Detecting Hierarchical and Overlapping Community Structures in Social Networks Using a One-Stage Memetic Algorithm

Chun-Cheng Lin[1], Der-Jiunn Deng[2](✉), Jung-Chao Wu[1], and Liang-Yi Lu[1]

[1] Department of Industrial Engineering and Management,
National Chiao Tung University, Hsinchu 300, Taiwan
cclin321@nctu.edu.tw, wu7419658@gmail.com

[2] Department of Computer Science and Information Engineering,
National Changhua University of Education, Changhua 500, Taiwan
djdeng@cc.ncue.edu.tw

Abstract. Detection of hierarchical and overlapping community structures for social networks is crucial in social network analysis. Previous strategies were focused on a two-stage strategy for separately detecting hierarchical and overlapping community structures. This paper develops a one-stage memetic algorithm for concurrently detecting hierarchical and overlapping community structures in social networks, where quality evaluation functions, community capacity, and hierarchical levels are taken into account to increase the solution quality. This algorithm includes a local search scheme to improve the solution searching ability. Through simulation, this algorithm shows pleasing quality.

Keywords: Hierarchical and overlapping community structure
Social network · Memetic algorithm

1 Introduction

Since the members with similar backgrounds and interests in social networks have frequent interactions with each other, they constitute a community. A social network with multiple communities is a community structure. Generally, the community structures of social networks to be investigated popularly are *overlapping community structures* and *hierarchical community structures*. Previous methods on detecting overlapping and hierarchical community structures were based on a two-stage strategy [1]. However, most previous works did not consider a one-stage strategy for this problem. Therefore, this work proposes a one-stage memetic algorithm (MA) for detecting overlapping and hierarchical community structures in social networks, while constraints of community size and hierarchical levels are considered. To design a one-stage method, this work proposes a novel optimization function of maximizing the quality of jointly overlapping and hierarchical community structures, in which the quality of overlapping community structures is based on the density function of links in each community (called the D value) [2]; and the quality of hierarchical community

B. Li et al. (Eds.): ChinaCom 2017, LNICST 237, pp. 182–188, 2018.
https://doi.org/10.1007/978-3-319-78139-6_19

structures is based on the EQ value proposed in [1]. Since the concerned community detection problem has been shown to be NP-hard [3], this work proposes a novel memetic algorithm (MA) for the problem, which is a genetic algorithm (GA) incorporated with local search scheme.

2 Literature Review

In detecting hierarchical community structures, Chen et al. [5] computed a similarity matrix that computes similarity of all adjacent vertices, and then merged vertices to form a hierarchical community structure according to the similarity matrix. Then, they applied the community density D [2] as the threshold value of determining the community structure at each hierarchical level. Shen et al. [1] proposed a so-called EAGLE algorithm to detect overlapping and hierarchical community structures, which first finds all maximal cliques in the input social network and lets them be overlapping communities; and then computes similarity of each pair of communities, and merge the communities with higher similarity to form a binary hierarchical structure. To obtain a good-quality community structure, they propose an *EQ* value to evaluate quality of overlapping community structure at each hierarchical level. They found the level with the community with the largest *EQ* value among all hierarchical levels, and let the level as the number of levels of the hierarchical structure.

3 The Concerned Problem

Consider to represent a social network topology as a graph G. The notations used for representing the input graph are summarized as follows. $G = (V, E)$: The social network is represented as graph G in which $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$ represent set of vertices and edges, respectively, in which v_i represents a vertex; e_i represents an edge. The number of vertices in graph G is n; and the number of edges in graph G is m.

The notations of the output of the problem are summarized as follows. Number of hierarchical levels of the hierarchical community structure of graph G is h. $H = \{H^1, H^2, \ldots, H^h\}$ represent the output overlapping and hierarchical community structure of graph G. $H^k = \{C_1^k, C_2^k, \ldots\}$ represent set of communities at the k-th hierarchical level, in which C_i^k denotes the i-th community at the k-th hierarchical level. This work allows $C_i^k \cap C_j^k \neq \emptyset$. $\left|C_i^k\right| = n_i^k$ denotes the number of vertices in the i-th community C_i^k at the k-th hierarchical level.

With the above notations, the concerned problem is described in detail as follows. Given the graph topology G, the problem is to detect $H = \{H^1, H^2, \ldots, H^h\}$ where $H^k = \{C_1^k, C_2^k, \ldots\}$ for graph G with the objectives: Maximize the D value and the *EQ* value i. Subject to the following constraints: $n_i^k \geq n_{threshold}$, $\forall k \in \{1, 2, \ldots, h\}$, $\forall i \in \{1, 2, \ldots, |H^k|\}$ where $n_{threshold}$ is a given lower bound of size of each community i at each hierarchical level k. $h_{threshold} \geq h$ where $h_{threshold}$ is a given upper bound of number of levels of the hierarchical community structure.

4 The Proposed MA

The proposed MA is detailed in Algorithm 1, whose main components are explained in detail as follows.

Solution Encoding and Population Initialization. This work is referred to the work [3] based on adjacency of links, which is more convenient to detect overlapping edges and vertices. In Algorithm 1, a population of η chromosomes are initialized randomly.

Algorithm 1 THE PROPOSED MA

```
Initialize the initial population $P_0$ and evaluate fitness of each chromosome in the population
Let the generation number $k$ be 0
while $k$ is less than the maximal generation number $\kappa$ do
    Conduct binary tournament selection to select $p_c \cdot \eta$ chromosomes
    Conduct one-point crossover on each pair of the $p_c \cdot \eta$ selected chromosomes to generate the offspring population $Q_k$
    Conduct local search on the offspring population $Q_k$
    Conduct mutation with mutation rate $p_m$ on the offspring population $Q_k$
    Conduct local search on the offspring population $Q_k$
    Evaluate fitness of each chromosome in the offspring population $Q_k$
    Replace the worst chromosomes in the current population $P_k$ by the offspring population $Q_k$
    Iteration number $k \leftarrow k + 1$
end while
Output the community structure corresponding to the chromosome with the best fitness
```

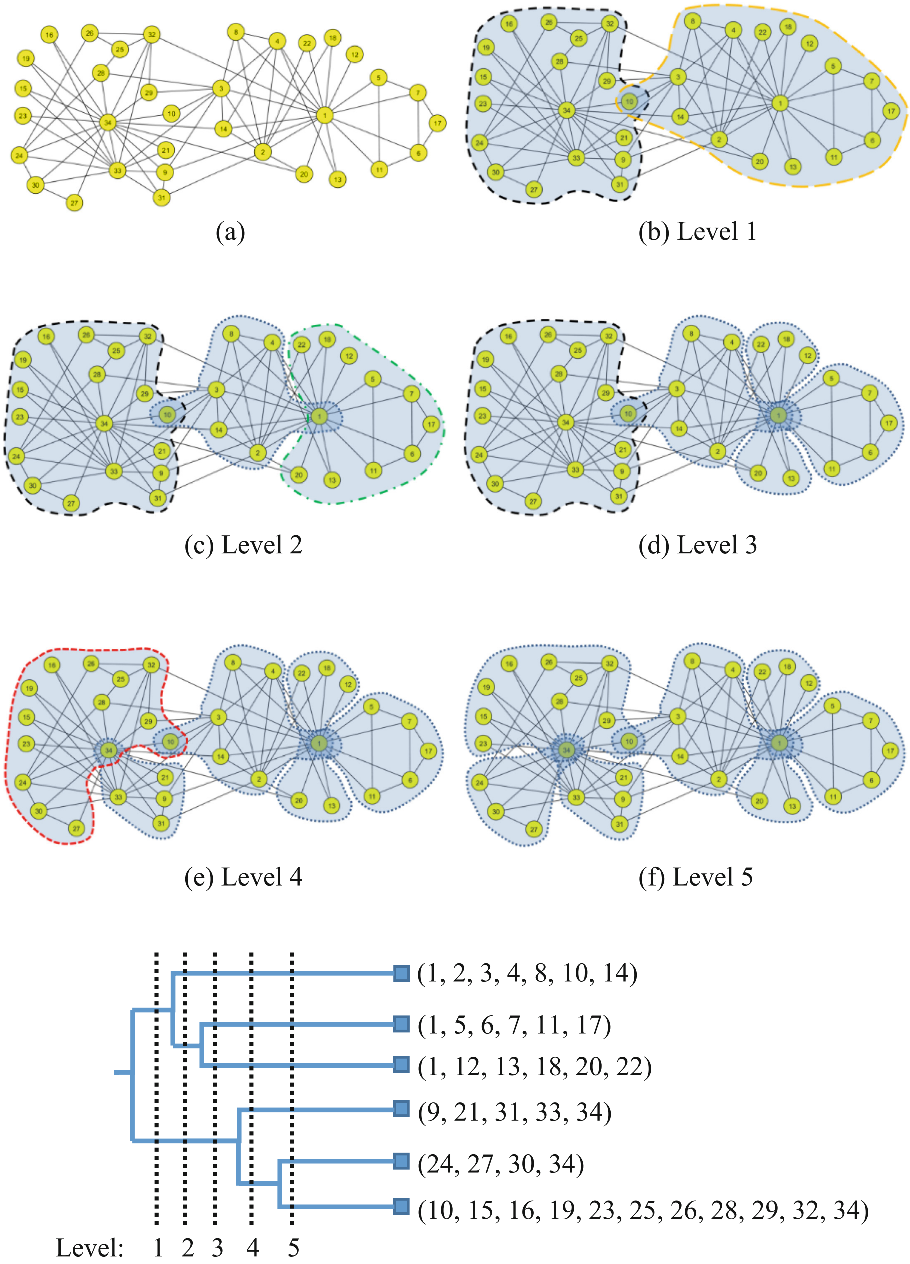
Fitness Evaluation. The concerned problem is to maximize both the D value and the EQ value. Therefore, after the overlapping and hierarchical community structure is decoded in the last subsection in which D and EQ_{max} values can be obtained, the fitness value corresponded to the chromosome instance is evaluated as the following weighted sum of D and EQ values:

$$Fitness = \lambda D + (1 - \lambda)EQ_{\max} \tag{1}$$

where λ is a weight falling within the range [0,1].

GA Operators. The conventional GA operators include parent selection, crossover, and mutation. The parent selection operator is the binary tournament selection. The crossover operator is one-point crossover. A mutation operator randomly selects a chromosome from the offspring population, then randomly selects a gene g_i from the chromosome, and then mutates gene g_i within the feasible range of g_i, which includes all possible adjacent edges of edge e_i.

Local Search. A local search is to randomly select a chromosome, then randomly select a gene from this chromosome based on roulette wheel selection, and then mutate the gene within its feasible range. The roulette wheel selection selects a gene based on the selection probability of each gene proportional to the number of feasible values.



(a)

(b) Level 1

(c) Level 2

(d) Level 3

(e) Level 4

(f) Level 5

(g)

Fig. 1. The overlapping and hierarchical community structure of the network for Zachary's karate club detected by the proposed method. (a) The network structure. (b)–(f) Overlapping community structures at each hierarchical level. (g) The hierarchical tree corresponding to the overlapping and hierarchical community structure.

5 Experimental Results

The experiment is conducted on a karate club network (Fig. 1, with 34 vertices and 78 edges) [6]. After lots of experimental trials, the setting of experimental parameters in the proposed MA is given as follows: $n_{threshold} = 4$; $h_{threshold} = 4$; number of iterations = 2000; weight λ in the fitness function = 0.3; number of chromosomes $\eta = 20$; crossover rate $p_c = 0.5$; mutation rate $p_m = 0.01$; number of local search at each iteration $l_{round} = 20$; number of genes in a chromosome by local search $l_{num} = 2$.

This subsection compares the convergence performance of the proposed MA and the original GA on this social network instances in Fig. 2. It is obvious from Fig. 2 that the proposed MA have better performance in convergence in this instance.

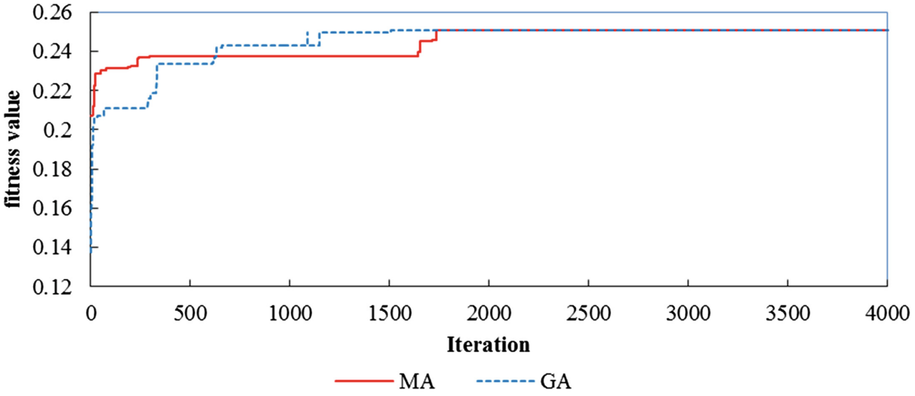


Fig. 2. Convergence comparison of the proposed MA and the original GA.

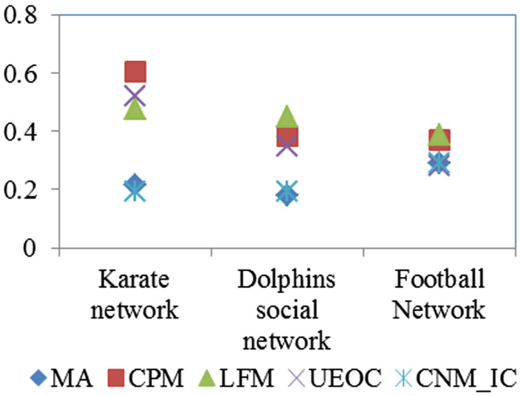


Fig. 3. Performance of the proposed MA and the other approaches for *AC* value (noting that a smaller *AC* value is better).

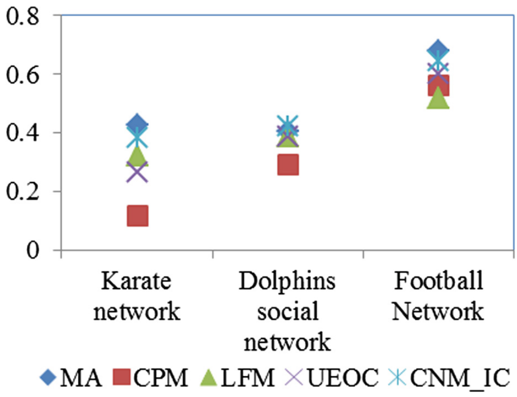


Fig. 4. Performance of the proposed MA and the other approaches for *EQ* value (noting that a larger *EQ* value is better).

Generally, the community structure cannot be realized from the network topology. Hence, some criteria are required to evaluate the performance of the results generated by different approaches. This section adopts two evaluation indices commonly. The first index is the average conductance function (AC) [4]. If the *AC* value is smaller, the community structure has a better quality. The second index is the *EQ* value. If the *EQ* value is greater, the quality of the overlapping community structure is better.

The two evaluation indices of the experimental results of the proposed MA and four previous approaches for three well-known network instances are compared in Figs. 3 and 4, in which the *EQ* value of the proposed MA is EQ_{max}. From Fig. 3, the proposed MA has better *AC* values than CPM, LFM, and UEOC in all network instances. From Fig. 4, the proposed MA has better *EQ* values than the other four approaches in three network instances.

6 Conclusion

This work has proposed a one-stage MA for detecting overlapping and hierarchical community structures in social networks which considers the constraints of community size and hierarchical level, both of which were not considered in previous works. The proposed MA includes a local search scheme to maximize two objectives *D* and EQ_{max}. Experimental results on three real social network instances show that the proposed MA performs better than the previous approaches.

References

1. Shen, H., Cheng, X., Cai, K., Hu, M.B.: Detect overlapping and hierarchical community structure in networks. Phys. A Stat. Mech. Appl. **388**(8), 1706–1712 (2009)
2. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)

3. Cai, Y., Shi, C., Dong, Y., Ke, Q., Wu, B.: A novel genetic algorithm for overlapping community detection. In: Tang, J., King, I., Chen, L., Wang, J. (eds.) ADMA 2011. LNCS (LNAI), vol. 7120, pp. 97–108. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-25853-4_8
4. Leskovec, J., Lang, K.J., Mahoney, M.: Empirical comparison of algorithms for network community detection. In: Proceedings of the 19th International World Wide Web Conference (WWW 2010), pp. 631–640 (2010)
5. Chen, J., Saad, Y.: Dense subgraph extraction with application to community detection. IEEE Trans. Knowl. Data Eng. **24**(7), 1216–1230 (2012)
6. Zachary, W.W.: An information flow model for conflict and fission in small groups. J. Anthropol. Res. **33**(4), 452–473 (1977)