A Model Based Metaheuristic for Hybrid Hierarchical Community Structure in Social Networks

Radhia Toujani, Jalel Akaichi

International Science Index, Computer and Information Engineering Vol:11, No:6, 2017 waset.org/Publication/10007188

***Abstract*—In recent years, the study of community detection in social networks has received great attention. The hierarchical structure of the network leads to the emergence of the convergence to a locally optimal community structure. In this paper, we aim to avoid this local optimum in the introduced hybrid hierarchical method. To achieve this purpose, we present an objective function where we incorporate the value of structural and semantic similarity based modularity and a metaheuristic namely bees colonies algorithm to optimize our objective function on both hierarchical level divisive and agglomerative. In order to assess the efficiency and the accuracy of the introduced hybrid bee colony model, we perform an extensive experimental evaluation on both synthetic and real networks.

***Keywords*—Social network, graph partition, community detection, agglomerative hierarchical clustering, divisive hierarchical clustering, similarity, modularity, optimization, metaheuristic, bee colony.

I. Introduction and Related Work

SOCIAL networks usually exhibit a hierarchy of communities which requires the appearance of algorithms to detect these communities and focus on their hierarchical relationships. Most of the existing hierarchical algorithms proposed for communities detection are based on either agglomerative or divisive principle. In fact, agglomerative hierarchical algorithms start with one community per vertex in the network and keep agglomerating vertices together to form increasingly larger communities. Nevertheless, the divisive hierarchical algorithms start with a single community and split the network into sub-partitions according to some criteria [21]. In our work, community detection method proceeds by successive combinations of the aggregation and decomposition operators and ends when a fixed partition is obtained. However, this later partition generated the problem of convergence to a locally optimal detected community. In fact, because the objective improves with each move and at each hierarchical level, eventually a local optimum will be achieved. In this current work, we aim to obtain a globally optimal hierarchical community structure. To achieve this purpose, we integrated metaheuristic, more precisely Bee Colony Optimization, into the introduced hybrid hierarchical model through the proposal of an objective function. Therefore, we review community detection methods relying on optimization. Indeed, the main objective of the introduced approaches in literature is the integration of optimization issue to attain optimal value of fitness function [11]. Actually, modularity optimizing community detection algorithms aim at determining the partition having maximum modularity. Several algorithms were proposed to approximate a reliable and accurate $Qmax$ [1]. In addition, the network modularity, developed by Girvan and Newman [2], [4], is extensively applied as a quality metric to evaluate a specific network partitioning in communities. Therefore, finding the highest modularity value is considered as a NP-hard problem because the possible partitions space enlarges more rapidly than any power of the system size [4]. The four well-known categories of modularity-optimizing community detection algorithms are spectral, greedy, simulated annealing and external optimization methods.

To enhance modularity, Newman introduced the first greedy agglomerative algorithm [2]. It represents a hierarchical clustering technique in which edges are progressively added during the greedy procedure.

Annealing [5] is a probabilistic process applied in various problems and fields to obtain global optimization. It represents the possible states space searching the maximum global optimum of a function F. Simulated annealing for modularity optimization was first used by Guimera et al. [6]. Their standard implementation [7] combines the local moves, in which a single node moves randomly from one cluster to another, and the global moves which contain the communities mergers and splits.

Extremal optimization is a heuristic search procedure, was introduced by Boettcher and Percus [8]. This technique relies on the local variables optimization. Then, Duch and Arenas [9] applied this method to optimize modularity. The latter is a sum over the nodes in the graph. The fitness measure of each node can be obtained by dividing the node local modularity by its degree which does not determine the modality measure. Girvan and Newman (GN) developed a divisive method [4] containing the edges removal depending on the values of their betweenness. To obtain efficient time complexity and to get an optimized division, authors integrated the Network Modularity (Q) into the iterative removing of edges with the greatest betweenness value [3]. Afterwards, Radicchi suggested a similar approach with GN [18] by applying the coefficient of edge-clustering as the novel metric. Indeed, the approach time complexity is equal to $o(n^2)$ which is inferior to that of GN. Clauset et al. developed a fast clustering algorithm in order to enhance the computation efficiency [2]

R. Toujani is with the BestMod Department, University of Tunis, Higher Institute of Management, Tunisia (e-mail: toujaniradia@gmail.com).

J. Akaichi is with the Department of Information Systems, King Khaled University, College of Computer Science, Tunisia (e-mail: jalel.akaichi@ku.edu.sa).

International Science Index, Computer and Information Engineering Vol:11, No:6, 2017 waset.org/Publication/10007188

with $O(n \log n)$ time complexity on sparse graph which uses a greedy strategy to get a maximal ΔQ by merging pairs of nodes iteratively until it becomes negative. Pons and Latapy [14] developed a clustering algorithm relying on the random walk method in order to measure the similarity between vertices. Their algorithm, having$O(n \log n)$ time complexity, applies the Network Modularity (Q) to show the end of the agglomerative process. Other important algorithms include Duch and Arenas's extremal optimization approach introduced in [15] with $O(n \log n)$ time complexity, Clauset's method for finding local community structures in [16], the agent-based algorithm proposed by Gunes and Bingol in [13], as well as the approaches based on the information theoretic framework in [12], [20].

To optimize modularity, the spectral eigen matrix values and vectors were used. For instance, Wang et al. [17] utilized community vectors in order to attain high-modularity partitions into communities number inferior to a given maximum. If the eigen vectors, corresponding to the two largest eigen values, are considered, then the graph can be split into three clusters. To obtain graph tri-partitions with large modularity along these lines, Richardson et al. [18] introduced a faster technique. Obviously, all the afore-mentioned approaches, having various backgrounds and valid scopes, are efficiently applied in community detection. Nevertheless, because the new social networks represent large sparse graphs with considerable overlapping between vertices groups [10], [16], the betweenness-based divisive algorithms will have unimportant computational efficiency. However, the fast agglomerative approach [4] cannot generally give an acceptable division because of its local optimization strategy.

II. Proposal

In this section, we define and formalize the introduced optimization based method hybrid hierarchical community structure.

A. Formalization

Social network can be modelized by a graph $G = (V, E, O)$, where V represents the users in the network, E denotes the different interactions between them and O a set of shared opinions between users V during their navigation in social network.

B. Useful Functions

1) The coefficient of Jacquard [22] is an index of the set neighbors intersection, obtained without applying any semantic analysis of their meaning. It represents the ratio between the cardinal (size) of the intersection of the considered sets and the Cardinal of the union of sets. It also measures the similarity between these two sets. In our case, the basic idea of computing similarity is:
Given two opinions sets Op_i and Op_j (Op_i represents opinions of user (V_i) while Op_j denotes opinions of user (V_j)), we use the coefficient of jaccard to measure the semantic and the structural similarity where we replace Op_i by N_i representing the neighbor node of user (V_i) and Op_j by N_j denoting the neighbor node of user (V_j). In fact, we define the index for determining the semantic similarity as:

$$JS(Op_i, Op_j) = \frac{Op_i \cap Op_j}{Op_i \cup Op_j} \tag{1}$$

2) Similarity-based Modularity (Q_s) Function [23] focus on similarity measure into modularity to ensure a good quality of graph partiton. Thus, the similarity of vertices within a cluster is higher than the similarity of vertices between clusters. Similarity-based Modularity (Q_s) is described as:

$$Q_s = \sum_{i=1}^{NC} (\frac{IS_i}{TS} - (\frac{DS_i}{TS})^2) \tag{2}$$

$IS_i = \sum_{u,v \in V_i} S(u,v)$, $DS_i = \sum_{u \in V_i, v \in V} S(u,v)$ and
$TS = \sum_{u,v \in V} S(u,v)$. where NC is the number of clusters, IS_i denotes the total similarity of vertices within cluster i; DS_i represent the total similarity between vertices in cluster i and any vertices in the graph; TS is the total similarity between any two vertices in the graph; $S(u,v)$ denotes the used similarity, V is the vertex set of the graph and V_i is the vertex set of cluster i.

C. The Objective Function

Our objective function is based on the concepts of similarity based modularity and the coefficient of Jaccard outlined in the previous section. Indeed, we defined define an initial partition which will be injected as input to the introduced hybrid method and we modify the (Q_s) function by adding the average of the Jaccard coefficient for measuring the structural and the semantic similarity between two nodes. Hence, each hierarchical level has its appropriate objective function.

1. Fitness of Ascendant Hierarchical Level

In agglomerative hierarchical level, the fitness function is:

$$AscQ_{JS} = \max \sum_{i=1}^{NC} (\frac{I(JS)_i}{T(JS)} - (\frac{D(JS)_i}{T(JS)})^2) \tag{3}$$

2. Fitness of Descendant Hierarchical Level

However, for the divisive hierarchical level, the fitness function is described in (4):

$$DescQ_{JS} = \min \sum_{i=1}^{NC} (\frac{I(JS)_i}{T(JS)} - (\frac{D(JS)_i}{T(JS)})^2) \tag{4}$$

3. Fitness of Hybrid Hierarchical Process

Consequently, the fitness in the hybrid process combines $AscQ_{JS}$ and $DescQ_{JS}$.

$$HQ_{JS} = AscQ_{JS} o DescQ_{JS} \tag{5}$$

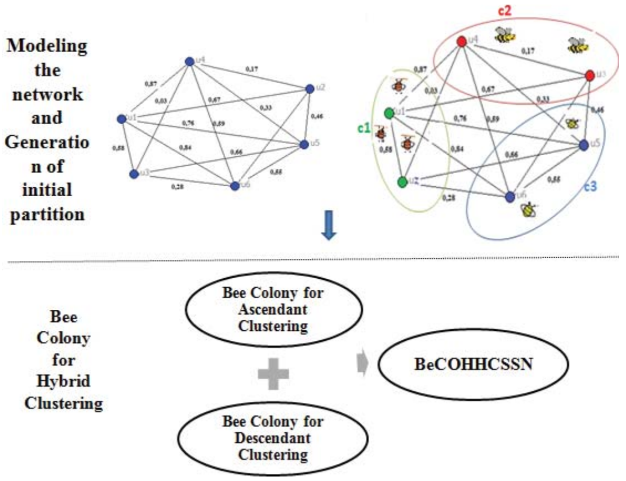

Modeling the network and Generation of initial partition
u1
u2
u3
u4
u5
u6
0,87
0,17
0,03
0,67
0,33
0,46
0,76
0,59
0,58
0,84
0,66
0,55
0,28
c1
c2
c3
Bee Colony for Hybrid Clustering
Bee Colony for Ascendant Clustering
Bee Colony for Descendant Clustering
BeCOHHCSSN

Fig. 1 Architecture of $BeCOHHCSSN$ Model

D. Bee Colony Optimization and Hybrid Clustering

Our model, called BEE Colony Optimization for Hybrid Hierarchical Community Structure in Social Network (BeCOHHCSSN) is used to optimize our objective functions that characterize the overall quality of a partitioning in both hierarchical levels.

As mentioned in Fig. 1, our model consists of two phases.

Phase 1: The first phase consist on building a weighted graph modeling the networks. In fact, we associate to each node a set of opinions shared between vertex using sentiment analysis method described in [24]. Then, we weighted every edges by the value of JS defined in (1). Then, we elaborate from this later network an initial solution Ccomposed of k sub-detected groups which will be considered as the starting point of our hybrid hierarchical clustering.

Phase 2: At this stage, the introduced metaheuristic process is launched to optimize objective functions in both hierarchical level. In fact, Bees Colony Optimization (BCO) is the used metaheuristic. It inspired by the natural foraging behavior of honey bees to find the optimal solution [25]. Each colony of honey bees spread in long distances and in multiple directions simultaneously to exploit a large number of food sources (flower). This optimization algorithm require an initialization procedure and a search for promising flower patches is iterated or until a higher quality of fitness is found. In fact, the proposed optimal hierarchical community detection approach is assimilated as bee colony optimization issue. The first introduced optimal hierarchical algorithm is the Ascendant Bee Colony algorithm ($AscBC$) relying on the aggregation operator. In fact, at each iteration, $AscBC$ merges the two communities having higher fitness. However, the second method is Descendant Bee Colony algorithm ($DescBC$)based on removing edges having less objective function value. Finally, the hybrid algorithm combines $AscBC$ and $DescBC$. Because this later algorithm focus on both maximizing the fitness on the aggregation process and minimizing the objective function on the decomposition operator, it is considered as a multi-objective optimization issue [19].

1) $AscBC$ algorithm: We summarize the steps of the introduced iterative $AscBC$ algorithm in the primordial stages described in Algorithm 1:

Algorithm 1 Ascendant Bee Colony Algorithm

```
Require: Input: graph G(V,E)
Ensure: Output: k sub-detected Colony Community
 1: $C = \{\{u_1\}, \{u_2\}, ..., \{u_n\}\}$
 2: Put the Queen of Bee in user u having highest connections
 3: while aggregation procedure is no longer feasible. do
 4:     repeat
 5:         Select u for neighborhood search.
 6:         Evaluate Fitness
 7:         Select the fittest bee from each patch.
 8:         Inform all bees by the change of structure
 9:         Put another colony to the next important non-visited node
10:     until (Every bees colony constructs its members)
11: end while
```

In first step, the $AscBC$ algorithm considers that each social network user constitutes a community and scout bees in the search space which is formed by sub-detected communities. Then, We put the queen of Colony on the most important user having highest connections. After that, our iterative process is lunched. Whether the aggregation procedure, based on merging the two communities having the highest fitness is feasible, we repeat these steps:

- An artificial bee visits another user for neighborhood search (line5)
- It evaluates the the fitness. In fact, for the $AscBC$ algorithm the objective function denotes the maximum $AscQ_s$ value.(line6)
- If this artificial bee found user ensuring higher fitness and decides to put it in its colony community, it acknowledges the queen which informs all bees by the change of structure. For communication within the colony, every bee uses a substance called pheromone to help the colony to send its bees to flower patches precisely. Hence, the essential information for colony communication are the direction in which it will be found, its distance from the hive and its fitness.(line 7 et 8)
- The bee with the highest fitness will be selected to form the next bee population (line9)
- We repeat the same steps until every bees colony constructs its members (line10)

2) $DescBC$ algorithm: The descendant bee colony algorithm is similar to the previously-described one. However, it is proceeded by an opposite hierarchical construction. It initially considers that all social network users constitute a community and begins with a partition containing a single community. The principle of this algorithm is to decompose sub-detected groups having the least fitness obtained through the introduced bee colony process (see Algorithm 2).

Algorithm 2 Descendant Bee Colony Algorithm

```
Require: Input: graph G(V,E)
Ensure: Output: k sub-detected Colony Community
 1: $C = \{u_1, u_2, u_3, ..., u_n\}$
 2: Put the Queen of Bee in user $u$ having least connections
 3: while Bursting procedure is feasible. do
 4:     repeat
 5:         Select $u$ for neighborhood search.
 6:         Evaluate Fitness
 7:         Select the fittest bee from each patch.
 8:         Informs all bees by the change of structure
 9:         Put another colony to the next important non-visited node
10:     until (Every bees colony constructs its members)
11: end while
```

In contrast to $AscBC$ algorithm, artificial bee separate, from its colony community, sub-detected groups having less fitness. After moving colony to the next less important non-visited node, we repeat the same steps (lines 5, 6, 7, 8) until every bees colony constructs its members.

3) The Hybrid Hierarchal Bee Colony Algorithm ($HHBCA$) exploits alternatively the two previously-mentioned algorithms and it can be summarized in these steps:

- $HHBCA$ requires the existence of an initial solution which can be defined by either the introduced $AscBC$ or $DescBC$.
- It proceeds by a successive combination of the introduced optimal decomposition and aggregation operators. In fact, these later operators are applied to the sub-detected colony community obtained through $AscBC$ or $DescBC$.
- The algorithm stops if detected colony community applying aggregation operator is the same one of obtained through $DescBC$

Thus, referring to the process of stabilization, we should apply $AscBC$o$DescBC$ or $DescBC$o$AscBC$ in a regular order until getting fixed optimum detected groups.

Cc_k denotes obtained Colony community at hierarchical level k (see Algorithm **??**).

Algorithm 3 $HHBCA$

```
Require: input: Graph G(V,E)
Ensure: output: k sub-detected colony communities
repeat
    repeat
        $Cc_k$ =AscBCoDescBC($Cc_k$).
    until ($AscBC$o$DescBC$ ($Cc_k$ ))=$Cc_k$
    repeat
        $Cc_k$ =DescBCoAscBC($Cc_k$).
    until ($DescBC$ o$AscBC$($Cc_k$ ))= $Cc_k$
until ($AscBC$o$DescBC$ ( $Cc_k$))=($DescBC$ o$AscBC$($Cc_k$))= $Cc_k$
```

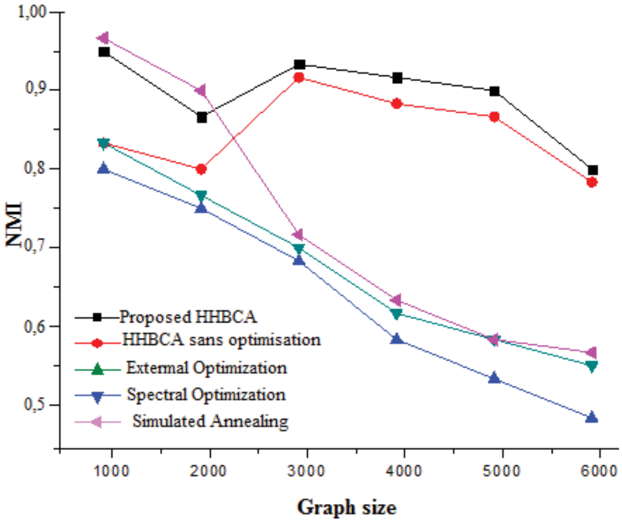


Fig. 2 Comparison of the clustering quality in terms of NMI for artificial network

III. Simulation Results

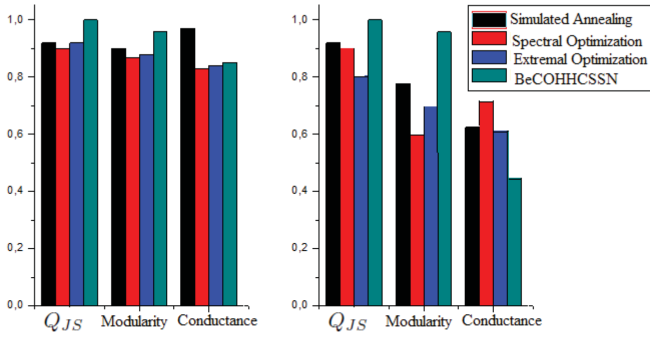
A. Evaluation on Artificial Networks

Comparing the computed partitions to the real structure of a network is the best way to evaluate the performance of different algorithms. Thus, to validate and test our model, we exploited randomly-generated graph using the LFR benchmarks [26]. In fact, to evaluate the efficiency of our method on this benchmark, we used the Normalized Mutual Information NMI to compare the computed partitions and the exact partitions of the network. In fact, the NMI is defined in this equation [27]:

$$NMI(A,B) = \frac{-2\sum_{a\in A}\sum_{b\in B}|a\cap b|\log(\frac{|a\cap b|n}{|a||b|})}{\sum_{a\in A}|a|\log(\frac{|a|}{n}) + \sum_{b\in B}|b|\log(\frac{|b|}{n})} \qquad (6)$$

where A is the real partitions of the network and B represents the partition obtained by the used algorithm. In fact, $NMI(A,B) = 1$ when both partitions A and B coincide and higher values are better.

As indicated in Fig. 2, in addition to the comparison of the introduced $HHBCA$ with its version without the use of optimization process, we compare the performance (in terms of NMI values) of the proposed model with the methods described in the literature namely Simulated annealing [5], Spectral Optimization [17] and Extremal Optimization [8] for different graph size. We notice that, although without integrating optimization process, the introduced $HHBCA$ displays better clustering quality. Furthermore, for a graph with 4000 nodes, $HHBCA$ version without the use of optimization outperforms the quality of Simulated annealing, Spectral Optimization and Extremal Optimization algorithms. Nevertheless, the use of metaheuristic namely Bee Colony Optimization in our clustering issue has a significant impact and lead to the generation of good results and higher clustering quality even in complex graph size. For example, $NMI = 0,87$ for a graph with 6000 nodes. Obviously, we notice that our $BeCOHHCSSN$ model performs almost perfectly (with NMI¿0.8) and generally outperforms the quality of the others algorithms, with the exception when the graph size is less than 2000 nodes, Simulated Annealing Algorithm has the best quality. Overall, we see that $BeCOHHCSSN$ displays higher clustering quality.



(a) Quality measures for Karate dataset

(b)Quality measures for football dataset

Fig. 3 Quality measures for the Real Networks

B. Evaluation on Real Networks

Secondly, we performed evaluations on other type of benchmark based on networks with known community structure. Moreover, we chose two networks previously considered in literature namely:

1) **karate**: Network of friendship relations between members of a US university karate club, known in literature as Zachary karate club [28]. This graph is well known and often used as a benchmark for community detection algorithms. The club consisted of 34 members and after internal disagreements it broke up in two groups.
2) **football**: Network of American football games between Division IA colleges during regular season Fall 2000 [4]. There are 115 teams, corresponding vertices, pairs of which are connected by an edge if they played each other. All teams are separated into 12 conferences. Conferences offer a natural community structure, as teams from one conference play more often one another than teams from a different conference.

Figs. 3 (a) and 3 (b) depict the quality measures for the karate and football datasets.

In addition to our objective function, we choose to evaluate the results with various evaluation criterions namely the conductance which measures, for a cluster, the ratio of internal to external connectivity with lower values indicating better clustering quality and the modularity which shows how separated are the different vertex types from each other with higher values indicating good graph partitioning. Generally, we notice that our model outperforms the quality of the simulated annealing, spectral optimization and extremal optimization algorithms with maximum values of modularity function, minimum values of conductance function and an important value of our objective function.

IV. Conclusion

In this paper, we are interested in the issue of hierarchical community detection in social networks. The considered hybrid hierarchical process combines the aggregation and decomposition operators until obtaining a fixed partition which generated the problem of a local optimum. Thus, the main contribution of our method is obtaining a globally optimal hierarchical community structure. For this reasons, we integrate metaheuristic, more precisely Bee Colony Optimization, into the introduced hybrid hierarchical model through the proposal of an objective function which measures the modularity of the semantics in the structural similarity for both hierarchical levels. On one hand, for the agglomerative process, our objective function consists on aggregate social network users having higher similarity based modularity value. On the other hand, for the divisive process we aim to decompose users with lower similarity based modularity value. In fact, the developed function measuring the similarities between social network users is based on common opinions to construct a community. In a future work, we will try to detect opinion leaders in each community, identify influential users and track the evolution of the communities structure.

References

[1] Benjamin H. Good, Y. A., and Aaron C. *Performance of modularity maximization in practical contexts.* Phys. Rev. E, 81:046106, Apr 2010.

[2] Clauset, M. E. J. Newman, and C. Moore *Finding community structure in very large networks.* , In Phys. Rev. E 70, 066111, 2004.

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

[5] Fortunato, S., *Physics Reports, 486(3-5), pp. 75 - 174.* 2010.

[6] Guimera, R., M. Sales-Pardo, and L. A. N. Amaral *Phys. Rev. E 70(2), 025101 (R)* 2004.

[7] Guimera, R., and L. A. N. Amaral *Nature 433, 895.* 2005.

[8] Boettcher, S., and A. G. Percus *Phys. Rev. Lett. 86, 5211,* 2001.

[9] Duch, J., and A. Arenas, *Phys. Rev. E 72(2), 027104,* 2005.

[10] Qi G. J., Aggarwal C. C., and Huang T., *Community Detection with Edge Content in Social Media Networks* IEEE 28th International Conference on Data Engineering, 2012.

[11] Chira C., Gog A., and Iclanzan D. *Evolutionary Detection of Community Structures in Complex Networks: a New Fitness Function* WCCI 2012 IEEE World Congress on Computational Intelligence, Brisbane, Australia, June, 10-15, 2012.

[12] Kulathumani, A. Arora, Sridharan M., and Demirbas M., *Trail: A Distance-Sensitive Sensor Network Service for Distributed Object Tracking* ACM Transactions on Sensor Networks, vol. 5, no. 2, article 15, pp. 140, Mar. 2009.

[13] Pothen H. *Path Selection for Social Network Evolution Map Formation of Start-up Enterprises* Technical Report, Norfolk, VA, USA, 1997.

[14] Huatao P. *Partitioning Algorithms with Applications to Scientific Computing* International Conference on Computer and Communication Technologies in Agriculture Engineering, IEEE, pp. 4750, 2010.

[15] Duch, J., and Arenas A., *Physical Review E, 72(2), 027104.* 2005.

[16] Danon, L., A. Daz-Guilera, J. Duch, and A. Arenas *Physical Review E, 72(2), 027104.* 2005.

[17] Wang, G., Shen Y., and Ouyang M., *Comput. Math. Appl. 55(12), 2746,* 2008.

[18] Richardson, T., Mucha P. J., and Porter M. A. *Phys. Rev. E 80(3), 036111* , 2009.

[19] Ehrgott M. and Gandibleux X. *Multiobjective combinatorial optimization* , In M. Ehrgott and X. Gandibleux, editors, Multiple Criteria Optimization State of the Art Annotated Bibliographic Surveys, volume 52, pages 369,444. Kluwer Academic Publishers, Boston, MA, 2002.

[20] Leskovec, J., Lang, K. J., and Mahoney, M. *Empirical comparison of algorithms for network community detection* , In WWW, 631640, 2010.

[21] Fasmer E. E. *Community Detection in Social Networks* Thesis University of Bergin,April2015.

[22] David L. and Jon K. *The link-prediction problem for social networks.* Journal of the American society for information science and technology, 58(7): 10191031, 2007.

[23] Feng, Xiaowei, Nurcan, Y., and Thomas A. J., S. *A Novel Similarity-based Modularity Function for Graph Partitioning* Proceeding DaWaK'07 Proceedings of the 9th international conference on Data Warehousing and Knowledge Discovery, Pages 385-396, Regensburg, Germany September 03 - 07, 2007.

[24] Toujani R. and Akaichi J. *Machine Learning and Metaheuristic For sentiment anal- ysis in social networks* Metaheuristic Internatianal Conference MIC'15, Morrocco, 2015.

[25] Sagayam R. and Akilandeswari K. *Comparison of Ant Colony and Bee Colony Optimization for Spam Host Detection* International Journal of Engineering Research and Development eISSN: 2278-067X, pISSN: 2278-800X, www.ijerd.com Volume 4, Issue 8, PP. 26-32, November 2012.

[26] San Fortunato*Benchmark graphs to test community detection algorithms.* http://sites.google.com/site/santofortunato/inthepress2, Date of access to the site: February 2016.

[27] Christos, G., Fragkiskos D. M., Dimitrios M. T. and Michalis, V. *CORECLUSTER: A Degeneracy Based Graph Clustering Framework*. Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[28] W. W. Zachary. *An information flow model for conflict and fission in small groups.* Journal of Anthropological Research, 33:452473, 1977.