Neurocomputing 000 (2017) 1–13

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Adaptive community detection in complex networks using genetic algorithms

Manuel Guerrero[a], Francisco G. Montoya[b,*], Raúl Baños[b], Alfredo Alcayde[b], Consolación Gil[a]

[a] CeiA3, Department of Informatics, University of Almería, Carretera de Sacramento s/n, Almería 04120, Spain
[b] CeiA3, Department of Engineering, University of Almería, Carretera de Sacramento s/n, 04120 Almería, Spain

ARTICLE INFO

*Article history:*
Received 24 November 2016
Revised 5 April 2017
Accepted 13 May 2017
Available online xxx

Communicated by W.K. Wong

*Keywords:*
Genetic algorithms
Network optimisation
Community detection
Modularity

ABSTRACT

Community detection is a challenging optimisation problem that consists in searching for communities that belong to a network or graph under the assumption that the nodes of the same community share properties that enable the detection of new characteristics or functional relationships in the network. A large number of methods have been proposed to address this problem in many research fields, such as power systems, biology, sociology or physics. Many of those optimisation methods use modularity to identify the optimal network subdivision. This paper presents a new generational genetic algorithm (GGA+) that includes efficient initialisation methods and search operators under the guidance of modularity. Further, this approach enables a flexible and adaptive analysis of the characteristics of a network from different levels of detail according to an analyst's needs. Results obtained in networks of different sizes and characteristics show the good performance of GGA+ in comparison with other five genetic algorithms, including efficient algorithms published in recent years.



## 1. Introduction

Network analysis is a powerful tool for describing many real systems, such as sociology [6], biology [23] or power systems [16], among many others. In particular, it has been proven that many real networks have a structure of modules which are characterized by groups of densely interconnected nodes. These modules or communities are subgraphs of the network such that nodes within a community are densely linked, while connections between subgraphs are sparser. These communities represent functional units in their networks, for example, a community in a social network corresponds to people that often communicate with other people of the same community, i.e., they share similar interests or backgrounds [23].

Having in mind the importance of Genetic Algorithms (GAs) [7] in computational optimisation, this paper presents a new generational genetic algorithm (GGA+) for solving the community detection problem, which is guided by the modularity index [15] and considers different *degrees of abstraction*. This concept provides the analyst the capacity to perform a flexible and adaptive analysis of the network using graphical information with different levels of detail.

The remainder of the paper is organized as follows: Section 2 briefly describes the problem of community detection in graphs, and some previous approaches applied to this problem. Section 3 describes in detail the generational genetic algorithm here proposed. Section 4 presents the empirical study, where GGA+ and other five genetic algorithms are evaluated using four networks of different sizes and characteristics. The conclusions of the work are provided in Section 5.

## 2. Related work

The study developed by Leonhard Euler in 1736 about the Seven Bridges of Königsberg problem laid the foundations of graph theory and initiated an ongoing study of the properties of graphs [3]. Graphs are used to model real systems in many areas, such as power systems [16], sociology [6], biology [23], physics [1], and informatics [2,8] thanks to the computers, which allow to efficiently representing, managing and processing large amounts of data, including graph-based structures.

Among other emerging scientific areas, the analysis of communities in complex systems has gained importance due to their application to multiple contexts, including collaboration networks, biological systems, social networks in the Internet, transportation

* Corresponding author.
*E-mail addresses:* mgl220@fm.ual.es (M. Guerrero), pagilm@ual.es (F.G. Montoya), rbanos@ual.es (R. Baños), aalcayde@ual.es (A. Alcayde), cgilm@ual.es (C. Gil).

http://dx.doi.org/10.1016/j.neucom.2017.05.029
0925-2312/

networks, or electrical networks [12]. All complex systems share a common characteristic: community structures [14] that consist of groups of nodes inside a network that are more densely connected than with the remaining nodes of the network. As the nodes that belong to the same community have a higher probability of shared properties, community detection can reveal new characteristics or functional relationships of a network. The community detection problem has been analysed by the scientific community [20,23] and consists in searching for the community structure that better represents the characteristics of a network. The complexity of this problem comes from the difficulty of determining the optimum community structure that best represents the characteristics of a network.

*Modularity* [15] has become one of the most extensively applied objective functions in community detection due to is simplicity and ease of calculation. Modularity provides a numerical value that represents the quality of the solution, such that the greater the value is, the more accurate the community structure. Modularity (Q) is defined as

$$Q = \frac{1}{2M}\sum\left(a_{ij} - \frac{K_i K_j}{2M}\right)\delta(i,j) \quad (1)$$

where $M$ represents the total number of edges in the network; the sub-indices $i$ and $j$ indicate two nodes of the network; $K_i$ and $K_j$ are the degree of the $i$th and $j$th nodes, respectively; the parameter $a_{ij}$ is the element of the $i$th row and the $j$th column of the adjacency matrix; and $\delta(i,j)$ represents the relationship between the $i$th node and the $j$th node, such that if node $i$ and node $j$ are in the same community, $\delta(i,j) = 1$; otherwise, $\delta(i,j) = 0$.

## 3. GGA+: a new generational genetic algorithm for community detection

Many heuristic and meta-heuristic methods have been proposed to solve community detection problems, including simulated annealing [10], swarm intelligence [19], and genetic and evolutionary algorithms [7]. This section presents a new generational genetic algorithm (GGA+) that is guided by the modularity index (MI) [15] and includes efficient strategies and search operators to detect communities in networks.

### 3.1. Initialisation

In many real-world situations, the number of community structures that form a network is known beforehand and, therefore, the search space to be explored by the algorithms can be reduced. In other cases, the number of community structures is initially unknown, but the algorithms can estimate a different number of community structures. Therefore, by determining beforehand the number of initial community structures to be detected, the search space can be reduced and the performance of the algorithm can be improved.

Despite random initialisation is often used, it can generate infeasible solutions, that is, nodes that are not interconnected due to a lack of relationship between them. To overcome this inconvenience, a *safe initialisation* is considered, such that each node $i$ is connected with a neighbouring node $j$ of the original graph [18]. Undesired clusters with disconnected nodes are prevented and, therefore, the search space of possible solutions is restricted to feasible individuals, which helps to improve the algorithm convergence.

To prevent the generation of individuals having unbalanced communities, the concept of *community density* is used, which is incorporated into the structure of individual data to balance the loads between communities (*balanced initialisation*). In this phase, the community density vector of the individuals of the population is configured. The number of nodes that each community contains (community size) is a function of the number of communities to be detected. Subsequently, these sizes are employed to initialise each component of the *community density* vector with the number of nodes in each community.

The genetic initialisation in GGA+ is based on the *safe and balanced initialisation* concept, in which a maximum node size is assigned to each community, as it is described below:

- Populate communities with neighbouring nodes: a community to be populated and the node $n_i$ are selected. Based on $n_i$, the community is populated with this node and its neighbours $nei_{ij}$ until the maximum community size is attained. If the community is not completed with the values $nei_{ij}$ of the node $n_i$, another node $n_{i+n}$ (not included in the nodes that already exist in the community) is selected to repeat the process. This procedure is repeated until the community is completed.
- Populate communities with adjacent neighbouring nodes: a community to be populated and a node $n_i$ are selected, and a neighbouring node $nei_{ij}$ of the node $n_i$ is subsequently selected. Based on the node $nei_{ij}$, the community is completed with this node and its neighbours $nei_{ijk}$ until the maximum community size is attained. If the community is not completed with the neighbouring nodes $nei_{ijk}$ of the node $nei_{ij}$, another neighbouring node $nei_{i(j+n)}$ of the node $n_i$ (not included in the nodes that already exist in the community) is selected to repeat the process. This procedure is repeated until the community is completed.

Fig. 1 displays an example that illustrates both concepts: safe initialisation and balanced initialisation. The number of communities to be detected in Fig. 1 (two communities in this case) is used to calculate the node size for each community in the balanced initialisation stage. Once the size of each community is defined, the safe initialisation stage begins; in this case, the *populate communities with neighbouring nodes* method is used to initiate the individual. Therefore, the community structure is generated by the combination of the safe and balanced initialisations, which coincides with the optimum problem solution in the case study presented in Fig. 1.

On the other hand, initialisation of the migration vector between boundaries is based on the idea of migration towards the most attractive destination, which coincides with the community that has the greatest number of nodes directly connected to a certain node. Given a node, the community that contains the highest number of nodes connected to this one will be selected as the destination boundary community to which the selected node will migrate.

The scheme described for the initialisation of the migration vector between boundaries is shown in Fig. 2. Fig. 2(d) shows the resulting vector, which elements represents nodes of the graph, its values, and the destination community to which each node could migrate. Table 1 shows the boundary community generation method, in which the columns *Communities 1* and *2* indicate, for the corresponding community, the number of nodes directly connected to the node of the column *Nodes*. The column *Migration Vector* shows the destination community selected for each node based on the highest recorded value among all community counters (in this example, the counters are communities 1 and 2).

### 3.2. Genetic operators

The algorithm here proposed (GGA+) uses genetic search operators that have been especially designed to obtain the maximum performance of the proposed data structure.

The crossover operator is based on the exchange of nodes between boundaries, such that it is selected a first node with a

Please cite this article as: M. Guerrero et al., Adaptive community detection in complex networks using genetic algorithms, Neurocomputing (2017), http://dx.doi.org/10.1016/j.neucom.2017.05.029