Community Detection in Complex Networks based on Multiobjective Honey Bee Mating Optimization

Bahak Amiri, Liaquat Hossain and John W Crawford
The University of Sydney, Sydney, NSW, Australia

Abstract—Detecting community structure is crucial for uncovering the links between structures and functions in complex networks. Most of contemporary community detection algorithms employ single optimization criteria (e.g., modularity), which may have fundamental disadvantages. This paper considers the community detection process as a Multi-Objective optimization Problem (MOP). To solve the community detection problem this study used improved honey bee mating optimization (HBMO) algorithm. In the proposed algorithm, an external repository is considered to save non-dominated (Pareto) solutions found during the search process. The efficiency of the proposed algorithm is studied by testing on several data sets. Numerical results show that the proposed evolutionary optimization algorithm is robust and suitable for community detection problem.

Keywords—complex network; community; multi-objective; honey bee mating optimization

I. INTRODUCTION

Most of real word networks possess inherent community structure, such as biological networks, web graphs and social networks. In network a community is a group of nodes with high dens connection within groups and sparse connection between groups. Communities in networks provide us information about how network function and topology affect each other[1].

In complex networks the number of communities is typically unknown and the communities are often of unequal size or density, and it has been shown that in complex networks communities have a hierarchical structure so we can say that finding communities in complex networks is a non trivial task [2]. Community detection problem has been introduced formally in 2002 [3] and recently has been attracted the attention of researches in deferent areas. The community detection problem can be considered almost like an optimization problem [4] and lots of studies have been done based on evolutionary methods like GA [5-9], SA [10] and collaborative evolutionary algorithms[11] and [12] to solve it. To consider community detection problem as an optimization problem we need an objective function to be improved like modularity Q that is used as the stopping criterion in GN [3]. Most of community detection problems are based on the single objective optimization and their differences are based on their objective functions. The single objective based community detection algorithms have some shortcomings such as: single-objective optimization algorithms attempt to optimize just one criterion, they may fail when the optimization criteria are inappropriate and also most of them require prior information like the number of communities, which is usually unknown for real networks.

To overcome these shortcomings the community detection problem can be considered as a multi-objective problem, so multiple objective functions can be considered to obtain more accurate and comprehensive community structure. In this paper we considered community detection problem as a multi-objective optimization problem and introduced an improved multi-objective algorithm based on honey bee mating optimization algorithm.

One of the recently proposed evolutionary algorithms that have shown great potential and good perspective for the solution of various optimization problems is honey bee mating optimization (HBMO). The HBMO algorithm has remarkable accuracy and calculation speed to deal with the optimization problem. Advantages of the HBMO algorithm are presented in [13, 14]. Refs. [15] and [14] have used the HBMO algorithm for solving optimization problems on two separate applications. In this paper, a multiobjective optimization is used for the placement and sizing of REGs by the improved HBMO algorithm. Original HBMO often converges to local optima. In order to avoid this shortcoming, in this paper a new mating process is proposed for rising accuracy of the algorithm. The proposed algorithm optimized two objective functions, the community score that measures the density of the clusters obtained and community fitness that minimizes the external links. A prior knowledge of a number of communities was not needed because this method returns a set of solutions where each of them correspond to different trade-offs between the two objectives, and gives a great chance to analyze the hierarchy of communities. The rest of the paper is organized as follows. Section 2 introduced the problem of community detection. The concept of a multi-objective optimization problem is reviewed in section 3. The original honey bee mating optimization algorithm is explained in section 4. In section 5 the proposed multiobjective algorithm used to detect community is presented, and then in section 6 the experimental results of the proposed algorithm in comparison with other approaches are shown.

II. COMMUNITY DETECTION
Community detection has been studied in many fields for many years such as computer science, physics and biology and lots of methods have been introduced in this field. In this study an undirected network \( G = (V, E) \) defined by a set of nodes [13] or vertices, and a set of links [16] connect two elements of \( V \). A community consists of vertices and an edge between these nodes, where the nodes often cluster into tightly knit groups with a high density and a lower density of between the group connections [17].

A network can be represented mathematically by an adjacency matrix \( A \), if there is an edge from \( v_i \) to \( v_j \), \( A_{ij} = 1 \) and \( A_{ij} = 0 \) otherwise. The degree \( k_i \) of a node \( i \), defined as \( k_i = \sum_j A_{ij} \). Let \( C \subset G \) the sub-graph where node \( i \) belongs to, the degree of \( i \) with respect to \( C \) can be split as \( k_i(C) = k_i^{in}(C) + k_i^{out}(C) \), where \( k_i^{in}(C) = \sum_{j \in C} A_{ij} \) is the number of edges connect the \( i \) to the other nodes in \( C \), and \( k_i^{out}(C) = \sum_{j \in C} A_{ij} \) is the number of edges connecting \( i \) to the rest of the network.

A sub-graph \( C \) is a community in a strong sense if \( k_i^{in}(C) > k_i^{out}(C) \), \( \forall i \in C \). A sub-graph \( C \) is a community if \( \sum_{i \in C} k_i^{in}(C) > \sum_{i \in C} k_i^{out}(C) \). The quality measure of a community \( C \) that maximizes the in-degree of the nodes belonging to \( C \) has been introduced in [8]. On the other hand, in [18] a criterion that minimizes the out-degree of a community is defined. We now recall the definitions of these measures first, and then show how they can be exploited in a multi-objective approach to find communities. In the following, without losing its generality, the network graph is assumed to be undirected.

Let \( \mu_i \) denote the fraction of edges connecting node \( i \) to the other nodes in \( C \). More formally, \( \mu_i = \frac{1}{|C|} k_i^{in}(C) \) where \( |C| \) is the cardinality of \( C \).

The power mean of \( C \) of order \( r \), denoted as \( M(C) \) is defined as

\[
M(C) = \frac{\sum_{i \in C} (\mu_i)^r}{|C|} \quad (1)
\]

Notice that in the calculation of \( M(C) \), since \( 0 \leq \mu \leq 1 \), the exponent \( r \) increases the weight of nodes having many connections with other nodes belonging to the same module, and diminishes the weight of those nodes having few connections inside \( C \).

The volume \( v_C \) of a community \( C \) is defined as the number of edges connecting vertices inside \( C \), i.e. the number of 1 entries in the adjacency sub-matrix of \( A \) corresponding to \( C \), \( v_C = \sum_{i \in C} \sum_{j \in C} A_{ij} \).

The score of \( C \) is defined as \( \text{score}(C) = M(C) \times v_C \). Thus, the score takes into account both the fraction of interconnections among the nodes (through the power mean) and the number of interconnections contained in the module \( C \) (through the volume). The community score of a clustering \( \{C_1, \ldots, C_k\} \) of a network is defined as

\[
CS = \sum_{i=1}^{k} \text{score}(C_i) \quad (2)
\]

The community score gives a global measure of the network division in communities by summing up the local score of each module found. The problem of community detection has been formulated in [8] as the problem of maximizing \( CS \).

In [18] the concept of community fitness of a module \( C \) is defined as

\[
P(C) = \sum_{i \in C} \left( k_i^{in}(C) + k_i^{out}(C) \right)^\alpha \quad (3)
\]

where \( k_i^{in}(C) \) and \( k_i^{out}(C) \) are the internal and external degrees of the nodes belonging to the community \( C \), and \( \alpha \) is a positive real-valued parameter controlling the size of the communities. The community fitness has been used by [18] to find communities.

### III. MULTI-OBJECTIVE OPTIMISATION

In a multi-objective optimization problem, the purpose is to optimize several conflicting objectives simultaneously while still meeting some constraints. The community detection problem algorithm can be formulated as a multi-objective problem because two objective functions are competing with each other. The first is maximizing the inter-connecting links and the second is minimizing connections between communities. The multi-objective problem can be described as follows [19]:

\[
min F = [f_1(X), f_2(X), \ldots, f_n(X)] \quad (4)
\]

where \( f_i(X) \) is the \( i \)-th objective function and \( X \) is the vector of the optimization variables, \( n \) is the number of objective functions.

The solution to the multi-objective optimisation problem is a set of Pareto points. In the multi-objective optimization problem, a solution \( X^* \in \Omega \) is a Pareto optimal if there is no solution \( X \in \Omega \) such that \( X \) dominates \( X^* \). \( \Omega \) is the set of all feasible values of \( X \).

The solution \( X_i \) is said to dominate the solution \( X_j \) if \( \forall j \in \{1, 2, \ldots, n\}, f_j(X_i) \leq f_j(X_j) \) \( \exists k \in \{1, 2, \ldots, n\}, f_k(X_i) < f_k(X_j) \) (5)

Solutions which dominate others but not themselves, are called non-dominated solutions.

### IV. HONEY BEE MATTING OPTIMIZATION

ALGORITHM

In order to deal with multiobjective problems, some modifications in the HBMO algorithm should be made. After the generation of the initial population and respective evaluation of objective functions, the selection of the “best solutions” (queens) should be made, but no longer based only on the comparison of single objective function values. Under a multiobjective approach, a new concept, such as the Pareto dominance concept, is needed for dealing with different solutions, i.e. classifying them as dominated or non-dominated solutions. The “best solutions” (queens) selected from the initial population are the non-dominated solutions.

Once identified the non-dominated solutions (queens), the iterative process is initiated in the same way as in the single objective case (mating flights, generation of new queens, improvement of the queens and of the new generation and selection of new queens). Each non-dominated solution will generate a certain number of solutions after each iteration. The criteria for generation...
and improvement of the solutions specially best solution (queen) are the same as employed in the uni-objective version. With the new generated solutions and the non-dominated solutions from the previous iteration, the new set of non-dominated solutions is identified, which forms the Pareto front. These new solutions are saved in the repository and will generate the new solutions in the next iteration. The process is repeated until the stop criterion is satisfied. Frequently, the number of solutions that belong to the Pareto front increases as the algorithm evolves, thus each non-dominated solution is a potential generator (queen) of new solutions in the next iteration of the algorithm. It is noted that these non-dominated solutions saved in the repository are not the final non-dominated solutions because the repository will be updated after generation of broods in the next iterations. Besides, our experiences in the implementation of the proposed algorithm shows that it is safe to say the repository of the non-dominated solutions will be significantly updated in each new iteration with respect to the previous ones in the initial iterations. However, after some iteration the results of the repository may be saturated. That is the non-dominated solutions may be remained unchanged. Indeed the new solutions in the higher number of the iterations may be equal to the same of the repository solutions or they will be dominated by the repository ones. Therefore, it can be concluded from this manner that the non-dominated solutions of the repository after some iteration are trustworthy.

V. ROPOSED ALGORITHM

Previously in this work the search capability of the honey bee mating optimization algorithm was specifically used to find communities in complex networks. The steps of the proposed community detection algorithm as shown in figure 1 are as follows:

Step 1: Initializing the problem and algorithm parameters
In this phase, as described above, we are interested in identifying a partitioning \( \{C_1, \ldots, C_k\} \) that maximizes the number of connections inside each community and minimizes the number of links between the modules. The first objective was fulfilled by the community score. The first objective function is therefore \( CS = \sum_{i=1}^{k} \text{score} (C_i) \). The second objective was carried out by the community fitness by summing up the fitness of all the \( C_i \) modules. The parameter \( \alpha \), that tunes the size of the communities, has been set to 1 because in most cases the partitioning found for this value are relevant [18]. The second objective is therefore \( \sum_{i=1}^{k} P (C_i) \).

Our partitioning algorithm uses the locus-based adjacency representation proposed in [20] and used by [20] for multi-objective clustering. In this graphic representation, an individual of the population consists of \( N \) variable \( x_1, \ldots, x_N \) and for each variable there is a set of possible range of values based on the adjacency matrix. For example, if node 1 has a connection with nodes 3, 5, and 6, the possible range of values for \( x_1 \) will be \{3, 5, and 6\}. For the isolated node \( k \) in the network, the possible range of values can be \( \{1, 2, \ldots, k-1, k+1, \ldots, N\} \).

Variables and values represent nodes of the graph \( G = (V, E) \) modeling a network \( N \), and a value \( j \) assigned to the \( i^{th} \) variable is interpreted as a link between the nodes \( i \) and \( j \) of \( V \). This means that in the clustering solution found, \( i \) and \( j \) will be in the same cluster. However, a decoding step is needed to identify all the components of the corresponding graph. The nodes participating with the same component are assigned to one cluster. As observed in [20], the decoding step can be done in linear time. The main advantage of this representation is that the number of communities will be automatically determined by the number of components contained in an individual, and will be determined by the decoding step.

Step 2: Generate initial population

Step 3: Begin with \( i = 1 \).

Step 2: Pick randomly two candidates for selection \( X_1 \) and \( X_2 \).

Step 3: Pick randomly a comparison set of individuals from the population.

Step 4: Compare each candidate, \( X_1 \) and \( X_2 \), against each individual in the comparison set for domination using the conditions for domination given in Eqs. (4) and (5).

Step 5: If one candidate is dominated by the comparison set while the other is not, then select the later for reproduction and go to Step 7, else proceed to step 6.

Step 6: If neither or both candidates are dominated by the comparison set, then use sharing to choose winner.

Step 7: If the criteria \( i = N \) is reached, stop selection procedure, else set \( i = i + 1 \) and go to Step 2.

VI. Experimental results

In this section effectiveness of the proposed multi-objective honey bee mating optimization algorithm (MHBMO) has been compared with Clauset, Newman and Moore (CNM) [21] and MOGA-Net [8] using some real world datasets and synthetic benchmark datasets. The effectiveness of stochastic algorithms is greatly dependent on the generation of initial solutions and therefore, for every dataset, algorithms have individually performed 100 times to test their own effectiveness, and each time with randomly generated initial solutions. Our algorithm was implemented into Matlab 7.1. All the experiments were conducted on a computer with Intel Core 2 Duo, 2.66 GHz, 4 GB RAM.

A. Evaluation Criteria

To evaluate the quality of the proposed community detection method we used Normalized Mutual Information (NMI) and Modularity (Q). The Normalized Mutual Information (NMI) is a similarity measure proven by Danon et al [22] to be reliable. Given two partitions \( A \) and \( B \) of a network in communities, let \( C \) be the confusion matrix whose element \( C_{ij} \) is the number of nodes of community \( i \) of the partition \( A \) that are also in the community \( j \) of the partition \( B \). The normalized mutual information \( I(A, B) \) is defined as:

\[
I(A, B) = \sum_{i,j} \frac{C_{ij}}{\min(n_i, n_j)} \cdot \log \left( \frac{\min(n_i, n_j)}{n_{ij}} \right)
\]
\[ I(A, B) = \frac{-2 \sum_{i=1}^{C_A} \sum_{j=1}^{C_B} C_{ij} \log \left( \frac{C_{ij} N}{C_i C_j} \right)}{\sum_{i=1}^{C_A} C_i \log \left( \frac{C_i}{N} \right) + \sum_{j=1}^{C_B} C_j \log \left( \frac{C_j}{N} \right)} \]  

where \( C_i \) (or \( C_j \)) is the number of groups in the partition \( A \) (or \( B \)), \( C_{ij} \) is the sum of the elements of \( C \) in row \( i \) (column \( j \)), and \( N \) is the number of nodes. If \( A = B \), \( I(A, B) = 0 \). The modularity of Newman and Girvan [23] is a well known quality function used to evaluate the goodness of a partition. Let \( k \) be the number of modules found inside a network, the modularity is defined as:

\[ Q = \sum_{s=1}^{k} \left[ \frac{l_s}{m} - \frac{(d_s)^2}{2m^2} \right] \]  

where \( l_s \) is the total number of edges joining vertices inside the module \( s \), and \( d_s \) is the sum of the degrees of the nodes of \( s \). The first term of each summand of the modularity \( Q \) is the fraction of edges inside a community and the second one is the expected value of the fraction of edges that would be in the network if they fell at random without regard to the community structure. Values approaching \( I \) indicate a strong community structure.

B. Real World Networks

The Zachary’s Karate Club network was generated by Zachary, who studied the friendship of 34 members of a karate club over a period of two years [24]. During this period, because of disagreements, the club divided into two groups almost of the same size.

The Bottlenose Dolphins network: A network of 62 bottlenose dolphins living in Doubtful Sound, New Zealand, was compiled by Lusseau after studying their behavior for seven years. A tie between two dolphins was established by their statistically significant frequent association. The network split naturally into two large groups where the number of ties was 159 [25].

The American College Football network: comes from the United States college football. The network represents the schedule of Division I games during the 2000 season. Nodes in the graph represent teams and edges represent the regular season games between the two teams they connect. The teams are divided into conferences. On average the teams played 4 inter-conference matches and 7 intra-conference matches, thus they tend to play between members of the same conference. The network consists of 115 nodes and 616 edges grouped in 12 teams [3].

The political books compiled by V. Krebs: The nodes represent 105 books on American politics brought from Amazon.com, and the edges join pairs of books frequently purchased by the same buyer. Books were divided by Newman [28] according to their political alignment (conservative or liberal), except for a small number (13) having no clear affiliation. The e-print Arxiv: initiated in Aug 1991, has become the primary mode of research communication in multiple fields of physics, and some related disciplines.

It is a network of 9000 scientific paper and their citations (9000 nodes and 24000 links) [29]. The webpage network: were obtained from the complete map of the nd.edu domain, which contains 325,729 documents and 1,469,680 links [26].

C. Results

A comparison of results for running different algorithms on each real world dataset that mentioned in last section is illustrated in Tables 1–8.

Table 1. Modularity result obtained by the three algorithms on Zachary’s Karate Club data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.4161</td>
<td>0.4161</td>
<td>0.4161</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.3811</td>
<td>0.3708</td>
<td>0.3621</td>
<td>0.0100</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.4151</td>
<td>0.4149</td>
<td>0.4148</td>
<td>0.0010</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Modularity result obtained by the three algorithms on Bottlenose Dolphins data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.5233</td>
<td>0.5233</td>
<td>0.5233</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.4605</td>
<td>0.4486</td>
<td>0.4367</td>
<td>0.0110</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.5048</td>
<td>0.5038</td>
<td>0.5029</td>
<td>0.0090</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Modularity result obtained by the three algorithms on American College Football data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.5812</td>
<td>0.5804</td>
<td>0.5793</td>
<td>0.0010</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.5433</td>
<td>0.5188</td>
<td>0.5046</td>
<td>0.0237</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.5148</td>
<td>0.4978</td>
<td>0.4784</td>
<td>0.0158</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Modularity result obtained by the three algorithms on Krebs’ books on American politics data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.5162</td>
<td>0.5162</td>
<td>0.5162</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.4934</td>
<td>0.4715</td>
<td>0.4522</td>
<td>0.0186</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.5176</td>
<td>0.5136</td>
<td>0.5075</td>
<td>0.0039</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Modularity result obtained by the three algorithms on Arxiv data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.7854</td>
<td>0.7811</td>
<td>0.7776</td>
<td>0.0042</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.7721</td>
<td>0.7415</td>
<td>0.7112</td>
<td>0.0304</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.7911</td>
<td>0.7226</td>
<td>0.7743</td>
<td>0.0083</td>
<td></td>
</tr>
</tbody>
</table>

Table 6. Modularity result obtained by the three algorithms on Web nd.edu data

<table>
<thead>
<tr>
<th>Method</th>
<th>Modularity</th>
<th>Best</th>
<th>Average</th>
<th>Worst</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHBMO</td>
<td>0.9286</td>
<td>0.9274</td>
<td>0.9260</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>CNM</td>
<td>0.9274</td>
<td>0.8852</td>
<td>0.8501</td>
<td>0.0411</td>
<td></td>
</tr>
<tr>
<td>MOGA-Net</td>
<td>0.9304</td>
<td>0.9187</td>
<td>0.9073</td>
<td>0.0116</td>
<td></td>
</tr>
</tbody>
</table>

For Zachary’s Karate Club data the MHBMO algorithm provided a value of 0.4161 in all runs, but on the other hand the CNM and MOGA-Net algorithms attained 0.3811 and 0.4151 respectively, as shown in Table 1. The MHBMO found 4 communities for this dataset in all runs. For Bottlenose Dolphins data (Table 2) the
MHBMO algorithm attained the values of 0.5233 for modularity in all runs and four communities were detected by the MHBMO. The best modularity values provided by MOGA-Net and CNM were 0.4605 and 0.5048. The MHBMO algorithm detected 11 communities for the American College Football data (Table 3) attained the best value of 0.5812 for modularity. The CNM and MOGA-Net algorithms provided the best values of 0.5433 and 0.5148 in terms of modularity. The Community detection problem, Krebs’ books on the results of the American Politics data given in Table 4 shows that the MHBMO provided an optimum value of 0.5162 for modularity. For the community detection problem, MHBMO algorithm detected 11 communities for the American College Football data. The MHBMO algorithm provided by MOGA-Net and CNM were 0.4605 and 0.7854 and 0.9286 for modularity. The MHBMO algorithm was much more stable than the other algorithms, as can be observed from Tables 1 to 6. The results illustrate that the proposed MHBMO community detection approach can be considered as a viable and an efficient heuristic to find optimal or near optimal solutions to the problem of community detection in networks.

VII. CONCLUSION
This paper presents a multiobjective community detection algorithm based on the improved honey bee mating optimization. The proposed algorithm has several advantages compared to other optimization techniques in that it does not require a complex calculus, thus it is free from divergence and there is no need to set initial values for the decision variables. The proposed algorithm for community detection can be used when the number of clusters is unknown a priori. To evaluate the performance of the proposed algorithm, it was compared with the MOGA-Net and CNM algorithms. The algorithm was implemented and tested on several real world datasets, and showed that it was quite efficient at discovering the community structure of complex networks. Thus, this proposed algorithm can be considered as a viable and an efficient heuristic to find the optimal or near optimal solutions to clustering problems.

REFERENCES


