2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

Evolutionary clustering algorithm for community detection using graph-based information

Gema Bello-Orgaz and David Camacho

Abstract— **The problem of community detection has become highly relevant due to the growing interest in social networks. The information contained in a social network is often represented as a graph. The idea of graph partitioning of graph theory can be apply to split a graph into node groups based on its topology information. In this paper the problem of detecting communities within a social network is handled applying graph clustering algorithms based on this idea. The new approach proposed is based on a genetic algorithm. A new fitness function has been designed to guide the clustering process combining different measures of network topology (Density, Centralization, Heterogeneity, Neighbourhood, Clustering Coefficient). These different network measures have been experimentally tested using a real-world social network. Experimental results show that the proposed approach is able to detect communities and the results obtained in previous work have been improved.**

I. Introduction

Community detection is a great importance problem applied in disciplines such as sociology, biology and computer science, whose information is often represented as graphs [1]. A Social Network can be represented by a graph, where the vertices are individuals, and the edges are the relationships among them. Then this graph representation can be clustered into node groups based on the topology information of the graph, where each cluster includes strongly interconnected vertices.

Therefore, the problem of detecting communities within a social network can be handled using graph clustering algorithms. There is no a single definition accepted of a cluster in a graph, and the variants used in the literature are numerous. But these kinds of algorithms are typically based on the topology information of the graph or network. Related to the graph connectivity, each cluster should be connected; it means that should be several paths connecting each pair of vertices within the cluster. It is generally accepted that a subset of vertices forms a good cluster if the induced sub-graph is dense, and there are few connections from the included vertices to the rest of the graph [2]. Considering both features, connectivity and density, a possible definition of a graph cluster could be a connected component or a maximal clique [3]. This is a sub-graph into which no vertex could be added without losing the clique property. On the other hand, it is not always clear that a vertex should be assigned only to a unique cluster. In some domains could be interesting that a vertex belongs to several clusters. To solve this problem, fuzzy clustering algorithms applied to graphs [4] and overlapping approaches [5] have been proposed.

This family of algorithms can be improved using Genetic Algorithms (GAs) to decrease their high computational complexity when they are applied to networks of very large sizes. Several of these evolutionary clustering algorithms use a single optimization criteria as the objective function, such as modularity [6]. There are also GAs where the community detection is solved as a multiobjective optimization problem, generally using two criteria to optimize [7], [8]. This paper aims to analyse the possible combination of several metrics from Network Topology (Density, Centralization, Heterogeneity, Neighbourhood, Clustering Coefficient) in order to find new approaches to improve detection community algorithms using GAs. To this purpose a new fitness function has been designed enabling combine various measures of network topology to guide the algorithm. The measures used in the fitness function and their weights can be changed in the algorithm settings. Then the algorithm is applied to a real-world social network and a detailed study of the most appropriate network metrics will be carried out.

The rest of the paper is structured as follows. Section 2 describes the related work concerning clustering, genetic algorithm and community detection algorithms. Section 2 presents the genetic algorithm, the encoding designed and the new fitness function implemented. Section 4 provides a description of the dataset used, the experimental setup of the algorithm and a complete experimental evaluation of it. Finally, in Section 5 the conclusions and some future research lines of the work are presented.

II. Related Work

Nowadays, the Community Detection Problem in Complex Networks has been the subject of numerous studies in the field of Data Mining and Social Network Analysis. Several methodologies have been applied to find optimal groups of nodes into communities. Usually these methods require a vast amount of memory and computational time to process large-scale networks in real world domains such as the World Wide Web, citation networks, transportation networks, and social and biochemical networks among others [1].

The goal of community detection problem is similar to the idea of graph partitioning of graph theory [9][10]. In computer science, the process of identifying the underlying structure of the data in terms of grouping the elements is called clustering, and a cluster in a graph could be called a community. Clustering [11] is an unsupervised classification

Gema Bello-Orgaz and David Camacho are with the Department of Computer Science, Universidad Autónoma de Madrid, 28049, Madrid, Spain (email: {gema.bello, david.camacho}@uam.es).

This work was supported by Spanish Ministry of Science and Education under Project Code TIN2010-19872 and Savier Project (Airbus Defence & Space, FUAM-076915)

978-1-4799-1488-3/14/$31.00 ©2014 IEEE

technique where a set of elements, usually represented as vectors in a multi-dimensional space, are grouped into clusters (or groups). The elements include in the same cluster should be similar, and elements include in different clusters should be dissimilar. For this reason, it is necessary to define a measure of similarity which establishes a rule for assigning elements to a particular cluster or group.

Graphs are structures formed by a set of vertices (also called nodes) and a set of edges which are connections between pairs of vertices. Graph clustering [12][13] is the task of grouping the vertices into clusters considering the edge structure of the graph. There should be many edges within each cluster and relatively few between the clusters.

One of the most well-known algorithms for Community Detection was proposed by Girvan and Newman [14]. This method uses a new similarity measure called "edge betweenness" based on the number of the shortest paths between all vertex pairs. This algorithm has a high complexity, for this reason Newman reformulated the modularity measure in terms of eigenvectors. The new characteristic matrix for the network is called modularity matrix [15]. This algorithm, based on modularity, has been employed by many authors to study community structures of complex networks, and it shows excellent performance when the size of the network is small. The main disadvantage of this algorithm is the high computational complexity on networks of very large sizes. Subsequently, this modularity measure was modified trying to reduce the computational demands significantly through several new approaches [9][16][17]. On the other hand, other algorithms have been designed to detect overlapping communities such as Cfinder algorithm based on the k-cliques of a graph [18]. But, the complexity of this procedure can be also high, and the computational time needed to find all k-cliques of a graph is an exponentially growing function related to the graph size.

Several clustering algorithms such as K-means or fuzzy c-means [19] [20] [5] [21] [22] have been improved using genetic algorithms. This kind of algorithms have been usually employed in optimization problems [23], where the fitness function tries to find the best solution among a population of possible solutions which are evolving. In clustering approaches, the encoding and optimization algorithm are used to look for the best set of groups optimizing a particular feature of the data.

Regarding to the measures used to find clusters, there are two main approaches [12]: the first one computes some values from the vertices and then classify them into clusters, and the second one compute a fitness measure over the set of possible clusters choosing the optimal among all cluster candidates. For the first one approach, there are several similarity, or distance, metrics that can be applied such as the Euclidean Distance, the Jaccard index or Cosine similarity, among others [24]. The other approaches compute a fitness measure (variants of density, measures that are based on the fraction or number of edges present in the induced subgraph, reaching the maximum value for cliques, etc...) over the set of possible clusters [25]. Then the group of clusters optimizing the measure used are chosen.

Many evolutionary clustering algorithms use a single optimization criteria as the objective function, being the modularity one of the most criteria used [6]. There are also works where the community detection is solved as a multi-objective optimization problem. In Gong et al. [7] a multi-objective evolutionary algorithm based on decomposition is proposed to maximize the density of internal degrees, and minimize the density of external degrees simultaneously. The multiobjective genetic algorithm for networks (MOGA-Net) [8] optimizes two objective functions to identify densely connected groups of nodes having sparse inter-connections. The first objective function uses the concept of community score [26] to measure the quality of the network division into communities. The second defines the concept of fitness [27] of the nodes belonging to a module and iteratively finds modules having the highest sum of node fitness. Finally, Amiri et al. [28] design three objective functions to guide a multiobjective evolutionary algorithm measuring quality, separation and overlapping communities respectively.

The new algorithm proposed is based on an evolutionary overlapping clustering approach using a fitness function combining different measures of topology network. Overlapping clustering algorithms can be classified into two main approaches [29]: soft (each object fully belongs to zero or more clusters) and fuzzy (each object belongs to zero or more clusters with a membership probability). One of the first approximations was fuzzy K-means [30], which can also benefit from combining with a genetic approach [31][32]. The new algorithm designed is related to soft computing allowing each node in the graph to belong to one or more subgraphs, and no membership probability is considered.

III. Clustering Genetic Algorithm for Community Detection

The Clustering Genetic Algorithm for Community Detection (CGACD) uses a genetic algorithm to find the best K communities in a network where any particular node could belong to different communities. In a previous work [5], the algorithm K-adaptive GCF has been designed to detect overlapping communities. It includes the value of K in the evolutionary process through a new encoding. This section describes the algorithm, including the encoding and the new fitness function designed. This function is a combination of several metrics from the network topology in order to tune up the community sets selection.

A. Encoding

A possible solution for the problem should contain a group of communities, for this reason, the genotypes (chromosomes) are represented as a set of vectors of binary values. Each allele represents a community composed by a set of binary values, one for each node in the network. For these binary vectors the value 1 meaning the node belongs to the community and value 0 the opposite. Therefore, the number of binary vectors (communities) contained in the

chromosome (group of communities) will be the number of detected communities (K), see Fig. 1

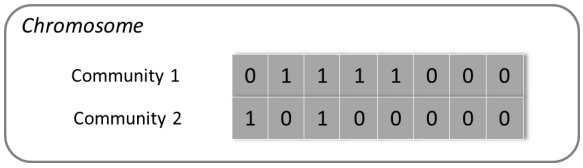

Fig. 1. Chromosome representing a group of communities in a network. Each allele represents a particular community where its binary vector represents the nodes of the network, and their belonging (or not) to the current community. In this example the solution contains 2 vectors representing two different communities detected.

B. The Algorithm

In the clustering genetic algorithm, the population evolves using a standard GA as Algorithm 1 shown, where the genetic operators work as follows:

- **Crossover**. A randomly crossover point is chosen. Then every community preceding this point is copied from both parents to create a new child. And every community succeeding this point is copied to create a second new child. See lines 13 and 14 in Algorithm 1.
- **Mutation**. Some values of the vectors representing the communities are randomly chosen and they change their values (with a predefined mutation probability) from 1 to 0 or vice versa. See lines 15 and 16 in Algorithm 1.

The algorithm performs an Elitism selection where the n-best chromosomes to the population are copied to the new population (line 9 in Algorithm 1). It prevents losing the n-best found solutions.

C. The Fitness Function

The new fitness function designed combines metrics from Network Topology. It tries to find a set of communities where their members are highly connected between them and they have similar behaviour. To combine several network metrics in a single fitness function, a weighted function based on these metrics chosen is designed. The measures used in the fitness function and their weights can be changed in the algorithm setting, and this function is calculated as follows:

$$F = \sum_{i=1}^{n} w_i * Metric_i \tag{1}$$

Where $Metric_i$ is a network metric extracted from graph theory to guide the evolutionary algorithm, and w_i are the weights given to each metric: $w_i \in (0, 1)$. In order to compute the metrics chosen and their weights, a preliminary analysis of them is done in section IV of experimental results. This fitness function will be based on the following metrics: neighbourhood, clustering coefficient, density, centralization and heterogeneity. Previous metrics derive from graph theory, for this reason, some of the basic concepts and metrics used in own fitness designed are briefly introduced.

Algorithm 1: Clustering Genetic Algorithm for Community Detection (CGACD)

Input: A network $N = (V, E)$ where V is a set of vertices denoted by $\{v_1, \ldots, v_n\}$ and E is a set of edges E denoted by e_{ij} representing whether there is a connection between the vertices v_i and v_j. And positive numbers $generations$, $population$, μ, λ and $mutprobability$

Output: The chromosome $C_i = \{g_1, g_2, \ldots, g_n\}$ such that $Fitness(C_i)$ is minimized

1 $C \leftarrow$ randomly generated set of $population$ chromosomes
2 $i \leftarrow 1$
3 $convergence \leftarrow 0$
4 **while** $i \leq generations \wedge convergence = 0$ **do**
5 $F \leftarrow \emptyset$
6 **for** $j \leftarrow 1$ ***to*** $population$ **do**
7 $F \leftarrow Fitness(C_j)$
8 $Cbest \leftarrow SelectNBest(\lambda, F)$
9 $C \leftarrow Cbest$
10 **for** $j \leftarrow \mu$ ***to*** λ **do**
11 $p1 \leftarrow$ randomly selected chromosome from $Cbest$
12 $p2 \leftarrow$ randomly selected chromosome from $Cbest$
13 $c1 \leftarrow Crossover(p1, p2, 1)$
14 $c2 \leftarrow Crossover(p1, p2, 2)$
15 $c1 \leftarrow Mutation(c1, mutprobability)$
16 $c2 \leftarrow Mutation(c2, mutprobability)$
17 $C \leftarrow C \cup \{c1, c2\}$
18 $i \leftarrow i + 1$
19 $convergence \leftarrow CheckConvergence(Cbest)$
20 **return** *SelectBest(C,F)*

1) Graph: A graph $G = (V, E)$ is a set of vertices or nodes V denoted by $\{v_1, \ldots, v_n\}$ and a set of edges E where each edge is denoted by e_{ij} if there is a connection between the vertices v_i and v_j. Graphs can be directed or undirected. If all edges satisfy the equality $\forall i, j$, $e_{ij} = e_{ji}$, the graph is said to be undirected.

2) Neighbourhood: Any algorithm working with the vertices of a graph needs to analyse each node neighbours. If the edge $e_{ij} \in E$ and $e_{ji} \in E$ we say that v_j is a neighbour of v_i. The neighbourhood of v_i Γ_{v_i} is defined as

$$\Gamma_{v_i} = \{v_j \mid e_{ij} \in E \text{ and } e_{ji} \in E\} \tag{2}$$

Then, the number of neighbours of a vertex v_i is

$$k_i = |\Gamma_{v_i}| \tag{3}$$

3) Clustering Coefficient (CC): Once the most general and simple concepts from graph theory are defined, we can

proceed with the definition of some basic measures related to any node in a graph.

Let A be an adjacency matrix with elements a_{ij}. And let Γ_{v_i} be the neighbourhood of the vertex v_i. If k_i is considered as the number of neighbours of a vertex, we can define the clustering coefficient ***(CC)*** of a vertex as follows:

$$C_i = \frac{1}{k_i(k_i-1)} \sum_{j,h} a_{jh} a_{ij} a_{ih} a_{ji} a_{hi} \tag{4}$$

The Local CC measure provides values ranging from 1 to 0. Where 0 means that the node and its neighbours do not have clustering features, so they do not share connections between them. Therefore, value 1 means that they are completely connected. Finally, if we want to study a general graph, we should study its Global CC that can be defined as:

$$C = \frac{1}{|V|} \sum_{i=0}^{|V|} C_i \tag{5}$$

Where $|V|$ is the number of vertices.

4) Density: The connectivity of a node is the size of its neighbourhood. The average number of neighbours indicates the average connectivity of a node in the network. A normalized version of this parameter is the network density. The density is a value between 0 and 1. A network containing no edges and isolated nodes has a density of 0. On the other hand, the density of a clique (that is a subset of its vertices such that every two vertices in the subset are connected by an edge) is 1. It can be defined as the mean off-diagonal adjacency as follows:

$$Density = \frac{\sum_i \sum_{j!=i} a_{ij}}{n(n-1)} \tag{6}$$

where n is the number of the nodes within the network.

5) Centralization: This network measure is also known as degree centralization used as an index related to the connectivity distribution. Networks whose topologies resemble a star have a centralization close to 1, whereas decentralized networks are characterized by having a centralization close to 0. Its value is given by:

$$Centralization = \frac{n}{n-2} * (\frac{max(k_i)}{n-1} - Density) \tag{7}$$

6) Heterogeneity: Finally, the heterogeneity of the degree distribution has been the focus of considerable research in recent years. Many measures of network heterogeneity are based on the variance of the connectivity. In this work, this measure notices the tendency of a network to contain "hub" nodes. It can be defined as the coefficient of variation of the connectivity distribution:

$$Heterogeneity = \frac{\sqrt{variance(k_i)}}{mean(k_i)} \tag{8}$$

IV. Experimental Results

In this work, the data from the Eurovision Song Contest have been selected to solve the problem of European Country Communities detection. This song contest provides an active forum where countries are free to give opinions about the rest of the participants. The votes emitted in a particular year of this song contest can be easily represented as a graph. In this representation, the nodes are the participant countries and the edges connect the countries which have exchanged points, see Fig.2.

There are several social network datasets to test the algorithm, but the Eurovision dataset is a real-world social network that has been studied using different data mining techniques since the nineties. Several studies on the Eurovision contest have been able to group the participating countries into blocs or communities of like behaviour using clustering methods [33], [34], regression analysis [35], dynamical networks [36], or analytical identification of statistically significant tends [37]. All of these works were able to group the participating countries into blocs of like behaviour and therefore they demonstrates a community structure in the Eurovision dataset. For this reason this dataset has been selected to carry out the experimental phase of the algorithm. The data used in this work have been extracted from The Eurovision's official website [38]

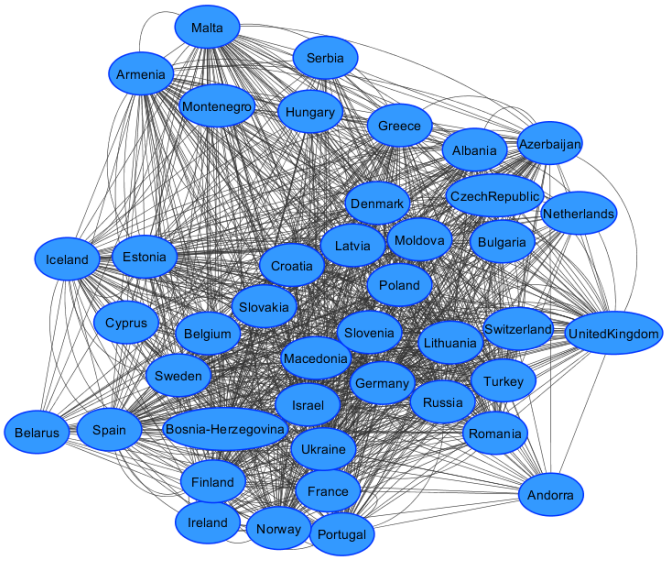

Fig. 2. Graph representing the votes emitted in the Eurovision Song Contest (2009 year).

Table I shows the parameters of the genetic algorithm used throughout the experimental phase. These parameters were obtained experimentally by performing several tests with different range of values. The dataset used is the same as in the preliminary analysis of networks metrics shown in next subsection. $\mu+\lambda$ is the selection criteria used, where λ is the number of offspring (population size), and μ is the number of the best parents that survive from the current generation to the next.

TABLE I

GENETIC PARAMETERS OF CLUSTERING ALGORITHM

Mutation probability	0.03
Generations	100
Population size	500
Selection criteria ($\mu + \lambda$)	$500 + 50$

A. *Preliminary analysis of network metrics*

In a previous work [39] a data analysis about the Eurovision vote dataset was performed using the clustering coefficient. This analysis confirms the existence of clusters or communities in this dataset. Specifically, the 2009 year dataset has the greatest difference in clustering coefficient, meaning that this year contains a large set of different communities. Hence, this year has been selected to perform a detailed study for different network measures which can be late used to tune up the fitness function designed. These network measures were described in section III.

The concept of good partition in groups for a set is sometimes quite subjective. There are two objective functions in clustering literature to measure the cluster quality: intra-cluster distance (elements within a cluster should be close) and inter-cluster distance (elements from different clusters should be away). These two distance measures have been calculated to compare the results obtained by each network measure used in the fitness function. The values obtained for each measure are shown in Table II.

The algorithm goal is to create communities consistent internally, but clearly different from each other. Members within a community should be as similar as possible (lower intra cluster distance), and members in one community should be as dissimilar as possible from members in other communities (higher inter cluster distance). Analysing the distance measures shown in Table II, it can be noticed that the fitness function based on Density measure obtains the better results (with a lower intra cluster value and the higher inter cluster value).

On the other hand, using the Density measure the algorithm detects small communities without overlapping. Heterogeneity measure achieves the best result related to these features. In this case the communities detected are bigger with great overlapping. Therefore, both measures have been combined in the new weighted fitness function trying to detect communities with better results considering all the features.

Once the best network measures for the fitness function are selected (Density and Heterogeneity), it is necessary to perform the estimation related to the best weight for each measure within the function. For this purpose, a comparative assessment of weights for both measures is carried out. The results obtained are shown in Fig. 3.

As can be seen in Fig. 3, with Density equals to 0.9 and Heterogeneity equals to 0.1 (0.9D-0.1H) the intra cluster distance is minimized and the inter cluster measure takes a high value. The intra cluster distance progressively is worse

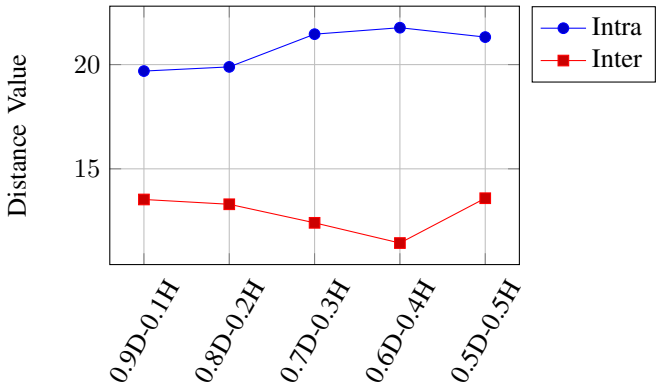

Fig. 3. Comparative assessment of weights for the network measures in the fitness function. The x-values represent the weight considered for Density (i.e 0.9D) and for Heterogeneity (i.e 0.1H) measures in the fitness function.

while its value increases. Therefore, these values have been finally selected to fix the fitness function.

B. *Comparative assessment of algorithms*

Finally, the results obtained in a previous work [5] have been compared with the results obtained using the new fitness approach. In the previous study the Clique Percolation Method (CPM) and The Genetic-based Community Finding Algorithm (GCF) have been applied to community detection. The year chosen for the analysis in this previous work was 2006 (from the Eurovision Song Contest too).

CPM [40] finds communities using k-cliques (where k is a fixed value of connections in a graph) which are defined as complete (fully connected) subgraphs of k vertices. It defines a community as the highest union of k-cliques. Otherwise GCF algorithm [5] is a previous version of this new approach, and it uses the Euclidean distance between nodes and the clustering coefficient to guide the genetic clustering algorithm.

Using the main conclusions from the preliminary analysis for network metrics shown in Fig.3, some new experimental tests have been carried out. Experiments have been performed using the new fitness function wherein the weights for Density and Heterogeneity have been fixed to 0.9 and 0.1 respectively. The results obtained for each algorithm are shown in Table III.

TABLE III

VALUES OF INTRA AND INTER CLUSTER DISTANCES USING DIFFERENT ALGORITHMS TO 2006 YEAR DATASET.

	CPM	**GCF**	**CGACD**
Intra Distance	20,75	18,5	**16,19**
Inter Distance	11,81	**13,56**	*13,32*
RunTime(s)	**2,34**	6,44	*5,49*

CPM algorithm obtains the best result at runtime, but the worst results at the distance metrics measuring the cluster quality. A real-world social network of small size has been chosen to test the algorithm, because it allows easily interpret

TABLE II

COMPARATIVE ASSESSMENT OF CLUSTER DISTANCES (INTER AND INTRA CLUSTER) FOR NETWORK MEASURES IN THE 2009 CONTEST YEAR.

Network Measure	Community	Countries	Intra Distance	Inter Distance
Density	1 2 3	Denmark Greece Norway Romania Albania Russia	**19,25**	**15,79**
Centralization	1 2	Belgium Ireland Ukraine Hungary Albania Serbia	21,32	14,66
Heterogeneity	1 2 3 4 5 6	Norway Sweden Armenia Albania Moldova Israel Denmark Finland Ukraine Turkey Azerbaijan Norway Croatia Estonia Sweden Albania Moldova Israel Denmark Finland Lithuania Ukraine Turkey Germany Azerbaijan Croatia Sweden BosniaandHerzegovina Armenia Albania Malta Russia Finland Romania Lithuania Ukraine Germany Azerbaijan Croatia Sweden BosniaandHerzegovina Malta Moldova Denmark Finland Lithuania Ukraine Iceland Germany Norway Armenia Moldova Israel Denmark Finland Spain Iceland Turkey Azerbaijan Estonia BosniaandHerzegovina Albania Moldova Israel Denmark Ukraine Iceland Germany Azerbaijan	20,72	4,92
Neighbourhood	1 2 3	France Norway Croatia Estonia Sweden Armenia Malta Portugal Israel Denmark Finland Lithuania Greece Iceland Azerbaijan UnitedKingdom France Norway Croatia Estonia Sweden Albania Malta Russia Portugal Israel Finland Lithuania Turkey France Norway Croatia Sweden BosniaandHerzegovina Malta Russia Portugal Israel Denmark Romania Lithuania Ukraine Spain Turkey Azerbaijan	21,55	3,99
CC	1 2 3	Norway Ukraine Azerbaijan Moldova Ukraine Azerbaijan Russia Ukraine Azerbaijan	21,27	7,01

the results obtained (community structure, overlapping, size, etc...) and the quality of the clusters found. But, in order to assess the effectiveness of the method, it would be necessary to test the algorithm using complex networks in future work. This work is focused on the analysing of the quality of the clusters detected by the distance metrics. Regarding the distance results shown in Table III, it can be noticed that the genetic algorithm using the new fitness function obtains the better results. This new approach takes the lowest intra cluster value and a high inter cluster result (very close to the best value obtained by GFC algorithm).

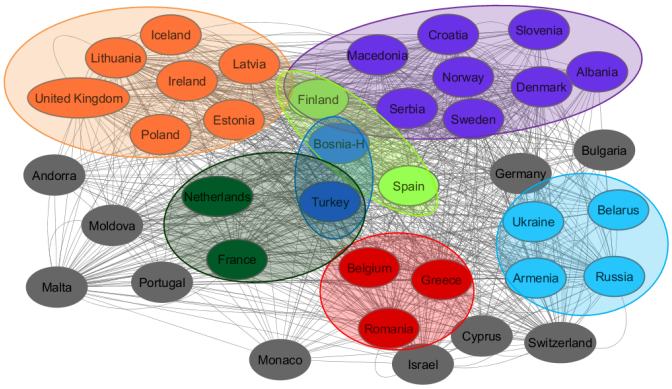

Fig. 4. CPM communities detected for 2006 year.

Additionally, from Fig.4 to Fig.6 the communities detected are plotted in a graph representation. This representation provides a better appreciation of the community structure and size. Analysing these features there are various remarkable aspects. In Fig.4, it can be noticed that the CPM method detects big communities with great overlapping. But the intra and inter cluster distance take the worst values, meaning that there are fewer connections between nodes within communities, and also these communities are not well differentiated. On the other hand, the communities resulting from the genetic algorithms are smaller as it can be seen in Fig. 5 and Fig. 6. This result could be expected for GCF algorithm due to the definition of its fitness function. This function uses the distances between centroids and the distance between the nodes belonging to a group to guide the community detection. However, the new fitness function designed ignores the distance measures; it is only based on network topology metrics and improves the results obtained.

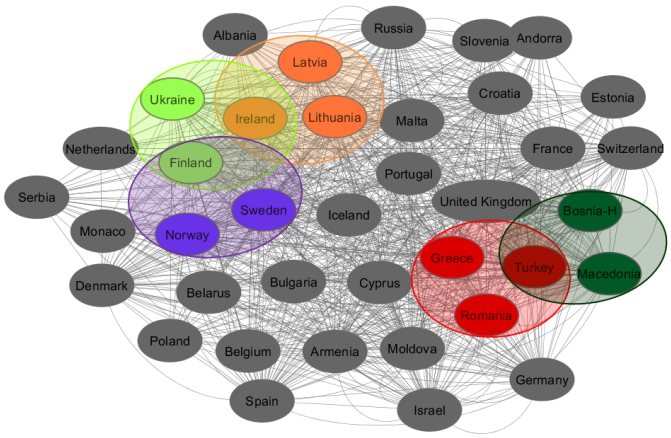

Fig. 5. GCF communities detected for 2006 year.

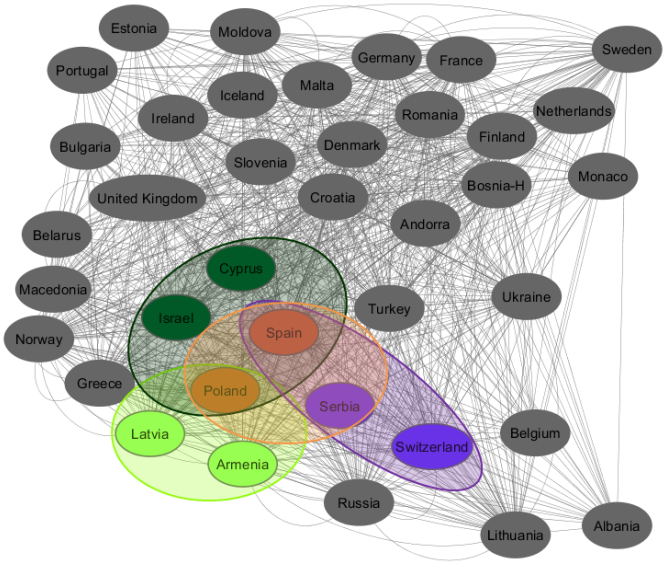

Fig. 6. CGACD communities detected for 2006 year.

V. Conclusions and Future Work

A new fitness function has been designed and implemented to improve a genetic clustering algorithm for community detection. This new fitness function approach has been inspired by network topology analysis, and it is based on the use of network measures (Density, Centralization, Heterogeneity, Neighbourhood, Clustering Coefficient) to guide the search. The measures used in the fitness function and their weights can be changed in the algorithm settings. Then the algorithm is applied to a real-world social network (Eurovision dataset). In order to choose the better measures for the fitness function, a comparative assessment of network measures has been carried out. Additionally, the experimental results obtained are compared to other clustering algorithms (CPM methods and GCF algorithm). These results show that the new approach is able to reach better results than the other approaches studied. Using the new fitness function, the algorithm detects communities (clusters) with an appropriate size, reduced overlapping, members very similar, and close distances between those clusters detected.

Finally some improvements can be made in the algorithm. In the future work, the algorithm will be experimentally tested in order to assess the effectiveness of the method on complex networks. To this purpose, complex graphs with known community structure will be generated. Additionally, the algorithm will extend to multi-objective approaches using specialized fitness functions and introducing methods to promote solution diversity.

References

[1] S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, 2003.

[2] R. Kannan, S. Vempala, and A. Veta, "On clusterings-good, bad and spectral," in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, ser. FOCS '00. Washington, DC, USA: IEEE Computer Society, 2000, pp. 367–.

[3] I. M. Bomze, M. Budinich, P. M. Pardalos, and M. Pelillo, "The maximum clique problem," in *Handbook of Combinatorial Optimization*. Kluwer Academic Publishers, 1999, pp. 1–74.

[4] Y. hong Dong, Y. Zhuang, K. Chen, and X. Tai, "A hierarchical clustering algorithm based on fuzzy graph connectedness." *Fuzzy Sets and Systems*, vol. 157, no. 13, pp. 1760–1774, 2006.

[5] G. Bello-Orgaz, H. Menendez, and D. Camacho, "Adaptive k-means algorithm for overlapped graph clustering," *International Journal of Neural Systems*, vol. 22, no. 05, pp. 1 250 018 1–19, 2012.

[6] R. Shang, J. Bai, L. Jiao, and C. Jin, "Community detection based on modularity and an improved genetic algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 5, pp. 1215–1231, 2013.

[7] M. Gong, L. Ma, Q. Zhang, and L. Jiao, "Community detection in networks by using multiobjective evolutionary algorithm with decomposition," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 15, pp. 4050–4060, 2012.

[8] C. Pizzuti, "A multiobjective genetic algorithm to find communities in complex networks," *Evolutionary Computation, IEEE Transactions on*, vol. 16, no. 3, pp. 418–430, 2012.

[9] A. Clauset, "Finding local community structure in networks," *Phys. Rev. E*, vol. 72, p. 026132, Aug. 2005.

[10] F. Santo, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[11] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1988.

[12] S. S. Elisa, "Survey: Graph clustering," *Comput. Sci. Rev.*, vol. 1, no. 1, pp. 27–64, Aug. 2007.

[13] H. Menéndez and D. Camacho, "A genetic graph-based clustering algorithm," in *Intelligent Data Engineering and Automated Learning - IDEAL 2012*, ser. Lecture Notes in Computer Science, H. Yin, J. Costa, and G. Barreto, Eds. Springer Berlin / Heidelberg, 2012, vol. 7435, pp. 216–225.

[14] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[15] C. G. Wang Xutao and L. Hongtao, "A very fast algorithm for detecting community structures in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 384, no. 2, pp. 667–674, Oct. 2007.

[16] M. E. Newman, "Modularity and community structure in networks," *Proc Natl Acad Sci U S A*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.

[17] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, pp. 066 133+, Jun. 2004.

[18] I. F. Gergely Palla, Imre Derényi and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, June 2005.

[19] U. Maulik, S. Bandyopadhyay, and S. B, "Genetic algorithm-based clustering technique," *Pattern Recognition*, vol. 33, pp. 1455–1465, 2000.

[20] M. Naldi, S. Salcedo-Sanz, L. Carro-Calvo, L. Laura, A. Portilla-Figueras, and G. F. Italiano, "A traffic-based evolutionary algorithm for network clustering." *Appl. Soft Comput.*, vol. 13, no. 11, pp. 4303–4319, 2013.

[21] Y. Li, J. Chen, R. Liu, and J. Wu, "A spectral clustering-based adaptive hybrid multi-objective harmony search algorithm for community detection," in *Evolutionary Computation (CEC), 2012 IEEE Congress on*. IEEE, 2012, pp. 1–8.

[22] H. Menéndez, D. F. Barrero, and D. Camacho, "A multi-objective genetic graph-based clustering algorithm with memory optimization," in *2013 IEEE Conference on Evolutionary Computation*, vol. 1, June 20-23 2013, pp. 3174–3181.

[23] A. A. Freitas, "A review of evolutionary algorithms for data mining," in *In: Soft Computing for Knowledge Discovery and Data Mining*, 2007, pp. 61–93.

[24] Luciano, F. A. Rodrigues, G. Travieso, and V. P. R. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, Aug. 2006.

[25] T. Washio and H. Motoda, "State of the art of graph-based data mining," *SIGKDD Explor. Newsl.*, vol. 5, no. 1, pp. 59–68, Jul. 2003.

[26] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.

[27] C. Pizzuti, "Ga-net: A genetic algorithm for community detection in social networks," in *Parallel Problem Solving from Nature–PPSN X*. Springer, 2008, pp. 1081–1090.

[28] J. Liu, W. Zhong, H. A. Abbass, and D. G. Green, "Separated and overlapping community detection in complex networks using multiobjective evolutionary algorithms," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010, pp. 1–7.

[29] E. Hruschka, R. Campello, A. Freitas, and A. de Carvalho, "A survey of evolutionary algorithms for clustering," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 39, no. 2, pp. 133 –155, march 2009.

[30] M. Oussalah and S. Nefti, "On the use of divergence distance in fuzzy clustering," *Fuzzy Optimization and Decision Making*, vol. 7, pp. 147–167, June 2008.

[31] S. Bandyopadhyay, "Genetic algorithms for clustering and fuzzy clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 6, pp. 524–531, 2011.

[32] J. Liu and W. Xie, "A genetics-based approach to fuzzy clustering," in *Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE International Conference on*, vol. 4, 1995, pp. 2233–2240 vol.4.

[33] G. Yair, "Unite unite europe' the political and cultural structures of europe as reflected in the eurovision song contest," *Social Networks*, vol. 17, no. 2, pp. 147–161, 1995.

[34] A. Ochoa Ortíz, A. E. Muñoz Zavala, and A. Hernández Aguirre, "A hybrid system using pso and data mining for determining the ranking of a new participant in eurovision," in *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, ser. GECCO '08. New York, NY, USA: ACM, 2008, pp. 1713–1714.

[35] V. Ginsburgh and A. Noury, "The eurovision song contest. is voting political or cultural?" *European Journal of Political Economy*, vol. 24, no. 1, pp. 41–52, 2008.

[36] D. Fenn, O. Suleman, J. Efstathiou, and N. Johnson, "How does europe make its mind up? connections, cliques, and compatibility between countries in the eurovision song contest," *Physica A: Statistical Mechanics and its Applications*, vol. 360, no. 2, pp. 576–598, February 2005.

[37] D. Gatherer, "Comparison of eurovision song contest simulation with actual results reveals shifting patterns of collusive voting alliances," *Journal of Artificial Societies and Social Simulation*, vol. 9, no. 2, p. 1, 2006.

[38] "Eurovision song contest," 2011, http://www.eurovision.tv.

[39] G. Bello, H. Menéndez, and D. Camacho, "Using the clustering coefficient to guide a genetic-based communities finding algorithm," in *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, ser. Lecture Notes in Computer Science, H. Yin, W. Wang, and V. Rayward-Smith, Eds., vol. 6936. Springer Berlin / Heidelberg, 2011, pp. 160–169.

[40] I. Derényi, G. Palla, and T. Vicsek, "Clique Percolation in Random Networks," *Physical Review Letters*, vol. 94, no. 16, pp. 160 202–1 – 160 202–4, Apr 2005.