Physica A 482 (2017) 218–227

Contents lists available at ScienceDirect

Physica A

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

A simple and efficient algorithm for modeling modular complex networks

Mateusz Kowalczyk, Piotr Fronczak *, Agata Fronczak

Faculty of Physics, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland

HIGHLIGHTS

- Fast method for generating networks with community structure.
- Quality demonstrated analytically and supported by numeric examples.
- Heterogeneity in node degrees and community sizes.

ARTICLE INFO

Article history:
Received 31 August 2016
Received in revised form 15 January 2017
Available online 26 April 2017

Keywords:
Complex networks
Community structure
Algorithms

ABSTRACT

In this paper we introduce a new algorithm to generate networks in which node degrees and community sizes can follow any arbitrary distribution. We compare the quality and efficiency of the proposed algorithm and the well-known algorithm by Lancichinetti et al. In contrast to the later one, the new algorithm, at the cost of accuracy, allows to generate two orders of magnitude larger networks in a reasonable time and it can be easily described analytically.

1. Introduction

Community structure is considered to be, next to the small-world effect and scale-free degree distribution, one of the most important topological properties of real networks. By community (also called cluster, module, or block) in a network we understand a group of nodes more densely connected to each other than to nodes outside the group. For example, in social networks, communities correspond to groups of people sharing the same interests [1], in the Internet, they consist of the sets of web pages on the same topic [2], while in cellular and metabolic networks, communities are functional modules of interacting proteins [3].

In the science of complex networks, community detection has become one of the most dominant research topics over the last decade. As a consequence, a large number of algorithms have been proposed for the analysis of community structure in network [4–7]. To evaluate these algorithms efficiently, synthetic networks with a well-defined community structure (benchmarks) had to be proposed. The advantage of such models is that, unlike in real networks, one can easily vary the model parameters and compare the recovered community structure with the predefined one.

One of the first models of networks with community structure, with a long tradition of study in the social sciences and computer science [8–13], is the so-called blockmodel. In its classical version [8], each of N nodes is assigned to one of K blocks (communities) of equal size, and undirected edges are independently drawn between pairs of nodes with probabilities that

* Corresponding author.
E-mail address: fronczak@if.pw.edu.pl (P. Fronczak).

http://dx.doi.org/10.1016/j.physa.2017.04.111
0378-4371/

are a function only of the group membership of the nodes. Unfortunately, the Poisson-like degree distribution makes this model unsuitable for the further analysis, since most of real networks exhibit power laws in their degree distributions.

Lancichinetti et al. [14] proposed an efficient numerical construction procedure for benchmark graphs that is free of this defect. The method accounts for the heterogeneity in the distributions of node degrees as well as community sizes. Its efficiency has been tested and proved in typical cases, however, further in this paper we show that in a certain range of parameters efficiency of the algorithm drops significantly. Moreover, the complexity of the proposed procedure does not allow for the analytic tractability.

In contrast to Lancichinetti's procedure, Fronczak et al. [15] provided an exponential random graph formulation [16–20] for blockmodel that is solvable for its parameter values in closed forms. Two kinds of the network structural Hamiltonians have been considered: the first one corresponding to the classical blockmodel, and the second one corresponding to its degree-corrected version. In both cases, a number of analytical predictions about various network properties was given. In particular, it was shown that in the degree-corrected blockmodel, node degrees display an interesting scaling property, that is similar to the scaling feature of the node degrees in fractal (self-similar) real-world networks. Unfortunately, the method is computationally inefficient since it is based on Markov chain Monte Carlo algorithm.

In this contribution we propose a simple, analytically tractable, and fast algorithm for generation of networks with community structure and heterogeneity of both node degrees and community sizes. The method allows to generate, in a reasonable time, networks that are orders of magnitude larger than those generated by the previous approaches. It also allows for closed-form parameter solutions.

In outline, the paper is as follows. First, we introduce a new method (KA; the meaning of this abbreviation is "Kowalczyk's et al. algorithm") for generating clustered networks and derive their main properties. Next, we review Lancichinetti's algorithm (LA). We describe its sub-procedures and their time complexity. This allows us to point the range of parameters for which the algorithm efficiency drastically drops down. Finally, we discuss all the major pros and cons of the both approaches. In the Appendix, we provide detailed listings of the both algorithms.

2. Derivation of the new algorithm

In this section, we present a simple algorithm to generate networks with community structure, which, despite its simplicity, has not been considered, at least to our knowledge, in previous studies. The algorithm is an extension of the model for generating uncorrelated networks with a given sequence of expected degrees $\{\langle k_1\rangle, \langle k_2\rangle, \ldots, \langle k_N\rangle\}$ (see eg. Eq. (15) in [21] and Eq. (48) in [17]). In such a prototype network, there is at most one link between any pair of nodes, and there are no self-loops connecting nodes to themselves. If a_{ij} is an entry of the adjacency matrix underlying the network, and $a_{ij} \in \{0, 1\}$, where $a_{ij} = a_{ji}$ and $a_{ii} = 0$, then the expected value of the entry, $\langle a_{ij}\rangle$, can be expressed in terms of the probability, p_{ij}, that the vertices i and j are connected, namely

$$\langle a_{ij}\rangle = 1 \cdot p_{ij} + 0 \cdot (1 - p_{ij}) = p_{ij}. \tag{1}$$

Simultaneously, given the expected node degrees, the average number of connections, which obviously cannot be greater than one, may be estimated as the expected number of successes in $\langle k_i\rangle$ attempts of i to connect to j, where the probability of success for one trial is $\langle k_j\rangle(\sum_{j\neq i}\langle k_j\rangle)^{-1}$, i.e.

$$\langle a_{ij}\rangle = \langle k_i\rangle \frac{\langle k_j\rangle}{\sum_{j=1}^{N}\langle k_j\rangle - \langle k_i\rangle} \simeq \frac{\langle k_i\rangle\langle k_j\rangle}{\langle k\rangle N}. \tag{2}$$

By comparing Eqs. (1) and (2), one gets a simple expression for the probability of a connection:

$$p_{ij} = \frac{\langle k_i\rangle\langle k_j\rangle}{\langle k\rangle N}. \tag{3}$$

In analogy to the above derivation, in networks with community structure, one can write similar relations for the probabilities p_{ij}^{int} and p_{ij}^{ext}, that there is an internal or external connection between two nodes, i and j, belonging to the same or to different communities. If it is not clear, let us explain that internal connections are those that are between nodes belonging to the same community. Accordingly, the external connections are those that are between nodes belonging to different clusters.

Thus, let $\langle k_{i,r}^{int}\rangle$ represent the expected internal degree of a node i belonging to the rth community. Correspondingly, let $\langle k_{i,r}^{ext}\rangle$ be the expected number of its external connections. Then:

$$p_{ij}^{int} = \langle k_{i,r}^{int}\rangle \frac{\langle k_{j,r}^{int}\rangle}{\sum_{j=1}^{c_r}\langle k_{j,r}^{int}\rangle - \langle k_{i,r}^{int}\rangle} \simeq \frac{\langle k_{i,r}^{int}\rangle\langle k_{j,r}^{int}\rangle}{2\langle E_r^{int}\rangle}, \tag{4}$$

and

$$p_{ij}^{ext} = \langle k_{i,r}^{ext}\rangle \frac{\langle k_{j,s}^{ext}\rangle}{\sum_{s\neq r}\sum_{j=1}^{c_s}\langle k_{j,s}^{ext}\rangle} \simeq \frac{\langle k_{i,r}^{ext}\rangle\langle k_{j,s}^{ext}\rangle}{2\langle E^{ext}\rangle}, \tag{5}$$