Heuristic algorithm for determination of local properties of scale-free networks

Marija Mitrović, Aleksandar Belić

Institute of Physics, Pregrevica 118, Belgrade, Serbia

Abstract. A large number of complex networks, both natural and artificial, share the presence of highly heterogeneous, scale-free degree distributions. We investigate if two networks are similar or not by examining their local characteristics. The local structure of networks is measured by counting the frequency of subgraphs of a given type. The exhaustive search of sub graphs in a large network is computationally prohibitively expensive. For this reason the World Wide Web, social networks, and biological networks of higher organisms, require the development of efficient heuristic approaches allowing counting of sub graphs by an incomplete search.

Here we report on the comparison of efficiencies of two heuristic algorithms, based on random and preferential choices of the subset of studied nodes. Both were applied to scale-free networks, generated using the Barabasi-Albert algorithm, to count sub graphs with up to four nodes. The random method, though simpler, is found to be inaccurate.

1 INTRODUCTION

Complex Networks are everywhere[1,2]. Many phenomena in nature can be modeled as networks, such as brain structures, protein-protein interaction networks, social interactions, written human language network [4] and the Internet and WWW [3]. All such systems can be represented in terms of nodes and edges indicating connections between nodes. For example in science collaboration network nodes represent individual scientists which are connected if two of them have written an article together.

An important characteristic of these networks is that they are not random, but have a rather more structured architecture. The structures of different networks are very similar. They all have power law degree distribution, i.e. they are scale-free. Further more despite the large size of these networks there is usually a relatively short path between any two nodes. This so called small world property is the best exemplified by the well known 6-degree of separation property of the human friendship network.

Degree distribution, clustering coefficient and the diameter of the network are global characteristics of the network. The local structure of networks is measured by counting the frequency of subgraphs of a given type.

2 MOTIVATION

Global properties of both artificial and natural networks have been extensively studied. Similarity of two networks is also subject of research. We investigate if two networks are similar or not by examining their local characteristics. Finding subgraphs exhaustively

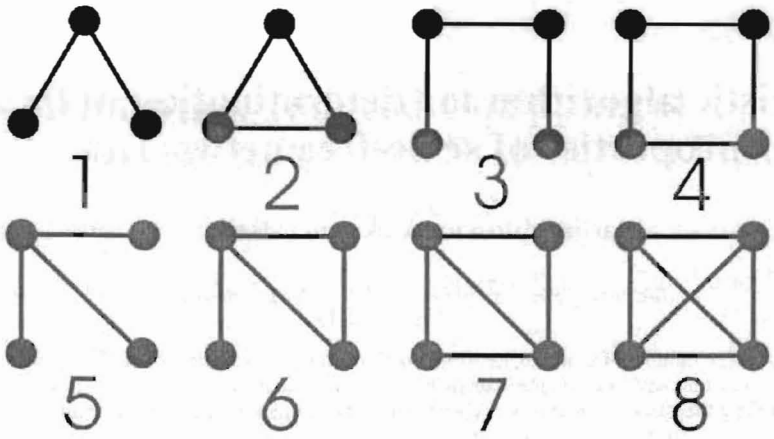

FIGURE 1. Type of subgraphs up to order four

in a large network is computationally intensive. The presence of large number of nodes in real-world networks, such as IT, social and biological networks of higher organisms, demands developing of heuristic algorithms.

3 METHODS

3.1 Definitions

Networks are represented by graphs. A graph is denoted by G, or $G(V,E)$, where V is a set of nodes and $E \subseteq V$xV is set of edges of G. The number of nodes $|V|$ is denoted by n. If there is an edge between nodes u and v, than we say that v is a first neighbor of node u, and vice versa. The degree of the node, k_i is the number of his first neighbors.

A subgraph $H(V_1,E_1)$ of G is a graph whose nodes and edges belong to G,i.e. $V_1 \subset V$ and $E_1 \subset E$. Order of subgraph H, k, is defined by number of nodes of subset V_1.There are different types of subgraphs of the same order. For example, for order three there are two different types, while for order four there are six of them (Figure 1).

3.2 Model Networks

We generate networks of interacting nodes using preferential attachment (SF) algorithm, introduced by Albert and Barabasi [1]. These networks have power-law degree distribution, which is well known property of many real world networks. The algorithm of the SF model is the following:

(1) *Growth*: Starting with a small number n_0 of nodes, at every time step we add a new node with $m(\leq n_0)$ edges that link the new node to m different nodes already present in the system.

(2) *Preferential attachment*: When choosing the nodes to which the new node connects, we assume that the probability Π that a new node will be connected to node i depends on the degree k_i of node i, such that $\Pi(k_i) = \frac{k_i}{\sum_j k_j}$

After t time steps this algorithm results in a network with $n = t + n_0$ nodes.

3.3 Algorithms

We counted the subgraphs of order up to four using two approaches, random and preferential. In both methods we chose a subset $S \subset V$ of nodes. We count all the subgraphs of order three and four that contain at least one node from the subset. The way of choosing the nodes in subset is different for this two approaches. Random method amounts to choosing nodes at random, with equal probability. It is simpler and faster, but much less accurate, since ut to 80% of the nodes need to be included in the search in order to get reasonably good accuracy in estimating the number of subgraphs.

In order to improve the algorithm we need to use the knowledge about global properties of the network. Scale-free networks have power-law degree distribution, which means that there are few nodes in network with high degree, while most of other nodes have just m first neighbors. This means that popular nodes are parts of most of the subgraphs. If we chose node with probability which is proportional to its degree $\Pi(k_i)$, than we need to consider less nodes to estimate the number of subgraphs since we will account all of the most important ones. We will spend less time on less important nodes - an idea closely related to the importance sampling approach in Monte Carlo simulations. Such preferential algorithm is much faster than random and exact search. In Figure 2 it is shown that we need typically to search only 20% of the nodes to get an accurate estimate of the number of subgraphs.

In order to compare these two algorithms we defined, for every type of subgraphs, ratios $a_i = \frac{P_i}{N_i}$ (N_i is a number of subgraphs of type i obtained by the exact counting, and P_i number of subgraphs found by heuristic search algorithm). These ratios are functions of the fraction of searched nodes q.

4 RESULTS AND CONCLUSIONS

Here we report on the comparison of efficiencies of two heuristic algorithms, based on random and preferential choices of the subset of studied nodes. Both were applied to scale-free networks, generated using the Barabasi-Albert algorithm, to count subgraphs with up to four nodes. The random method, though simpler, is found to be slower, which was expected. We used the fact that SF networks have hubs and derived more faster heuristic algorithm for determination of local properties of complex networks.

One typical case is shown in Figure 2, where the two algorithms were applied to scale-free network with n=15 000 nodes. It can be seen that preferential algorithm essentially converged to the exact value obtained by exhaustive count already after 20% of the

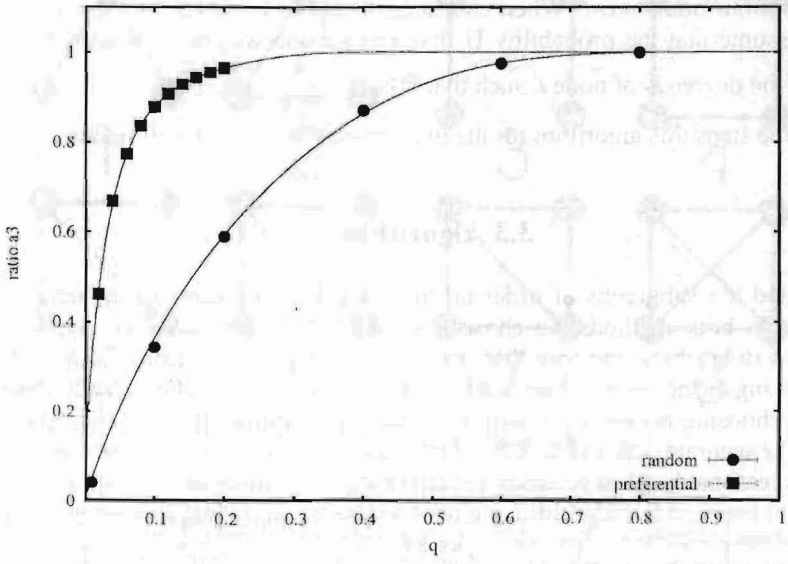

FIGURE 2. Dependence of the ratio a_3 on the fraction of the searched nodes q, for preferential and random algorithms for scale-free networks of size n=15 000. The data is obtained by averaging over 100 different networks. The error bars are smaller than the symbols used.

nodes were considered. On the other hand, in order to get an accurate estimate using random algorithm, full 80% of the nodes needed to be searched, which is not a significant improvement over exhaustive count.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Science of the Republic of Serbia, under project no OI141035. Marija Mitrović has received support under FP6 project CX-CMS. The presented numerical results were obtained on the AEGIS e-Infrastructure whose operation is supported in part by FP6 projects, EGEE-II and SEE-GRID-2.

REFERENCES

1. A. Barabasi A.-L., R. Albert, Reviews of Modern Physics, (2002) vol. 74, Issue 1, 47-97
2. N. Przulj, G. Corneil, and I. Jurisica, Bioinformatics (2006), vol. 22, number 8, 974-980
3. Bosiljka Tadić, Computer Physics Communications 147 (2002), 586-589
4. A.P. Masucci and G.J. Rodgers, Physical Review E 74 (2006), 026102