Community Detection in Social Networks with Genetic Algorithms

Clara Pizzuti
ICAR-CNR
Via Pietro Bucci, 41C
87036 Rende (CS), Italy
pizzuti@icar.cnr.it

ABSTRACT

A new genetic algorithm to detect communities in social networks is presented. The algorithm uses a fitness function able to identify groups of nodes in the network having dense intra-connections, and sparse inter-connections. The variation operators employed are suitably adapted to take into account the actual links among the nodes. These modified operators makes the method efficient because the space of possible solutions is sensibly reduced. Experiments on a real life network show the capability of the method to successfully identify the network structure.

Categories and Subject Descriptors

H.2.8 [**Database Managment**]: Database Applications — *Data Mining*; I.2.2 [**Artificial Intelligence**]: Automatic Programming; I.5.3 [**Computing Methodologies**]: Pattern Recognition—*Clustering*

General Terms

Algorithms

Keywords

Genetic Algorithms, Data Mining, Clustering, Social Networks.

1. INTRODUCTION

The study of networks is an active research topic because of their capability of modelling many real world complex systems. Collaboration networks, the Internet, the world-wide-web, biological networks, social networks are just some examples. An interesting property to investigate, typical to many networks, is the *community structure*, i.e. the division of networks into groups (also called clusters) having dense intra-connections, and sparse inter-connections [1, 6, 2, 8, 3, 5].

A social network $\mathcal{SN}$ can be modelled as a graph $G = (V, E)$ where V is a set of objects, called nodes or vertices, and E is a set of links, called edges, that connect two elements of V. The problem of detecting k communities in a network, where the number k is unknown, can be formulated as finding a partitioning of the nodes in k subsets having a high density of edges within them and a lower density of edges between groups.

In this paper we propose an algorithm to discover communities in networks by employing genetic algorithms. The approach defines a quality metric of a network partitioning in communities based on the number and topology of the links present among the nodes constituting a community, and tries to optimize this quantity by running the genetic algorithm. The algorithm uses a graph-based representation [7] for the individuals of the population in which a chromosome consists of N genes, each gene can assume allele values j in the range $\{1, \ldots, N\}$. Genes and alleles represent nodes of the graph $G = (V, E)$ modelling a social network $\mathcal{SN}$, and a value j assigned to the ith gene 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. A decoding step, however, is necessary to identify all the components of the corresponding graph. The nodes participating to the same components are assigned to one cluster. A main advantage of this representation is that the number k of clusters is automatically determined by the number of components contained in an individual and determined by the decoding step.

All the dense communities present in the network structure are obtained at the end of the algorithm by selectively exploring the search space, without the need to know in advance the exact number of groups. Specialized variation operators allow to reduce the space of the possible solutions thus improving the convergence of the method. Experiments on a real life network show the capability of the genetic approach to correctly detect communities with results comparable to the state-of-the-art approaches.

2. EXPERIMENTAL RESULTS

In this section we study the effectiveness of our approach on a real-world network, *American College Football*, for which the partitioning in communities is known, and compare our results with those reported by Girvan and Newman in [4].

The *American College Football* network used for testing the method 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 in conferences. The teams on average played 4 inter-conference matches and 7 intra-conference matches, thus teams tend to play between members of the same conference. The network consists of

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

Table 1: Results obtained by our method and Girvan and Newman's algorithm for the American College Football network.

Conference	Num. Teams	Num. Correct grouping	avg num of misclassified teams	GN results
Atlantic Coast	9	10/10	-	ok
Big East	8	8/10	1	ok
Big Ten	11	6/10	1.5	ok
Big Twelve	12	10/10	-	ok
Conference USA	10	2/10	2.6	1
Independents	5	0/10	3.5	5
Mid-American	13	7/10	5.5	6
Mountain West	8	6/10	1	ok
Pacific Ten	10	9/10	1	ok
Southeastern	12	6/10	3	ok
Sunbelt	7	0/10	3.5	3
Western Athletic	10	0/10	2	2

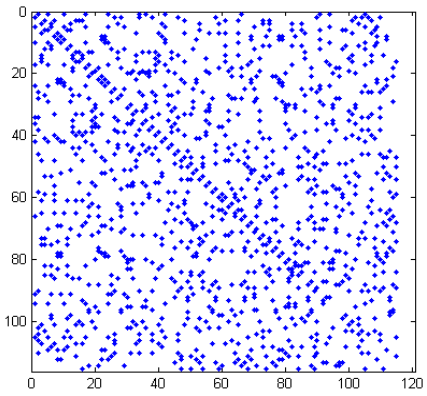

Figure 1: American College Football connections.

115 nodes and 616 edges. Figure 1 displays the connections between the 115 teams, i.e. the adjacency matrix. The figure points out the rather complex structure of the links. The application of the genetic algorithm to this network produced very good results. We run the method 10 times and in table 1 we report, besides the name of the conference and the number of teams it is composed, the number of times the algorithm successfully identified the correct grouping. For example, our approach failed to find the Big East conference in two runs over the 10 executed. In these two runs only one team was assigned to the wrong group. On the table it is also reported the average number of misplaced teams when the community found is not exact. The same information regarding the Girvan and Newman's algorithm appears. In particular "ok" means that the group found is the true group, the integer reported, instead, means the number of teams the algorithm failed to assign to the correct community. The table shows that, over the 10 runs, *GA-Net* was not able to correctly group the teams of three conferences, namely Independets, Sunbelt and Western Athetlic. However, neither Girvan and Newmans's algorithm found them. Indeed this algorithm misplaced also one team in Conference Usa and split Mid-American in two groups. For Mid-American, our approach found the right grouping seven runs over ten. In [4] the authors note that in these cases the failure is due to the fact that the conference structure is not maintained because there is not a remarkable difference in the scheduling of games. The results obtained show the capability of genetic algorithms to effectively deal with community identification in networks.

3. REFERENCES

[1] U. Brandes, M. Gaertler, and D. Wagner. Experiments on graph clustering algorithms. In *Algorithms: ESA 2003, 11th Annual European Symposium*, pages 568–579, 2003.

[2] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.

[3] L. Danon, J. Duch, A. Arenas, and A. Díaz-Guilera. Community structure identification. *Large Scale Structure and Dynamics of Complex Networks: From Information Technology to Finance and Natural Science, World Scientific,*, pages 93–113, 2007.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. In *Proc. National. Academy of Science. USA 99*, pages 7821–7826, 2002.

[5] S. Lozano, J. Duch, and A. Arenas. Analysis of large social datasets by community detection. *European Physical Journal ST*, 143:257–259, 2007.

[6] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[7] Y.J. Park and M.S. Song. A genetic algorithm for clustering problems. In *Proc. of 3rd Annual Conference on Genetic Algorithms*, pages 2–9, 1989.

[8] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proc. Natl. Acad.Sci. USA (PNAS'04)*, 101(9):2658–2663, 2004.