Multi-objective Overlapping Community Detection by Global and Local Approaches

Darian H. Grass-Boada[1(✉)], Airel Pérez-Suárez[1], Andrés Gago-Alonso[1], Rafael Bello[2], and Alejandro Rosete[3]

[1] Advanced Technologies Application Center (CENATAV), Havana, Cuba
{dgrass,asuarez,agago}@cenatav.co.cu
[2] Department of Computer Science,
Universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba
rbellop@uclv.edu.cu
[3] Facultad de Ingeniería Informática, Universidad Tecnológica de la Habana
"José Antonio Echeverría" (CUJAE), Havana, Cuba
rosete@ceis.cujae.edu.cu

Abstract. Overlapping community detection on social networks has received a lot of attention nowadays and it has been recently addressed as Multi-objective Optimization Evolutionary Algorithms. In this paper, we introduce a new algorithm, named MOGLAOC, which is based on the Pareto-dominance based MOEAs and combines global and local approaches for discovering overlapping communities. The experimental evaluation over four classical real-life networks showed that our proposal is promising and effective for overlapping community detection in social networks.

Keywords: Social network analysis
Overlapping community detection
Multi-objective Optimization Evolutionary Algorithm

1 Introduction

The analysis of complex networks has been received a lot of attention nowadays due to its applications on several contexts which include bioinformatics, sociology and security, among others. Several data mining techniques have been applied in order to extract knowledge from social networks, specifically, the detection of communities plays an important role in the analysis of these networks [1]. This technique aims to organise the nodes of a network in groups or communities such that nodes belonging to the same community are densely interconnected but spare connected with the remaining nodes in the network [2].

Taking into account the NP-hard nature of the community detection problem, most reported approaches use heuristics in order to search for a set of nodes that optimises an objective function which captures the intuition of community [2]. Consequently, most of community detection algorithms focused on solving

M. Mendoza and S. Velastín (Eds.): CIARP 2017, LNCS 10657, pp. 272–280, 2018.
https://doi.org/10.1007/978-3-319-75193-1_33

a single-objective optimization problem; however, these approaches face some difficulties: (1) the optimization of only one function confines the solution to a particular community structure, (2) many of them may fail due to the *resolution limit problem* [3], and (3) returning one single partition may not be suitable when the network has many potential structures. To overcome the aforementioned problems, many community detection algorithms model the problem as a Multi-objective optimization problem, and specifically, they used Multi-objective Optimization Evolutionary Algorithms (MOEAs) to solve it.

Unfortunately, most of reported MOEAs focused on discovering disjoint communities, although according to Palla et al. in [4], most real-world networks have overlapping community structure. To the best of our knowledge, only the MOEAs proposed in [1,5–8] addressed the overlapping community detection problem.

The MEA_CDPs algorithm [1] uses an undirected representation of the solution and the classical NSGA-II framework with the reverse operator, in order to search for the solutions optimising three objective functions. On the other hand, iMEA_CDPs [6] uses the same representation as MEA_CDPs but it uses other objective functions. Besides, iMEA_CDPs proposes to employ the NSGA-II or the MOEA/D as the optimization framework, together with the PMX and simple mutation operators. IMOQPSO [7] uses a center-based representation of the solution together with a combination of QPSO and HSA optimization frameworks, in order to find a set of nodes that optimises two previously defined objective functions. OMO [5] employs a representation based on adjacencies between edges of the network together with two objective functions and the NSGA-II framework. Finally, MCMOEA [8] detects first the set of maximal cliques of the network and then, it builds the maximal-clique graph. Starting from this transformation, MCMOEA uses a representation based on labels and the MOEA/D framework in order to detect the communities optimizing two objective functions.

In this paper, we propose a new algorithm based on Pareto-dominance based MOEAs [12] which combines global and local approaches for discovering overlapping communities (MOGLAOC). Our proposal starts by detecting a set of seeds which are then used to build the overlapping communities. With this aim we introduced in the classical Pareto-dominance based MOEAs framework an *expansion*, *improving* and *merging* steps which allow our proposal to detect overlapping zones in the network, to improve the overlapping quality of these zones, and to merge communities having a high overlapping. The experimental evaluation of our proposal over four classical real-world social networks showed that it is promising and effective for overlapping community detection.

The remainder of this paper is organized as follow: in Sect. 2, we introduce the MOGLAOC algorithm. The experimental evaluation, showing the performance of our proposed algorithm, over four real-life networks is presented in Sect. 3. Finally, conclusions and future work are presented in Sect. 4.

2 The MOGLAOC Algorithm

Let $G = \langle V, E \rangle$ be a given network, where V is the set of vertices and E the set of edges among the vertices. A multi-objective community detection problem aims to search for a partition P^* of G such that:

$$F(P^*) = min_{P \in \Omega} \left(f_1(P), f_2(P), \ldots, f_r(P)\right), \quad (1)$$

where P is a partition of G, Ω is the set of feasible partitions, r is the number of objective functions, f_i is the ith objective function and $min(\cdot)$ is the minimum value. With the introduction of the multiple objective functions there is usually no absolute optimal solution, thus, the goal is to find a set of *Pareto* optimal solutions [2].

A commonly used way to solve a multi-objective community detection problem is by using MOEAs. The general Pareto-dominance based MOEAs framework consists on the following four steps: (a) to generate an initial population of chromosomes (i.e., solutions to the problem at hand), taking into account a predefined representation, (b) to apply the evolutionary operators over the current population in order to build the next generation and to move through the solution space, (c) to evaluate the current and new populations by using a predefined set of objective functions, and (d) to apply a predefined heuristic for keeping and improving the best solutions found so far. Usually, steps b, c and d are repeated a predefined number of times or until a specific stop criterium is fulfilled.

The contributions our proposal introduces to the Pareto-dominance based MOEAs framework for overlapping community detection are focused on the inclusion of an *expansion*, *improving* and *merging* steps, which in turn are inserted between the above mentioned steps c and d. Following, in Sect. 2.1 we briefly describe how our proposal addresses the general steps of the MOEAs framework and then, Sect. 2.2 describes in details the *expansion*, *improving* and *merging* steps our proposal introduces.

2.1 General Steps of MOGLAOC

The main idea of our proposal is to use the steps of classic Pareto-dominance based MOEAs framework in order to detect a set of disjoint seed clusters, where each seed cluster represents the set of objects that a community should not share with any other community.

In order to detect these seed clusters we use the PESA-II [9] as the optimization mechanism. Taking this into account, we use the locus-based adjacency graph encoding [10] for representing each chromosome of the initial population generated at step a; the decoding of a chromosome requires the identification of all connected components, which in turn will be our seed clusters. For generating a chromosome the ith genotype composing the chromosome is built by randomly selecting a neighbour of node i. The uniform two-point crossover operator [10] is selected for crossover and for mutation, some genes are randomly selected and substituted by other randomly selected adjacent nodes [11].

For evaluating the quality of a set of seeds clusters S we employ the *intra* and *inter* objective functions proposed in [2], which measure the intra-link and inter-link strength of S, respectively. These functions are defined as follows:

$$Intra(S) = 1 - \sum_{S_i \in S} \frac{|E(S_i)|}{m} \qquad Inter(S) = \sum_{S_i \in S} \left(\frac{\sum_{v \in S_i} |N(v)|}{2 \cdot m}\right)^2, \quad (2)$$

where $E(S_i)$ is the number of edges inside seed S_i, m is the total number of edges in the network, and $N(v)$ is the set of adjacent vertices of vertex v. In order to address the step (d) we use the mechanism PESA-II includes for keeping and improving the best seed clusters found so far.

2.2 Expansion, Improving and Merging Steps

Let $S = \{S_1, S_2, \ldots, S_k\}$ be the set of disjoint seed clusters represented by a chromosome of the current population. Overlapping vertices are supposed to be those vertices that belong to more than one community and in order to be correctly located inside a community they need to have edges with vertices in those communities. The *expansion* step aims to detect these vertices through a greedy randomise local search procedure (GRASP) over each S_i.

For detecting the zones containing overlapping vertices we soften the initial criterium used for building the seeds. With this aim each seed cluster S_i is processed for determining which vertices outside the community share a significant number of their adjacent vertices with the community; that is, the *potential* overlapping vertices.

Let $S_i \in S$ be a seed cluster and $\partial S_i \subseteq S_i$ the set of vertices of S_i having neighbours outside S_i. The strength of ∂S_i is denoted as $Str(\partial S_i)$ and it is computed as the ratio between the number of edges the vertices of ∂S_i have with vertices inside S_i, and the number of edges the vertices of ∂S_i have with vertices inside and outside S_i. The greater the value of $Str(\partial S_i)$ the greater the number of inner edges ∂S_i has and consequently, the better ∂S_i is. A vertex $u \notin S_i$ is considered a *candidate* to be included in S_i iff u is adjacent to at least one vertex in ∂S_i and $Str(\partial S_i^{'}) - Str(\partial S_i) > 0$, where $S_i^{'} = S_i \cup \{v\}$.

In order to iteratively expand a seed cluster S_i, the following steps are performed: (1) determining the set L of vertices which are *candidate* to be included in S_i, (2) applying the roulette wheel selection method over the set L, where the probability of being selected that a vertex $v \in L$ has is computed by using the increase v produces in $Str(\partial S_i)$, and (3) repeating steps 1 and 2 while $L \neq \emptyset$.

Once the *expansion* step finished, the *improving* step is performed in order to locally improve each overlapping zone detected. From our point of view, any overlapping vertex is expected to have adjacent vertices belonging to different communities and possibly, belonging to different overlapping zones. In this paper, we propose to measure the overlapping quality of a vertex belonging to an overlapping zone by using two properties we call *uniformity* and *simple betweenness*.

Let Z be an overlapping zone detected and $C_Z = \{C_1, C_2, \ldots, C_m\}$ the set of communities that set up Z. Let $v \in Z$ be an overlapping vertex. Let $N_{C_Z}(v)$ be the set of adjacent vertices of v that belong to at least one community in C_Z. Let $G_v = \{G_v^1, G_v^2, \ldots, G_v^l\}$ be the set of communities or overlapping zones containing the vertices in $N_{C_Z}(v)$. A property we will expect v satisfies is to have the vertices in $N_{C_Z}(v)$ equally distributed over the groups of G_v. The *uniformity* of v, denoted as $U(v)$, measures how much the distribution of vertices in $N_{C_Z}(v)$ deviates from the expected distribution of $N_{C_Z}(v)$ and it is computed as follows:

$$U(v) = 1 - \sum_{G_v^i \in G_v} abs\left(\frac{\left|N_{C_Z}(v) \cap G_v^i\right|}{|N_{C_Z}(v)|} - \frac{1}{|G_v|}\right), \tag{3}$$

where $abs(\cdot)$ is the absolute value. $U(v)$ takes values in $[0, 1]$ and the higher its value the better well-balanced v is.

Let $N'_{C_Z}(v)$ be the set of adjacent vertices of v that belong to at most one community in C_Z. Another property we would expect an overlapping vertex $v \in Z$ to have is to be an *intermediary* between any pair of its adjacent vertices in $N'_{C_Z}(v)$; that is, the shortest path connecting any pair of vertices $u, w \in N'_{C_Z}(v)$ should be the path made of the undirected edges (u, v) and (v, w). The *simple betweenness* of v, denoted as $SB(v)$, measures how much intermediary v is and it is computed as follows:

$$SB(v) = \frac{2 \cdot \sum_{i=1}^{|C_Z|-1} \sum_{j>i}^{|C_Z|} \left(1 - \frac{|E(C_i, C_j)|}{\left|N'_{C_Z}(v) \cap C_i\right| \cdot \left|N'_{C_Z}(v) \cap C_j\right|}\right)}{|C_Z| \cdot (|C_Z| - 1)} \tag{4}$$

where $E(C_i, C_j) = \left\{(u, w) \in E \mid u, w \in N'_{C_Z}(v) \wedge u \in C_i \wedge w \in C_j\right\}$ is the set of edges between vertices in $N'_{C_Z}(v)$, with one vertex in C_i and the other one in C_j. $SB(v)$ takes values in $[0, 1]$ and the higher its value the best intermediary v is.

We would like to highlight that both the uniformity and simple betweenness concepts can be straightforward generalised in order to be applied to an overlapping zone. Let $U_{ave}(Z)$ be the initial average uniformity of the vertices belonging to an overlapping zone Z. A vertex $v \in Z$ is a candidate to be removed from Z iff $U(v) < U_{ave}(Z)$. On the other hand, a vertex $u \in N(v|C_Z), v \in Z$, is a candidate to be added to Z iff $U(u) > U_{ave}(Z)$. Any addition or removal of a candidate vertex from Z that transforms Z into Z' is considered as *viable* iff $(U(Z') + SB(Z')) - (U(Z) + SB(Z)) > 0$.

Let $O = \{Z_1, Z_2, \ldots, Z_j\}$ be the set of all the overlapping zones detected after the expansion step. The heuristic proposed for improving these zones is as follows: (1) computing the initial average uniformity of each zone $Z_i \in O$, (2) detecting the set T of viable transformations to apply over O, (3) selecting and performing the transformation $t \in T$ which produces the higher improvement in its overlapping zone, and (4) to repeat steps 2 and 3 while $T \neq \emptyset$.

Let $C = \{C_1, C_2, \ldots, C_k\}$ be the set of communities detected after the improving step. Although it is allowable for communities to overlap, what is most important for each community is to have a subset of vertices that makes the community different from the remaining ones. The *merging* step aims to reduce the redundancy in the detected communities, by iteratively merging those communities having a high overlapping.

Let $C_i \in C$ a community. The *distinctiveness* of C_i, denoted as D_{C_i}, is computed as the difference between the number of edges of C_i composed of vertices belonging only to C_i, and the number of edges of C_i composed of at least one vertex C_i shares with another community. Let C_i and C_j be two communities

which overlap each other. C_i and C_j are candidates to be merged iff $D_{C_i} \leq 0$ or $D_{C_j} \leq 0$.

The heuristics for merging communities having a high overlapping degree is as follows: (1) detecting the set PC of pairs of communities which are candidate to be merged, (2) applying the roulette wheel selection method over the set PC, where the probability of selection of each pair is computed by using the highest absolute value of the distinctiveness of the two communities forming the pair, and (3) repeating steps 1 and 2 while $PC \neq \emptyset$. The set of overlapping communities remaining after this step is evaluated by using the intra and inter objective functions described in Sect. 2.1, in order to keep a set of non-dominated solutions.

3 Experimental Results

In this section, the results of several experiments testing the MOGLAOC algorithm are presented. The experiments were focused on: (1) to compare the accuracy attained by MOGLAOC against the one attained by MEA_CDP [1], IMOQPSO [7], iMEA_CDP [6] and OMO [5] algorithms, and (2) to evaluate the number of communities as well as the overlapping degree of these communities, for the best solutions found by our proposal for each network.

In our experiments we use four real-life networks: the American College Football network, the Zachary's Karate Club network, the Bottlenose Dolphins network, and the Krebs' books on American politics network; these networks can be downloaded from http://konect.uni-koblenz.de/networks. Table 1 shows the characteristics of these networks.

Table 1. Overview of the networks used in our experiments

Networks	# of nodes	# of edges	Ave. degree	# communities
American Cool. Football	115	613	10.66	12
Zachary's Karate Club	34	78	4.58	2
Bottlenose Dolphins	62	159	5.129	2
Krebs' books	105	441	8.4	3

In the first experiment we used the NMI external evaluation measure [6] for computing the accuracy of each algorithm. NMI takes values in $[0, 1]$ and it evaluates a set of communities based on how much these communities resemble a set of communities manually labelled by experts, where 1 means identical results and 0 completely different results. For each network, we executed MOGLAOC thirty times and computing the NMI value attained by the best solution of the Pareto front. Table 2 showed the average NMI value attained by each algorithm over the networks; the average values for MEA_CDP, IMOQPSO, iMEA_CDP and OMO algorithms were taken from their original articles. The "X" in Table 2 means that IMOQPSO does not report any results on the Krebs' books network.

Table 2. Best NMI average values attained by each algorithm. Highest values appears bold-faced

Networks	MEA_CDP	IMOQPSO	iMEA_CDP	OMO	MOGLAOC
American Cool. Football	0.495	0.462	0.593	0.33	**0.65**
Zachary's Karate Club	0.52	**0.818**	0.629	0.375	0.75
Bottlenose Dolphins	0.549	**0.886**	0.595	0.41	0.69
Krebs' books	0.469	X	**0.549**	0.39	0.449
Ave. ranking position	3.25	2.75	2.25	4.75	2.0

As it can be seen from Table 2, MOGLAOC attains comparable results with those of the related algorithms. Our proposal clearly outperforms the other algorithms in the American Cool. Football network which is a difficult network, while it is second in Zachary's Karate Club and Bottlenose Dolphins networks, outperformed only by the IMOQPSO, which in turn it includes some operations that make it a highly computational expensive algorithm. In the last row of Table 2 we also showed the average ranking position attained by each algorithm and as it can be observed, our proposal attains the best results. From the above experiments on real-world networks, we can say that MOGLAOC is promising and effective for overlapping community detection in complex networks.

In the second experiment, we compute the average number of communities MOGLAOC detects when it attains its highest NMI value, as well as the overlapping among these communities. The results of this experiment are showed in Table 3.

Table 3. Average number of communities and overlapping degree for the best solutions found by MOGLAOC

Networks	Ave # communities	Ave. overlapping degree
American Cool. Football	9.7	1.1
Zachary's Karate Club	2	1.108
Bottlenose Dolphins	2	1.2
Krebs' books	3.8	1.33

As it can be seen from Table 3, our proposal detects a number of communities which, in the average case, is close to the real number of communities existing in the networks. Moreover, our proposal do not produce solutions having high overlapping degree which means it is able to detect overlapping zones but it does not profit from this overlapping for boosting its accuracy.

4 Conclusions

In this paper, we proposed a new algorithm, named MOGLAOC, for discovering overlapping communities through a combination of global and local optimization approaches. MOGLAOC introduced three steps to the classical Pareto-dominance based MOEAs framework which allow it to detect overlapping zones, to optimise the quality of these zones and to reduce redundancy in the solutions. Unlike previously reported algorithms, our proposal defined two properties that should satisfy any vertex belonging to an overlapping zone.

The MOGLAOC algorithm was evaluated over four real-life networks in terms of its accuracy and it was compared against four algorithms of the related work. The experimental evaluation showed our proposal attains comparable results to that of the evaluated state-of-the-art algorithms. Moreover, this evaluation showed that MOGLAOC is promising and effective for overlapping community detection in complex networks. Another conclusion is that MOGLAOC detects a number of communities which is close to that existing in the networks. Besides, our proposal produces communities having low overlapping degree.

As future work, we would like to further evaluate the MOGLAOC algorithm over synthetical networks in order to have a better insight about its behaviour under different conditions.

References

1. Liu, J., Zhong, W., Abbass, H., Green, D.G.: Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms. In: IEEE Congress on Evolutionary Computation (CEC) (2010)
2. Shi, C., Yan, Z., Cai, Y., Wu, B.: Multi-objective community detection in complex networks. Appl. Soft Comput. **12**(2), 850–859 (2012)
3. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. Proc. Nat. Acad. Sci. **104**(1), 36–41 (2007)
4. Palla, G., Derényi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature **435**, 814–818 (2005)
5. Liu, B., Wang, C., Wang, C., Yuan, Y.: A new algorithm for overlapping community. In: Proceeding of the 2015 IEEE International Conference on Information and Automation Detection, pp. 813–816 (2015)
6. Liu, C., Liu, J., Jiang, Z.: An improved multi-objective evolutionary algorithm for simultaneously detecting separated and overlapping communities. Int. J. Nat. Comput. **15**(4), 635–651 (2016)
7. Li, Y., Wang, Y., Chen, J., Jiao, L., Shang, R.: Overlapping community detection through an improved multi-objective quantum-behaved particle swarm optimization. J. Heuristics **21**(4), 549–575 (2015)
8. Wen, X., Chen, W.N., Lin, Y., Gu, T., Zhang, H., Li, Y., Yin, Y., Zhang, J.: A maximal clique based multiobjective evolutionary algorithm for overlapping community detection. IEEE Trans. Evol. Comput. **21**(3), 363–377 (2016)
9. Corne, D., Jerram, N., Knowles, J., Oates, M.: PESA-II: region-based selection in evolutionary multi-objective optimization. In: Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation (GECCO 2001), San Francisco, CA, pp. 283–290 (2001)

10. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S.: A survey of multiobjective evolutionary clustering. ACM Comput. Surv. **47**(4), 61:1–61:46 (2015)
11. Pizzuti, C.: A multiobjective genetic algorithm to find communities in complex networks. IEEE Trans. Evol. Comput. **16**(3), 418–430 (2012)
12. Zhou, A., Qu, B.Y., Li, H., Zhao, S.Z., Suganthan, P.N., Zhang, Q.: Multiobjective evolutionary algorithms: a survey of the state of the art. Swarm Evol. Comput. **1**(1), 32–49 (2011)