Proceedings of the 4th Polish Symposium on Econo- and Sociophysics, Rzeszów, Poland, May 7–9, 2009

Genetic Algorithms Approach to Community Detection

P. Mazur[a], K. Zmarzłowski[a] and A.J. Orłowski[a,b]

[a]Katedra Informatyki SGGW, Nowoursynowska 166, 02-787 Warszawa, Poland

[b]Instytut Fizyki PAN, al. Lotników 32/46, 02-668 Warszawa, Poland

The so-called community detection problem is investigated within a framework of graph theory. Genetic algorithms approach is applied to the task of identifying possible communities. Results obtained for two different fitness functions are presented and compared to each other.

PACS numbers: 89.65.Gh, 89.65.Ef, 02.50.−r, 89.75.Fb

1. Introduction

We are surrounded by, immersed in, and have to deal regularly with many complex networks of various kinds and different origins. It is enough to mention computer or information networks, biological networks, and social networks of many types. No wonder that studies of complex networks focus so much attention of scientific communities of different subjects and backgrounds [1–4].

Networks can be, and in fact are, effectively modeled within graph theory, although such fundamental concepts as vertices (nodes) and edges (links) have to be always defined anew within the realm of particular application areas we are investigating. When studying such systems we are interested in both the structure of a network and in dynamical processes [5]. One of the important problems in studies of complex networks, especially if we deal with social networks, is finding out some (if any) underlying sub-structures. Searching for such a special group on nodes that, roughly speaking, has more connections inside itself than with the rest of the network, is often called the community detection problem. This problem has been extensively investigated over the last few years and many algorithms of different kinds and levels of sophistication have been developed, criticized, tested, and applied to various situations [6–11].

In this paper we focus our attention on one of possible approaches, namely on genetic algorithms. They form a set of procedures based on natural selection mechanisms and genetics and aim at finding exact or approximately solutions to optimization problems [12]. There are many versions of genetic algorithms developed for the task of community detection and here we concentrate on a very promising one proposed quite recently by Pizzuti [13].

2. Genetic algorithm and fitness functions

As any of different genetic algorithms, also that designed in [13] consists the initialization stage when the base population (set of solutions represented by chromosomes) is created and then performs several stages repeated cyclically (crossover, mutation) to find the best solution. Of course, to make a good use of such algorithms we need a special function (fitness function) which defines the quality of obtained solutions. The search procedure ends after the definite number of steps or after obtainment suitable value of the quality function.

Pizzuti's algorithm uses locus-based adjacency representation of the chromosome, i.e., the chromosome is represented as a vector of the length equal to the number of all nodes in the graph. Every index of this vector represents a specific node, and a value shown by this index is a vertex to which there exists a connection from the index node in the original graph. As a quality function Pizzuti has chosen a well-known parameter called a community score

$$CS = \sum_{k=1}^{m} \left[\frac{\sum_{i \in I_{S_k}} \left(a_{iJ_{S_k}} \right)^r}{|I_{S_k}|} \sum_{i \in I_{S_k},\ j \in J_{S_k}} (a_{ij}) \right].$$

Here m is the number of communities and $r = 0.5$. In our case I_{S_k} is equal to J_{S_k}, because we use only undirected graphs. Thus for partition k we should create sub-matrix $S(I_{S_k}, J_{S_k})$ where I_{S_k} is subset of rows of adjacency matrix corresponding to nodes belonging to community k. Moreover a_{ij} is a value obtained from sub-matrix S_k and $a_{iJ_{S_k}}$ is a sum of values from row i of sub-matrix S_k. In this work we compare results of Pizzuti's algorithm for two different fitness functions: originally used community score and a modularity function, introduced in [14]:

$$Q = \sum_{k=1}^{m} \left[\frac{l_k}{L} - \left(\frac{d_k}{2L} \right)^2 \right].$$

Here m is a number of communities, l_k is a number of links joining vertices inside the community k, d_k is a sum of degrees of the nodes in the community k, and L is a total number of links in the investigated network.

As a measure of similarity between the partitions we use normalized mutual information (NMI):

$$I(A,B) = \frac{-2\sum_{i=1}^{c_A}\sum_{j=1}^{c_B}\left(c_{ij}\log\left(c_{ij}N/c_{i\bullet}c_{\bullet j}\right)\right)}{\sum_{i=1}^{c_A} c_{i\bullet}\log\left(c_{i\bullet}/N\right)+\sum_{j=1}^{c_B} c_{\bullet j}\log\left(c_{\bullet j}/N\right)}$$

where A, B are partitions to compare, C is a matrix in which at the position (i,j) occurs number of nodes being both in the community i of the partition A and in the community j of the partition B, c_A and c_B are numbers of communities in partition A and B, respectively, $c_{i\bullet}$ ($c_{\bullet j}$) are the sums of the elements of the matrix C in row i (column j), and N is the total number of nodes.

3. Results

In all tests performed in this section, we use the same standard set of parameters for our genetic algorithm (crossover rate 0.8, mutation rate 0.2, and elite reproduction 10%) as in [13]. The population size is equal to 300 and the number of generations is restricted to 100.

We first test Pizzuti's algorithm, for two fitness functions, on the well-known Zachary's karate club network of acquaintance relationship between 34 members [15]. In both cases the algorithm finds 4 communities, which are subgroups of the real structure of this network as shown in Fig. 1. For community score we obtained NMI = 0.71 and for modularity NMI = 0.69. This difference is due to the method of calculation of NMI and we cannot clearly identify which partition is really better.

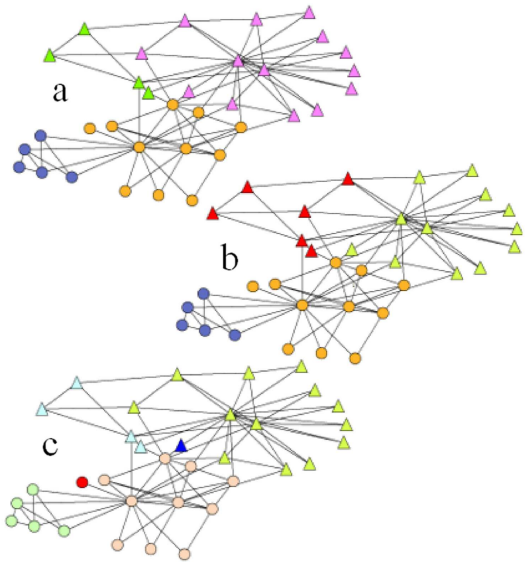

Fig. 1. Zachary's karate club network. Nodes are colored according to communities found by (a) Pizzuti algorithm using community score, (b) Pizzuti algorithm using modularity, (c) Pizzuti algorithm using community score and with the possibility of detecting one-node communities.

The next example is a network of transcontinental airline connections between Africa, Asia and North America. It consists of 140 countries and 1997 undirected edges as shown in Fig. 2. This network is created by adding a link between the two countries if and only if there was an airplane connection between them in April 7, 2009. We run algorithm 10 times for both fitness functions. The average normalized mutual information for community score is 0.726 (max = 0.799 and min = 0.68) and for modularity Q it is 0.84 (max = 0.924 and min = 0.8). In both cases, for maximal values of NMI we obtained lower values of the fitness function than for minimal values of NMI.

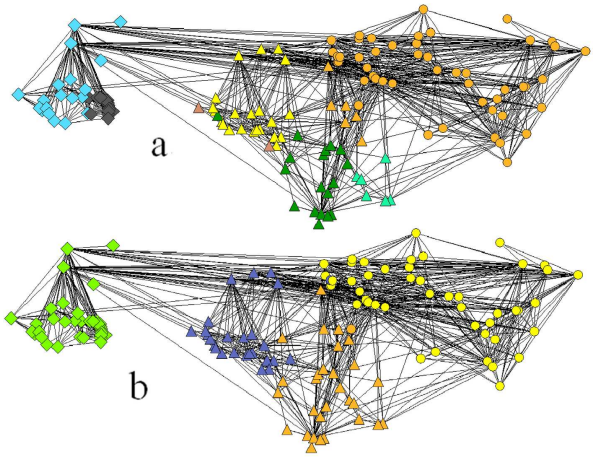

Fig. 2. Airline connection network between Africa, Asia and North America. Nodes are colored according to communities found by (a) Pizzuti algorithm using community score (fitness function = 918.67), (b) Pizzuti algorithm using modularity (fitness function = 0.4937).

In networks with a small number of nodes as exemplified in Fig. 3 a modification of Pizzuti's algorithm works better. As seen in Fig. 3a and b, the original algorithm (fitness function being the community score) finds only 4 communities. If we allow for the detection of one-node communities this algorithm is able to find a richer structure (i.e., more communities) as shown in Fig. 3c and d. The same happens for our first example of Zachary's karate club network, where with such a modified algorithm we find two more communities (fitness function being the community score) as shown in Fig. 1c.

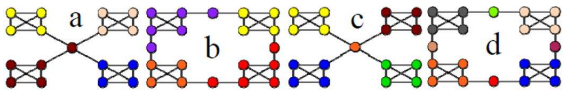

Fig. 3. Communities found by original Pizzuti algorithm using community score (a) and (b) and by modified Pizzuti algorithm using community score and with the possibility of detecting one-node communities (c) and (d).

4. Brief summary

Community detection problem, especially in social networks, is a very important one and finds a lot of interesting applications. Genetic algorithms approach seems to

be quite effective in addressing this issue. In this paper we have utilized a recently proposed genetic algorithm, accompanied by two different fitness functions being used in the context of finding out community structure: community score and modularity. We have also proposed a small modification of the original algorithm allowing for isolated single-node communities to be detected. Such a modification reasonably improves the revealed community structure for small networks. Unfortunately, in the case of large networks this is not a good strategy, as it increases the amount of possible solutions and hence may even worsen the performance of the genetic algorithm. We plan to further refine our approach and to apply it to more complicated real-life dynamical networks in forthcoming papers.

Acknowledgments

Work of one of the authors (AJO) was supported in part by PAN/CNRS Project PICS No. 4339 (2008-2010).

References

[1] *Analysis of Complex Networks. From Biology to Linguistics*, Eds. M. Dehmer, F. Emmert-Streib, Wiley-VCH, Weinheim 2009.

[2] *Handbook of Graphs and Networks. From the Genome to the Internet*, Eds. S. Bornholdt, H.G. Schuster, Wiley-VCH, Weinheim 2003.

[3] S.N. Dorogovtsev, J.F.F. Mendes, *Evolution of networks. From Biological Nets to the Internet and www*, Oxford University Press, Oxford 2003.

[4] A.L. Barabasi, R. Albert, *Rev. Mod. Phys.* **74**, 47 (2002).

[5] A. Barrat, M. Barthelemy, A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, Cambridge 2008.

[6] M. Girvan, M.E.J. Newman, *Proc. Natl. Acad. Sci. USA* **99**, 7821 (2002).

[7] S. Fortunato, *Phys. Rep.* **486**, 75 (2010).

[8] M.E.J. Newman, *Eur. Phys. J. B* **38**, 321 (2004).

[9] M.E.J. Newman, *Phys. Rev. E* **69**, 066133 (2004).

[10] S. Fortunato, M. Barthelemy, *Proc. Natl. Acad. Sci. USA* **104**, 36 (2007).

[11] A. Lancichinetti, S. Fortunato, F. Radicchi, *Phys. Rev. E* **78**, 046110 (2008).

[12] D.E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley, New York 1989.

[13] C. Pizzuti, *GA-Net: A Genetic Algorithm for Community Detection in Social Networks*, in: *Lecture Notes in Computer Sciences*, LNCS 5189, Springer Verlag, Berlin 2008, p. 1081.

[14] M.E.J. Newman, M. Girvan, *Phys. Rev. E* **69**, 026113 (2004).

[15] W.W. Zachary, *J. Anthropol. Res.* **33**, 452 (1977).