_i = cluster center i.
k = number of clusters.
This means that the center of every cluster center minimizes the objective function. This minimization is defined by equation (7) [23].

$$L = min \sum_{i=1}^{k} \sum_{x \in D} \|x - c_i\| \tag{7}$$

3 THE PROPOSED METHOD

The proposed method (K-median Modularity CSO) in this research consists of two main parts, clustering the data by using K-median and searching for the best modularity as a well community detection for social network data by applying CSO algorithm. Steps of proposed method are summarized down:

- **Step 1: Defining The initial Cluster Center**
  Randomly select k points from the data points to be the initial cluster centers.
- **Step 2: Grouping Data into Clusters**
  Place the data into clusters with the closest cluster center by using equation (7).
- **Step 3: Calculating The modularity**
  Modularity is the fitness function in this method. The value of modularity is calculated using equation (1).
- **Step 4: Clustering with CSO**
  It is not possible to directly apply the CSO algorithm into clustering problem, there are few adjustments needs to be done in CSO algorithm. These adjustments are reminded as follows:
  - Each cat passes to the seeking and tracing mode by deleting mixture ratio. This change is designed to reduce the time to find the best cluster centers.
  - In seeking mode, counts of dimension to change (CDC) were assumed always is equal 100% value, therefore all dimensions will be varied.

  We apply the CSO algorithm on cluster centers to get best modularity value and best cluster centers for each data.
- **Step 5: Repeat Step 4**
  Recurrence step 4 until it reaches the stop criteria.

The steps of proposed method can summarize in Fig. 1.

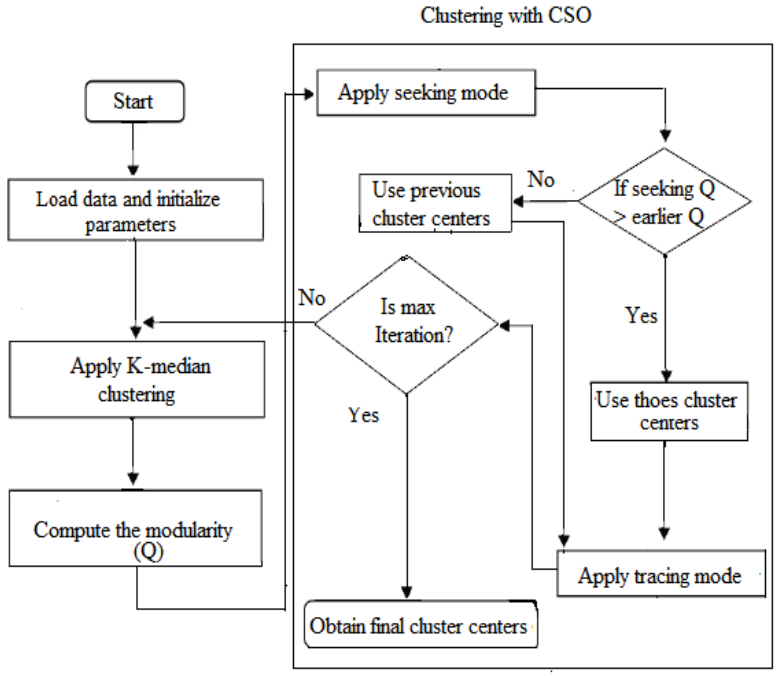

Fig. 1: The proposed method.

4 EXPERIMENTAL RESULTS

4.1 Datasets

In the section, the proposed method (K-median Modularity CSO) is applied on four real life social networks datasets:

- The Zachary Karate Club:
  It was studied by Wayne W. Zachary from 1970 to 1972 and was observed from the members of a university karate club. The graph of Karate network composes of 34 nodes and 78 edges. Where each member of the club was represented by node and each relation between two members in the club was represented by edge. The problem was often discussed the use of this dataset to find groups of people after a struggle arose between teachers, which led to the divide the karate club into two group [24].
- The Bottlenose Dolphins network:
  This was named by social network of bottlenose dolphins and was aggregated by Lusseau from 1994 to 2001. The network was built the basis of the behavior of 62 bottlenose dolphins that live in Doubtful Sound, New Zealand. So the nodes represent the bottlenose dolphins and edges represent a frequent associations [25].
- American College football network:
  It represented American football games between American colleges during a regular season in Fall 2000, as reorganized by M. Girvan and M. Newman. Each node in the network refers to team and edge refers to game of regular season between the two

teams they connect [26].

- The Polbooks network:
  It network of books about US politics showed at the time of the 2004 presidential election and marketed by the online bookseller Amazon.com. Edges between books act frequent co purchasing of books by the same shoppers. This network was combined by V. Krebs [27].

4.2 Results and Analysis

The results of modularity after applying the proposed method (K-median Modularity CSO) are compared to K-means Modularity PSO, K-means Modularity Bat optimization, K-means Modularity CSO, K-median Modularity PSO, K-median Modularity Bat optimization, GN [28], FN [28], BGLL [28], and HSCDA [28] on the real datasets. Since modularity is a famous community quality measure used widely in community detection, we used it as a quality measure for the result community structure of all other objectives. Also Normalized Mutual Information (NMI) is used to compare the accuracy of the outcome communities where it is used to compute the likeness between two parts. NMI is described in equation (8) [29].

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{C_X}\sum_{j=1}^{C_Y} C_{ij} log(C_{ij}N/C_iC_j)}{\sum_{i=1}^{C_X} C_i log(C_i/N) + \sum_{j=1}^{C_Y} C_j log(C_j/N)} \tag{8}$$

$NMI(X,Y)$: Denotes NMI for two parts X and Y.
C_{ij}: Number of nodes assigned to ith community in part X, and jth community in part Y.
C_i: Number of nodes in part X assigned to ith community.
C_j: Number of nodes in part Y assigned to jth community.
C_X: The Community Number in part X.
C_Y: The Community Number in part Y.
N: Total number of nodes.

If NMI is equal 1 it means that the two consequences consistent completely. If NMI is equal zero it means that the two consequences inconsistent completely.

In the experiments on real networks, we ran each method 50 times on each network and reported the average NMI and the average of modularity.

Table 1 contains the information of the four networks and some parameters of the proposed method. The results of the experiments are shown in Table 2. Where the table consists of comparison of modularity in Karate network, Dolphins network, Football network , and Polbooks network after applying the group of methods. The group of methods composes from K-means Modularity PSO, K-means Modularity Bat optimization, K-means Modularity CSO, K-median Modularity PSO, K-median Modularity Bat optimization, K-median Modularity CSO, GN [28], FN [28], BGLL [28], and HSCDA [28]. From this table, we can conclude that the modularity obtained by K-median Modularity CSO is better than that obtained by another methods except in case of dolphins network with HSCDA get better result. Table 3 illustrates the experimental results of NMI. It confirms that the proposed method (K-median Modularity CSO) outperforms the K-median Modularity PSO, and the K-median Modularity Bat optimization on four datasets.

Table 1: Parameters setting in the proposed method.

Data	Population Size	♯ of Iteration	♯ of Edge	♯ of Node	♯ of Community
Karate	30	100	78	34	4
Dolphins	40	120	159	62	4
Football	50	150	613	115	9
Polbooks	100	100	441	105	3

Table 2: Comparison of Modularity results

Data	Karate	Dolphins	Football	Polbooks
GN [28]	0.401[28]	0.519[28]	0.599[28]	0.516[28]
FN [28]	0.380[28]	0.489[28]	0.577[28]	0.502[28]
BGLL [28]	0.418[28]	0.518[28]	0.602[28]	0.498[28]
HSCDA [28]	0.419[28]	**0.527**[28]	0.602[28]	0.527[28]
PSObK-means	0.433	0.445	0.529	0.465
PSO K-median	0.442	0.461	0.566	0.480
BAT K-means	0.449	0.501	0.583	0.50
BAT K-median	0.470	0.472	0.614	0.51
CSO K-means	0.451	0.472	0.593	0.53
CSO K-median	**0.489**	0.502	**0.621**	**0.559**

Fig. 2 and Fig. 3 discuss the relation between modularity and community number for Karate network, Dolphins network, Football network, and Polbooks network respectively. It is obvious that when the community number is equal 4, the Karate network and Dolphins network get the maximum modularity. And when community number is equal 9, the Football network obtains the maximum modularity. Also when community number is equal 3, the Polbooks network obtains the maximum modularity.

As shown in Fig. 4, the proposed method (K-median Modularity CSO) has a quicker speed to converge at the best modularity Q. When the iteration number is equal to 100, Karate network and Polbooks network gain the best modularity Q. When the iteration number is equal to 120, Dolphins network obtains to the best modularity Q. When the iteration number is equal to 150, Football network gets the best modularity Q.

5 CONCLUSION

Community detection is great important in computer science, biology, physics and sociology to understanding of complicated systems. This problem is very arduous and not yet satisfactorily solved despite many methods have been proposed. The K-median Modularity CSO method is presented for finding community structure in social network. This method uses modularity measure as the fitness function in the optimization process by changing the

Table 3: Comparison of NMI results.

Data	Karate	Dolphins	Football	Polbooks
PSO K-median	0.61	0.0.50	0.79	0.52
BAT K-median	0.68	0.53	0.82	0.56
CSO K-median	**0.71**	**0.61**	**0.86**	**0.59**

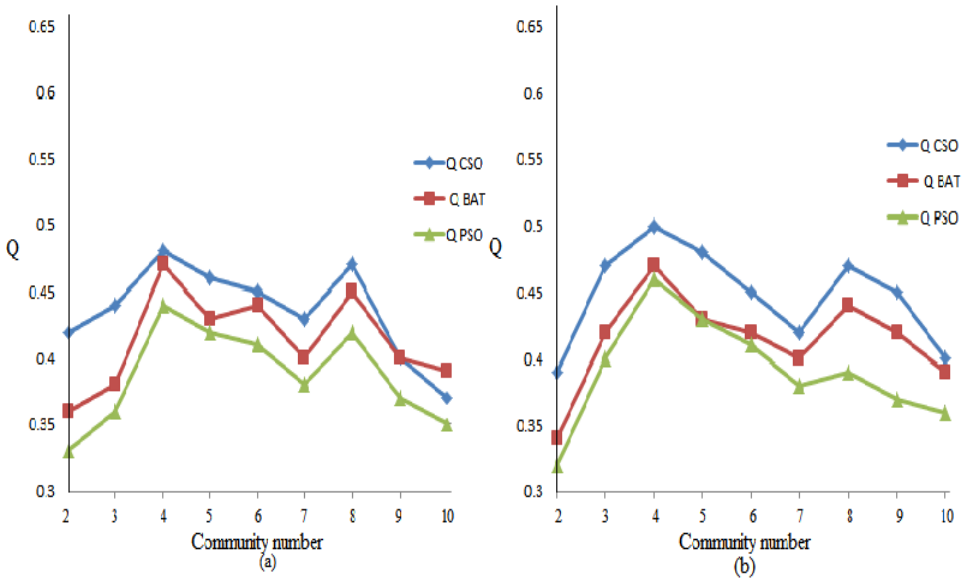

Fig. 2: Relation between modularity Q and Community number for Karate network and Dolphins network respectively.

cluster center. Experimental results show that the performance of The K-median Modularity CSO method is successfully finds an optimized community structure for different size of dataset and implementation of NMI measure on datasets confirmed this result. Besides, comparative the experimental results with different methods.

References

[1] D. Boyd, and N. Ellison, "Social Network Sites: Definition, History, and Scholarship.", Journal of Computer-Mediated Communication, vol. 13, pp. 210-230, 2008.

[2] Y. Chen, and X. Qiu, "Detecting Community Structures in Social Networks with Particle Swarm Optimization.", Springer Berlin Heidelberg, pp. 266-275, 2013.

[3] l. Danon, J. Duch, A. Diaz-Guilera, and A. Arenas, "Comparing community structure identification.", Journal of Statistical Mechanics, vol. 9, pp. 1-10, 2005.

[4] T. Alzahrani, and K. Horadam, "Community Detection in Bipartite Networks: Algorithms and Case studies.", Complex Systems and Networks, Springer Berlin Heidelberg, pp. 25-50, 2016.

[5] M. Girvan, and M. Newman, "Community structure in social and biological networks.", Proceedings of the National Academy of Sciences, vol. 99, pp. 7821-7826,

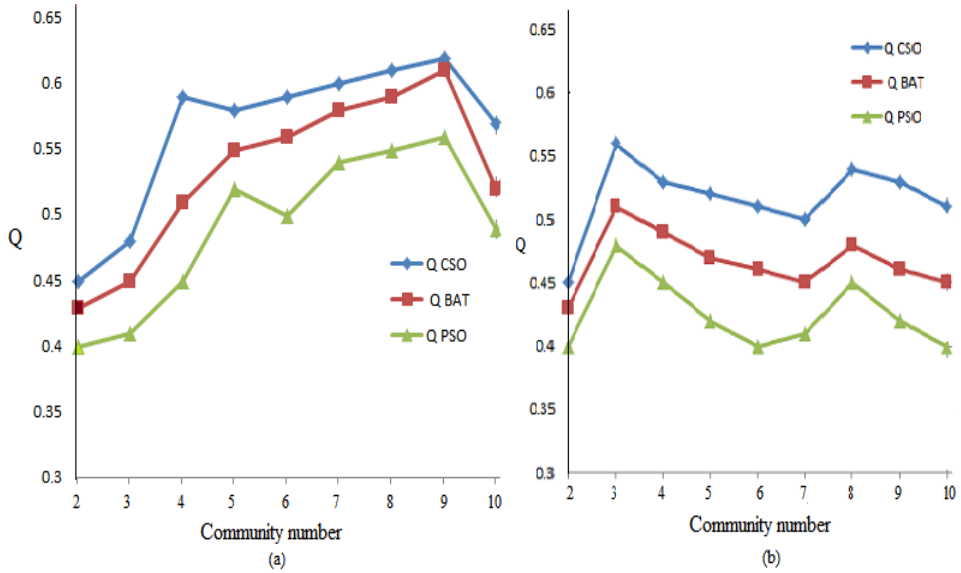

Fig. 3: Relation between modularity Q and Community number for Football network Polbooks network respectively..

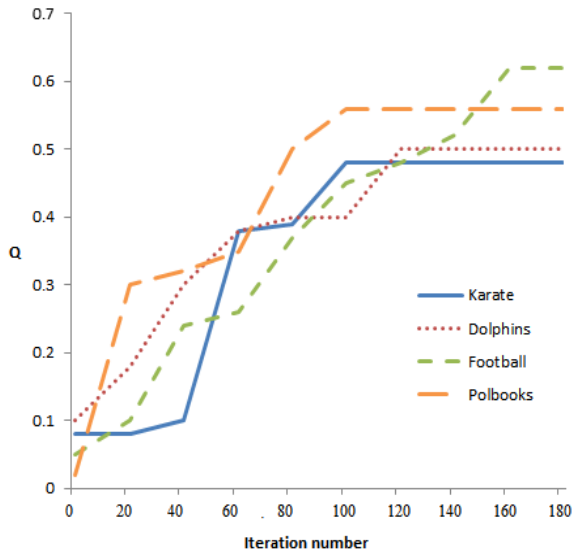

Fig. 4: Relationship between modularity Q and Iteration number in the proposed method.

2002.

[6] M. Newman, "Detecting community structure in networks.", The European Physical Journal B - Condensed Matter and Complex Systems, vol. 38, pp. 321-330, 2004.

[7] S. Nomura, S. Oyama, T. Hayamizu, and T. Ishida, "Analysis and improvement of HITS algorithm for detecting Web communities", Journal of Systems and Computers in Japan, vol. 35, pp. 32-42, 2004.

[8] M. Newman, "Spectral methods for network community detection and graph partitioning.", Journal of Physical Review E, vol. 88, pp. 1-11, 2013.

[9] M. Newman, and M. Girvan, "Finding and evaluating community structure in networks.", Physical Review, vol. 69, pp. 1-16, 2004.

[10] C. Pizzuti, "GA-Net: A Genetic Algorithm for Community Detection in Social Networks.", Springer Berlin Heidelberg, pp. 1081-1090, 2008.

[11] C. Honghao, F. Zuren, and R. Zhigang, "Community Detection Using Ant Colony Optimization.", IEEE Congress on Evolutionary Computation, pp. 3072-3078, 2013.

[12] Z. Masdarolomoor, R. Azmi, S. Aliakbary, and N. Riahi, "Finding Community Structure in Complex Networks Using Parallel Approach.", Embedded and Ubiquitous Computing (EUC), 2011 IFIP 9th International Conference, IEEE, vol. 10, pp. 474-479, 2011.

[13] A. Song, M. Li, X. Ding, w. Cao, and K. Pu, "Community Detection Using Discrete Bat Algorithm.", IAENG International Journal of Computer Science, vol. 43, pp. 1-7, 2016.

[14] Y. EL Barawy, L. EL Bakrawy, and N. Ghali, "K-means Clustering With Swarm Optimization For Social Network Community Detection.", Asian Journal of Mathematics and Computer Research, vol. 3, pp. 220-230, 2013.

[15] J. Yang, J. McAuley, and J. Leskovec, "Community Detection in Networks with Node Attributes.", Published in the proceedings of IEEE ICDM '13, vol. 10, pp. 1-10, 2014.

[16] N. Yahia, N. Saoud, and H. Ghezala, "Evaluating Community Detection Using a Bi-objective Optimization.", International Conference on Intelligent Computing, Springer Berlin Heidelberg, pp. 61-70, 2013.

[17] S. Chu, P. Tsai, and J. Pan, "Cat Swarm Optimization", Springer Berlin Heidelberg, pp. 854-858, 2006.

[18] S. Yousef, M. Khanesar, and M. Teshnehlab, "Discrete binary cat swarm optimization algorithm.", Control and Communication (IC4), 2013 3rd International Conference on. IEEE, pp. 1-6, 2013.

[19] K. Yugal, and G. Sahoo, "A hybrid data clustering approach based on improved cat swarm optimization and K-harmonic mean algorithm.", Journal of Information and Computing Science, vol. 9, pp. 196-209, 2014.

[20] S. Budi, and M. Ningrum, "Cat swarm optimization for clustering.", Soft Computing and Pattern Recognition, International Conference of. IEEE, pp. 54-59, 2009.

[21] P. Dalatu, "Time Complexity of K-Means and K-Medians Clustering Algorithms in Outliers Detection.", Global Journal of Pure and Applied Mathematics, vol. 12, pp. 4405-4418, 2016.

[22] H. Zhu, and Y. Shi, "Brain storm optimization algorithms with k-medians clustering

algorithms.", Advanced Computational Intelligence (ICACI), 2015 Seventh International Conference on, IEEE, pp. 107-110, 2015.

[23] C. Whelan, G. Harrell, and J. Wang, "Understanding the K-Medians Problem.", Proceedings of the International Conference on Scientific Computing (CSC), The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp. 219-222, 2015.

[24] http://konect.uni-koblenz.de/networks/ucidata-zachary, Accessed Novamber 2016.

[25] http://konect.uni-koblenz.de/networks/dolphins, Accessed Novamber 2016.

[26] http://www-personal.umich.edu/ mejn/netdata/, Accessed Novamber 2016.

[27] http://www.casos.cs.cmu.edu/computational tools/datasets/external/polbooks/index11.php, Accessed Novamber 2016.

[28] X. Bin, Q. Jin, Z. Chunxia, H. Xiaoxuan, X. Bianjia, S. Yanfei, "Hybrid Self-Adaptive Algorithm for Community Detection in Complex Networks", Hindawi Publishing Corporation, Mathematical Problems in Engineering, pp. 1-12, 2015.

[29] Y. Wang, J. Fang, and F. Wu, "Application of Community Detection Algorithm with Link Clustering in Inhibition of Social Network Worms.", International Journal of Network Security, vol. 19, pp. 458-468, 2017.