The object of this study is the models of complex networks. The task addressed is to construct a generative model of a growing network, which has two key features characteristic of real-world networks: scale-free property and homophily. Homophily of a network is understood as the tendency of nodes to group into communities. To combine the requirements for homophily and scale-free property, a two-level rule of preferential attachment has been devised. First, the color of a new node is chosen with a probability proportional to the volume of the corresponding community, and then, according to the usual rule of preferential attachment, neighboring nodes are chosen within the community. It has been shown that the computational complexity of generating a network model with n nodes is O(n). It has been proven that under such conditions, the distribution of degrees of nodes across the entire network is the same as in the classical non-homophilic Barabási-Albert network and does not depend on the number and structure of communities. Under the conditions of homophily, it is quite natural to generalize the requirement for scale-free property to the distribution of community sizes. It has been found that this distribution is determined by the intensity of the formation of new communities. The dependence of the expected time interval between the formations of successive communities on their index has been established. The measure of homophily of the generated networks – modularity – has been estimated; its dependence on the scaling of the community volumes was found. The model built allows step-by-step generation of growing scale-free networks that have a built-in mechanism for the formation of communities, which is of practical significance. Moreover, the proposed model could also be used in the opposite direction: given the structural parameters of the network, it is possible to restore the hidden rules by which this network evolves

Keywords: scale-free property, homophily, modularity, generative network models, node communities

UDC 004.03
DOI: 10.15587/1729-4061.2025.326092

CREATION OF THE GENERATIVE MODEL OF A SCALE-FREE NETWORK WITH HOMOPHILIC STRUCTURE

Vadim Shergin
PhD, Associate Professor*
Sergei Grinyov*
Larysa Chala
PhD, Associate Professor*
Serhii Udovenko
Corresponding author
Doctor of Technical Scienses, Professor
Department of Informatics and Computer Engineering
Semyon Kuznets Kharkiv National University of Economics
Nauky ave., 9a, Kharkiv, Ukraine, 61165
E-mail: serhiy.udovenko@hneu.net
*Department of Artificial Intelligence
Kharkiv National University of Radio Electronics
Nauky ave., 14, Kharkiv, Ukraine, 61166

Received 15.01.2025
Received in revised form 05.03.2025
Accepted 26.03.2025
Published 30.04.2025

How to Cite: *Shergin, V., Grinyov, S., Chala, L., Udovenko, S. (2025). Creation of the generative model of a scale-free network with homophilic structure. Eastern-European Journal of Enterprise Technologies, 2 (4 (134)), 14–22. https://doi.org/10.15587/1729-4061.2025.326092*

1. Introduction

One of the main problems in the theory of complex networks is the construction of a structural theory of networks [1]. This theory should establish a clear connection between statistical properties and the laws of network evolution. Direct problems (i.e., establishing the properties of networks that are formed according to given rules) are more or less successfully solved, while inverse problems, i.e., the formation of rules, the application of which makes it possible to obtain a network with the predefined properties, are much more complex and cover only a limited set of structural properties.

Since the basic property of complex networks is the principle of growth, it is even more relevant to devise evolutionary rules for network formation, i.e., those that make it possible to generate growing networks in which the necessary structural properties are supported. It is such network models that are generative.

One of the most common properties of real-world networks is a scale-free property (or scale invariance) [2] and heterogeneity of the structure: the presence of communities [3], i.e., homophily [4]. There is currently no model that would generate a growing network with given scaling and homophily parameters, so its construction is relevant from a theoretical perspective. However, the relevance of creating the specified model relates not only to achieving the ability to generate networks with the predefined properties (which in itself has practical meaning). Much more significant is the ability to apply the generative model in the opposite direction: using given structural parameters (scaling and homophily) of a real-world network, to restore the hidden rules by which the evolution of this network is taking place or has taken place.

Therefore, it is a relevant task to carry out studies aimed at creating a generative model of a scale-free network with a homophilic structure.

2. Literature review and problem statement

The mathematical tools of networks (graphs) is a widely used means of describing many objects and systems in science, society, nature, and technology. It is evident that real-world networks have a wide variety of individual properties, qualitative and quantitative characteristics, but there are certain properties that are characteristic of most existing networks.

The main ones are randomness, growth, scale-free property, and structure. Accordingly, the problem of creating models designed to describe networks with these properties arose. One of the first and simplest random graph models was the Erdős-Rényi model [5]. However, this model is not evolutionary (generative); the corresponding graph is not scale-free and its structural characteristics (assortativity, modularity, clustering) do not correspond to those observed in most existing networks.

It is known [6] that most real networks have a power-law distribution of nodes by the number of connections. Such a distribution defined the scale-free property (scale invariance) of a network. In [7] it was shown that the natural evolutionary mechanism of such networks is the principle of preferential attachment (PA). PA models, also known as Barabási-Albert (BA) networks, have become widespread because they correspond to three important properties. First, the generated networks are random; secondly, they are scale-free; and thirdly, such models are generative (i.e., they make it possible to generate networks under a step-by-step mode, starting from a small initial seed). However, a significant drawback of this class of models is the discrepancy between the structural properties of the generated networks and the properties of real networks. Thus, the assortativity of BA networks is asymptotically zero, while social networks have a pronounced positive assortativity, while biological and technical ones have negative assortativity [8]. There are link rewiring algorithms [9] that make it possible to achieve non-zero assortativity while preserving the power distribution, but these methods work under the post-processing mode of an already created network, i.e., they are not generative. The modularity of BA networks is also asymptotically zero [10], which indicates the absence of a natural internal structure, unlike real networks that have such a structure.

Most real-world networks are not homogeneous: nodes and connections between them form more or less well-defined subnetworks, or clusters, or communities. In practice, it is very convenient to associate such communities with colors [1], although the physical nature of such communities can be arbitrary. A key manifestation of such structural heterogeneity is that nodes are more closely connected to nodes of their own color than to nodes of other colors. This property (attraction to connect with their own kind) is termed homophily [1]. The most common quantitative criterion for the clarity of the network partition into communities is modularity, proposed in [11]. The modularity of a network is easily calculated and has a clear essence: the comparison of a network with a random Erdős-Rényi graph of the same size in terms of the densities of connections within and between communities. It is worth noting that the choice of the Erdős-Rényi model as a null model significantly limits the meaning of modularity in the case when communities have significantly different sizes [12]. And this case is characteristic of scale-free networks. Therefore, in [12], a modified modularity was proposed, which directly takes into account the size of communities.

In [1, 4, 10], alternative modularity quantitative measures of homophily, i.e., the structuring of networks into communities, are considered: conductivity and entropy ratio. In [10] it was experimentally confirmed that all three measures are equivalent in determining community structures in networks. In other words, if one of them indicates the statistical significance of the detected communities, then the other two also confirm this significance. However, the indicated synchronicity of the specified homophily measures has not been mathematically proven, which does not give grounds to arbitrarily switch from one measure to another.

Over a long time, the homophily of networks was studied sicut datum est, that is, as a property of already built, existing networks. In other words, the models that explicitly or implicitly underlie the analysis of modularity or conductivity were not generative. Also, homophily was estimated and analyzed separately from the distribution of nodes by degrees. An attempt to construct a generative model of a scale-free network with a homophilic structure of communities was made in [1, 4]. The model proposed by the authors makes it possible, starting from a small seed, to step by step grow the network, adding new nodes and edges, creating new communities. However, this model is purely empirical. There is no justification for the chosen probability of the appearance of a new community $p_k=(\log k)^{-a}$ (i.e., the probability that a new node k will become the firstborn of a new community). Moreover, the distribution of nodes of the resulting network by the number of connections is not power-law (although it is somewhat similar visually), and therefore the network is not scale-free.

Thus, the task of creation a model that would step-by-step generate a scale-free network with a homophilic structure (i.e., with "built-in", "natural" communities) remains an actual.

3. The aim and objectives of the study

The aim of our research is to create a generative model of networks that would have a scale-free property and homophily, inherent in most real-world networks. This could make it possible to both directly generate networks with given scaling and homophily parameters and to analyze existing networks in the reverse direction. That is, based on the given structural parameters of the network, restore the hidden rules by which the evolution of this network took place.

To achieve the goal, the following tasks were set:

– developing an algorithm for generating a homophily network model;

– determining the rules for coloring nodes that ensure scale-free property of the network with respect to the degrees of nodes;

– defining the rules for the appearance of new colors that ensure scale-free property of the network in terms of the volume of communities;

– estimating the homophily parameters of the network and establishing their relationship with the model parameters;

– implementing the model programmatically, checking its operability and compliance of the structural properties of the networks that are generated with the given requirements.

4. The study materials and methods

The object of research in our work is generative models of complex networks. The network is represented in the form of an undirected unweighted graph $G(V,E)$. The model is generative (evolutionary) if it makes it possible to model the step-by-step growth of the network:

$$E\{x_{n+1}\} = x_n + \Delta_n, \quad (1)$$

where x is the network parameter under study, Δ_n is its expected increment at step n, $E\{\cdot\}$ is the expectation operator (usually, for brevity, it is omitted; the same will be done in this paper).

The step (n) is usually identified with the node number; thus, the generative model (1) describes the dynamics of changes in some network parameter x with an increase in the number of nodes. Most often, such a parameter is the degree of the node.

Model (1) is a first-order difference equation, its use as a general form of the generative network model is the main assumption in our study.

In order for the dependence of the expected values of the studied parameter x on n to be power-law, it is necessary that the increment Δ_n in (1) satisfies the requirement:

$$\Delta_n = x_n \cdot \frac{a}{n+b}. \tag{2}$$

Dependence (2) has zero memory depth, i.e., Δ_n depends only on the current value of x_n and does not depend on the history of the parameter x under study. This is an essential, but standard, assumption when developing network models.

From (1) and (2), the equation of the dynamics of parameter x follows:

$$x_{n+1} = x_n + x_n \cdot \frac{a}{n+b}, \tag{3}$$

whose solution under the initial condition x_k is:

$$x_n = x_k \cdot \frac{\Gamma(k+b)}{\Gamma(k+a+b)} \cdot \frac{\Gamma(n+a+b)}{\Gamma(n+b)}, \tag{4}$$

where $\Gamma(y)$ is Euler's gamma function; for natural numbers, m, $\Gamma(m+1)=m!$.

It is worth noting that dependence (4) is often simplified as power-law because $x_n \sim n^a$ as $n \to \infty$. The parameter a is called the scaling index. However, this simplification does not imply the approximate nature of (4) or the imperfection of the model (1) to (3), but the use of the terms "power-law dependence", "power-law distribution" (which are quite natural for continuous time functions $x(t)$) for discrete time sequences x_n. In this sense, it is appropriate to term the dependences of form (4) "discrete-power-law".

An important partial case of the model (1) to (4) is the well-known preferential attachment model and its simplest variant – the Barabási-Albert model (BA-model) [7], which describes the increment of the degree of a network node. It is the BA model that is proposed as the basis of the model being built. Assumptions were accepted (quite natural for PA models) that each new node is connected by $m>1$ edges to existing ones, and the seed is a complete graph K_m, all nodes of which belong to one (primary) community. In this case, the total number of connections in a network with $n>m$ nodes is:

$$\mathrm{vol}(G_n) = m(m-1) + 2m(n-m) = 2m(n-n'), \tag{5}$$

where $n' = \frac{m+1}{2}$.

According to the PA principle, the probability that the free end of a newly generated edge will be attached to node i of a network G of n nodes is proportional to the degree $deg_{i,n}$ of node i at step n:

$$\pi_{i,n} = \frac{deg_{i,n}}{\mathrm{vol}(G_n)}. \tag{6}$$

An important assumption of the BA model is that the attachment of each of the m newly generated edges is considered as independent events. In this case, taking into account (5), the expected increase in the power of node i is equal to:

$$\Delta_{i,n} = m \cdot \pi_{i,n} = m \cdot \frac{deg_{i,n}}{\mathrm{vol}(G_n)} = \frac{deg_{i,n}}{2(n-n')}. \tag{7}$$

The basic equation for the dynamics of node degree growth in the BA model takes the following form:

$$deg_{i,n+1} = deg_{i,n} + \frac{deg_{i,n}}{2(n-n')}, \tag{8}$$

under initial conditions $deg_{i,i}=m$ for $i>m$, $deg_{i,m}=m$ for $i \le m$.

According to model (5)–(8) (which is a partial case of the more general model (1)–(4)), the expected values of the degrees of the network nodes are equal to:

$$deg_{i,n} = m \cdot \frac{\Gamma(i-n') \cdot \Gamma((n-n'+1/2)}{\Gamma(i-n'+1/2) \cdot \Gamma(n-n')}, i \ge m. \tag{9}$$

Dependence (9) is discrete-power, for large i and n it approaches the usual power law $deg_{i,n} \sim \sqrt{n/i}$, that is, it has a rank scaling index $\rho=1/2$.

The suitability of BA model (5)–(9) for modeling each of the communities of a homophilic scale-free network (while preserving the scaling $\rho=1/2$) is the main hypothesis of our study.

Modularity is used as a measure of homophilicity of networks [11]. If all nodes of the network are divided into k communities (colored with k colors), then the modularity of the network is defined as:

$$\mu(G_n) = \sum_{j=1}^{k} \left(e_{j,n}^{in} - e_{j,n}^2 \right) = 1 - E_n^{ext} - \sum_{j=1}^{k} e_{j,n}^2, \tag{10}$$

where $e_{j,n}$ is the relative volume of community j (i.e., the sum of powers of nodes of color j, normalized to $\mathrm{vol}(G_n)$ – the volume of the entire network), $e_{j,n}^{in}$ is the relative number of connections between nodes within community j, $e_{j,n}^{ext}$ is the relative number of connections between nodes of the community j and nodes of other communities (it is obvious that $e_{j,n} = e_{j,n}^{in} + e_{j,n}^{ext}$), E_n^{ext} is the relative number of connections between nodes of different communities throughout the network, the "relativity" of all these volumes means their normalization to $\mathrm{vol}(G_n)$ – the volume of the network (i.e., the total number of connections).

It is obvious, that modularity (10) significantly depends on the null model of the network. According to the assumption by the authors of this measure [11], such a model is the Erdős-Rényi random graph model. An alternative assumption is the model of connections considered in [12]. According to it, the modularity criterion is modified:

$$\mu^*(G_n) = \sum_{j=1}^{k} \left(e_{j,n}^{in} - p_{j,n} e_{j,n} \right), \tag{11}$$

where $p_{j,n} = \frac{n_j - 1}{n-1}$, n_j is the size of community j (i.e., the number of nodes of color j), n is the size of the network.

It is proposed to use both of these assumptions, i.e., criteria (10) and (11).

The software implementation of the designed network models and methods for assessing their homophily has been

performed in the Python programming language. Network visualization is performed using Gephi.

5. Results of research on the homophilicity property of scale-free networks

5. 1. Development of an algorithm for generating a homophilic network model

Models based on preferential attachment, in particular the BA model (5)–(9), are not homophilic, but they can be used to model individual communities. Let each node i during creation receive a color (i.e., community number c_i): one of K existing ones, or become the firstborn of a new community K+1. In the last case, the next m–1 nodes also receive color K+1. If the new node n+1 received one of the existing colors $c_{n+1} \le K$, then it chooses partners according to the PA rule, but only among nodes of its color. The specified modification of the PA rule (6) takes the following form:

$$\pi_{i,n} = \frac{deg_{i,n}}{L_{k,n}} \cdot \delta\left(c_i, c_{n+1}\right), \tag{12}$$

where $L_{k,n}$ is the sum of powers of nodes of color k in a network of n nodes, i.e. the volume of this community; in turn, $\sum_k L_{k,n} = \mathrm{vol}\left(G_n\right)$.

If the number of nodes of color c_{n+1} is less than m (which happens when the first m nodes of a new color are added), then the missing nodes are selected according to the general rule (6), i.e., ignoring the color.

Therefore, the generation of the proposed homophilic network model can be carried out according to the algorithm, the pseudocode of which is shown in Fig. 1.

```
GenerateBAwithHomophily(nMax, m, pnew, select_color):
G = complete_graph(m)
K = 1 # number of communities
colors[1:m] = 1 # initialize seed nodes color
S[1] = list(1,…,m) # initial community
n = m
while n < nMax-1 :
    with probability pnew(n,K) :
        K = K + 1 # adding new color
        # adding m new nodes n+1, n+2,…,n+m
        for j = 1 to m :
            colors[n+j] = K
            S[K] = [n+j]
            targets = pref_attach(G, m+1-j)
            targets += [n+j]*(j-1)
            # add node and edges to graph
            G.add_node_and_edges(n+j, targets)
            n++
    otherwise :
        # add new node with number n+1
        col = select_color(G,S,K) # choose existing color
        colors[n+1] = col
        S[col].append(n+1)
        targets = pref_attach (S[col], m)
        # add node and edges to graph
        G.add_node_and_edges(n+1, targets)
        n++
return G
```

Fig. 1. Pseudocode of the homophilic network generation algorithm

The input parameters of this model (and the algorithm in Fig. 1) are the network size, the number of edges of the new node, as well as the functions *pnew* and *select_color*. The first of them calculates the probability $p_{K,n}^{new}$ of assigning a new node n+1 to a new community K+1. Function *select_color* assigns one of the existing colors $k \in \{1, \ldots, K\}$ to this node, based on the probabilities $p_{k,n+1}$.

It is the probabilities $p_{K,n}^{new}$, $p_{k,n+1}$ that determine whether the generated network will be scale-free; if so, with what scaling, and what expected values of the homophily indices (10), (11) it will have.

5. 2. Definition of rules for coloring existing nodes

Let there be K colors (i.e., communities) in the network G_n. Two cases are possible:

(i) the newborn node becomes the firstborn of a new community (the probability of this is $p_{K,n}^{new}$);

(ii) the newborn node is assigned to one of the existing colors.

According to the algorithm (Fig. 1) of network generation in case (ii), which is the main one, the proposed modified (12) PA-rule is applied. Then the expected increase in the degree of node i is equal to:

$$\Delta_{i,n} = m \cdot \pi_{i,n} \cdot p_{c_i,n+1} = m \cdot \frac{deg_{i,n}}{L_{c_i,n}} \cdot p_{c_i,n+1}, \tag{13}$$

where $p_{ci,n+1}$ is the probability that node n+1 will receive color c_i.

In order for the basic condition (2) of scale-free property of the network to be met and for the rank scaling index of BA networks ($\rho = 1/2$) to be preserved, it is necessary to choose $p_{ci,n+1}$ so that expression (13) coincides with (7). From this requirement it follows that the rule for coloring a new node with one of the existing colors (i.e., the essence of the *select_color* function from Fig. 1) takes the following form:

$$p_{k,n+1} = \frac{L_{k,n}}{\mathrm{vol}\left(G_n\right)}. \tag{14}$$

Thus, while the number of communities (colors) in the network remains unchanged, the application of the attachment rule (12) together with the coloring rule (14) leads to the same expression (6) for the expected increase in the degree of a node, and also the same equation for the dynamics of the growth of the degrees of nodes (8) as the usual BA model.

If node n+1 becomes the firstborn of a new color (case (i)), then not one, but m nodes of this color are created sequentially. The degrees of these nodes after m steps will be equal to m. The first node has m free (i.e., not connected to nodes of its color) ends, the second – m–1, the last – one. All these $m(m+1)/2$ free ends are connected to the nodes of the network according to the usual PA rule (6). Then the expected increase in the degree of node i after the adding of m nodes of a new color will decrease compared to the usual case (ii) by $(m+1)/(2m)$ times, but will remain discrete-power.

Considering that the creation of a new color is a rare event (i.e., $K \ll n$), the dependence of node degrees on the network size asymptotically approaches the discrete-power dependence of the BA-model (9). Thus, we can assume that under the conditions of coloring the nodes according to rule (14), the generated homophilic network (12) is asymptotically scale-free with respect to the node degrees.

5. 3. Definition of rules for the emergence of new community

Since traditional models based on PA-like rules generate networks that are not homophilic, the scale-free property of such networks is considered only with respect to the distribution

of node degrees. However, in the case of modeling a structured network, it is quite natural to consider the distribution of community volumes $L_{k,n}$ and, accordingly, the requirement for its scale-free property.

Let t_k be the time point preceding the appearance of color k. Since the time count is synchronized with the number of nodes, nodes with numbers t_k+1, ..., t_k+m will become the core of community k.

In the time interval between t_k+m+1 to t_{k+1} (i.e., after the completion of the formation of the core of color k and before the appearance of color k+1), the increments of degrees of all nodes of the network are determined according to (7). Summing them along the communities, we can derive the equation of the dynamics of increase in community volumes:

$$L_{j,n+1} = L_{j,n} + 2m \cdot p_{j,n+1} = L_{j,n} + \frac{L_{j,n}}{n-n'}, \quad (15)$$

$$j = 1,\ldots,k,\ n = t_k + m + 1,\ldots,t_{k+1}.$$

Equation (15) satisfies the general model (3) (for a=1, b=–n'), therefore its solution, according to (4), takes the following form:

$$L_{j,n} = L_{j,t_k+m} \cdot \frac{n-n'}{t_k+m-n'}. \quad (16)$$

As was proved in chapter 5.2, during the formation of the next, k+1, community $m(m+1)/2$ links are joined to the existing communities according to the modified PA rule (12). Thus, the volumes of these communities at step t_{k+1}+m will be:

$$L_{j,t_{k+1}+m} = L_{j,t_{k+1}} + \frac{m(m+1)}{2} \cdot p_{j,t_{k+1}+1} =$$
$$= L_{j,t_{k+1}} \left(1 + \frac{m+1}{4(t_{k+1}-n')}\right). \quad (17)$$

Taking into account (16), the volumes of communities j=1...k at the moment of completion of the formation of color k+1 is equal to:

$$L_{j,t_{k+1}+m} = L_{j,t_k+m} \cdot u(k),\ j = 1,\ldots,k, \quad (18)$$

where:

$$u(k) = \left(1 + \frac{m+1}{4(t_{k+1}-n')}\right) \cdot \frac{t_{k+1}-n'}{t_k+m-n'} = \frac{t_{k+1}-n'/2}{t_k+m-n'}. \quad (19)$$

The volumes of the cores immediately after their formation are:

$$L_{j,t_j+m} = 2m^2 - m(m+1)/2 = m(3m-1)/2. \quad (20)$$

Expanding equation (18), and taking into account (16) and (20), we can find the volume of the community $j \le k$ at an arbitrary time point n=t_k+m+1, ..., t_{k+1}:

$$L_{j,n} = \frac{m(3m-1)}{2} \cdot \frac{n-n'}{t_k+m-n'} \cdot \prod_{i=j}^{k-1} u(i). \quad (21)$$

Now we can formulate the condition for the scale-free property of communities with respect to their volume. According to (2), the ratio between the volumes of consecutive communities must satisfy the following condition:

$$\frac{L_{j+1,n}-L_{j,n}}{L_{j,n}} = \frac{-\alpha}{j+\alpha+\delta}, \quad (22)$$

for some $\alpha > 0$, $\delta > -1$.

Substituting (21) in (22), one can get:

$$\frac{L_{j+1,n}-L_{j,n}}{L_{j,n}} = \frac{1-u(j)}{u(j)} = \frac{-\alpha}{j+\alpha+\delta} \Rightarrow u(j) = \frac{j+\delta+\alpha}{j+\delta}. \quad (23)$$

The requirement for the relationship between the moments t_k and t_{k+1} follows from (23) and (29):

$$t_{k+1} = t_k + \frac{\alpha(t_k+c_1)}{k+\delta} + c_2, \quad (24)$$

where:

$$c_1 = \frac{m-1}{2},\ c_2 = \frac{3m-1}{4}.$$

In this case, the t_1 value should be conditionally set such that the volume of the first community at step t_1+m is the same as for all other communities t_k+m, i.e. (20). It follows that t_1=(m+1)/4. With this initial condition, the solution to the difference equation (24) takes the following form:

$$t_k = \begin{cases} -c_1 + \frac{c_2}{\alpha-1}\left(\frac{\Gamma(\delta+1)\cdot\Gamma(k+\delta+\alpha)}{\Gamma(\delta+\alpha)\cdot\Gamma(k+\delta)} - (k+\delta)\right), & \alpha \ne 1, \\ -c_1 + c_2(k+\delta)\left(\psi^{(0)}(k+\delta+1) - \psi^{(0)}(\delta+1)\right), & \alpha = 1, \end{cases} \quad (25)$$

where $\psi^{(0)}(x)$ is the digamma function (logarithmic derivative of the gamma function).

Dependences (25) at $k \to \infty$ take on the asymptotic form:

$$t_k \approx \begin{cases} \text{const} \cdot (k+\delta)^{\alpha}, & \alpha > 1, \\ c_2(k+\delta)\cdot\ln((k+\delta+1)), & \alpha = 1, \\ c_2/(1-\alpha)\cdot(k+\delta), & 0 < \alpha < 1. \end{cases} \quad (26)$$

Expression (26) clearly illustrates that for $\alpha > 1$ the sequence of moments of appearance of new colors is asymptotically power-law with respect to the color number. In this case, α is a scaling index, and the parameter δ determines not only the shift in power laws but also the intensity of the appearance of new colors compared to the growth rate of the number of network nodes.

The measure of the intensity of the appearance of colors is the time difference between the moments of appearance of the next and current colors. It is equal to:

$$\Delta t_k = t_{k+1} - t_k =$$
$$= \begin{cases} \frac{c_2}{\alpha-1}\left(\frac{\alpha\cdot\Gamma(\delta+1)\cdot\Gamma(k+\delta+\alpha)}{\Gamma(\delta+\alpha)\cdot\Gamma(k+\delta+1)} - 1\right), & \alpha \ne 1, \\ c_2\left(\psi^{(0)}(k+\delta+1) - \psi^{(0)}(\delta+1) + 1\right), & \alpha = 1, \end{cases} \quad (27)$$

or asymptotically (at $k \to \infty$):

$$\Delta t_k \approx \begin{cases} \text{const} \cdot (k+\delta)^{\alpha-1}, & \alpha > 1, \\ c_2 \cdot \ln(k+\delta), & \alpha = 1, \\ c_2/(1-\alpha), & 0 < \alpha < 1. \end{cases} \quad (28)$$

The plots of Δt_k (27) for k=1,...,20, m=5 and different values of α and δ are shown in Fig. 2, 3.

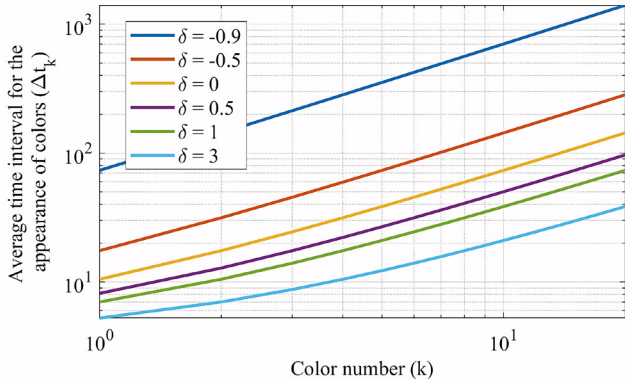

Fig. 2. Dependence of time intervals between the moments of community emergence on their serial number for $m=5$, $\alpha=2$

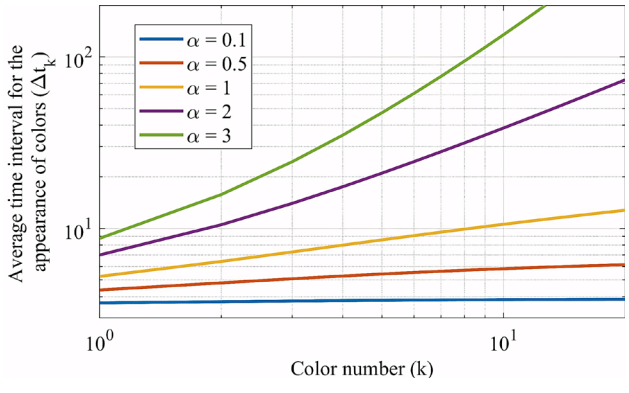

Fig. 3. Dependence of time intervals between the moments of community emergence on their serial number for $m=5$, $\delta=1$

The process of the emergence of a new color corresponds to the Bernoulli scheme – a sequence of attempts until the first successful one. Then, from (27), (28) it follows that for $\alpha>1$, the probability of the emergence of a new color at step n is asymptotically power-law:

$$p_{k,n}^{new}=\frac{1}{\Delta t_k}\approx \text{const}\cdot\left(k+\delta\right)^{1-\alpha}, \tag{29}$$

and in the case of $\alpha<1$, the colors appear with an asymptotically constant probability, which is equal to $p_{k,n}^{new}\approx\left(1-\alpha\right)/c_2$.

5. 4. Evaluation of the modularity of the modeled network

According to (23), (24), expression (21), which describes the number of connections in the community $j\le k$ at an arbitrary time point $t_k+m<n\le t_{k+1}$, can be represented in the form:

$$L_{j,n}=c_2\frac{2m\left(n-n'\right)}{t_k+c_1}\cdot\frac{\Gamma\left(k+\delta+\alpha\right)\cdot\Gamma\left(j+\delta\right)}{\Gamma\left(k+\delta\right)\cdot\Gamma\left(j+\delta+\alpha\right)}. \tag{30}$$

It follows that the relative volumes of communities are:

$$e_{j,n}=\frac{L_{j,n}}{\text{vol}\left(G_n\right)}=\frac{c_2}{t_k+c_1}\cdot\frac{\Gamma\left(k+\delta+\alpha\right)\cdot\Gamma\left(j+\delta\right)}{\Gamma\left(k+\delta\right)\cdot\Gamma\left(j+\delta+\alpha\right)}. \tag{31}$$

One can see that due to the very design of the algorithm for constructing a homophilic scale-free network (Fig. 1), the relative volumes of communities (31) are simultaneously the probabilities $p_{j,n+1}$ (14) of a new node obtaining color j.

If $\alpha>1$, then these volumes/probabilities asymptotically converge to:

$$e_j^{as}=\left(\alpha-1\right)\cdot\frac{\Gamma\left(\delta+\alpha\right)\cdot\Gamma\left(j+\delta\right)}{\Gamma\left(\delta+1\right)\cdot\Gamma\left(j+\delta+\alpha\right)}. \tag{32}$$

In particular, the relative volume of the first (and largest) community is $e_1^{as}=\frac{\alpha-1}{\alpha+\delta}$. From this it follows that by setting the weight of the first community e_1^{as} and the scaling index of community volumes $\alpha>1$, we can determine the parameter δ:

$$\delta=\frac{\alpha-1}{e_1^{as}}-\alpha. \tag{33}$$

Since each community, except the first one, is connected by $m(m+1)/2$ edges to the nodes of the previous communities (and this is the only reason for the emergence of inter-community connections), the relative number of such connections in the network is equal to:

$$E_n^{ext}=\frac{\left(k-1\right)m\left(m+1\right)}{\text{vol}\left(G_n\right)}=\frac{\left(k-1\right)\left(m+1\right)}{2\left(n-n'\right)}. \tag{34}$$

Thus, the modularity (10) of the model network is:

$$\mu\left(G_n\right)=1-\frac{\left(k-1\right)\left(m+1\right)}{2\left(n-n'\right)}-\sum_{j=1}^{k}e_{j,n}^2. \tag{35}$$

But in the case when α is not an integer, the sum of the squares of the relative volumes of communities does not have an analytical expression through elementary mathematical functions (although it is expressed in a rather cumbersome way through the hypergeometric function ${}_3F_2$).

The graphical representation of dependence (35) for some sets of model parameters is shown in Fig. 4, 5.

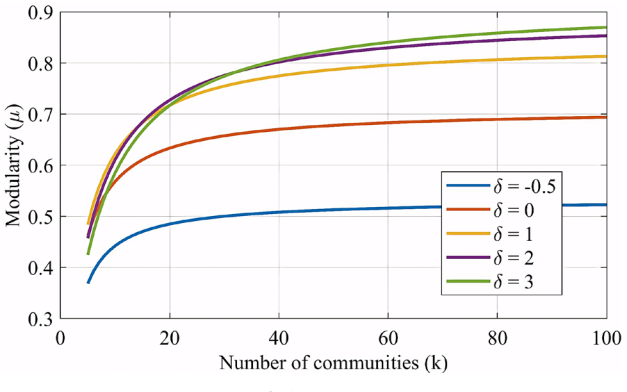

Fig. 4. Dependence (35) of the network modularity on the number of communities at $\alpha=3$ and varying δ

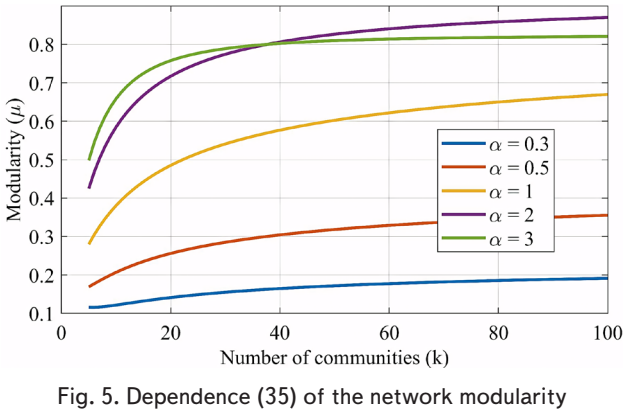

Fig. 5. Dependence (35) of the network modularity on the number of communities at $\delta=0$ and varying α

The calculation of the modified modularity (11) faces a similar problem since the community sizes n_j are asymptotically distributed according to a discrete power law similar to (31), (32). Therefore, the summation of $n_{j,n}e_{j,n}$ also leads to hypergeometric functions.

5. 5. Example of simulation of homophilic scale-free networks

The proposed model was implemented in software in order to demonstrate the main properties such as scale-free property and homophilicity. The algorithm (Fig. 1) was used as a basis. The assignment of an existing color to a new node (function *select_color*) is based on (14). The probability of a new color (i.e., function *pnew*) was calculated according to (27), (29).

A network of $n=10000$ nodes was generated. The size of the core was $m=5$. It was obtained (Fig. 6) that the rank distribution of node degrees coincides with the corresponding distribution for the BA model, i.e., it does not depend on the parameters of the community structure. It is discrete-power with a scaling index $\rho=1/2$.

To investigate the scale-free property of communities, *nSamples*=10 networks were generated (with parameters $\alpha=4n=10000$, $m=5$, $\alpha=4$, $\delta=0$). The distributions of community volumes and sizes (i.e., the number of links and nodes) are shown in Fig. 7.

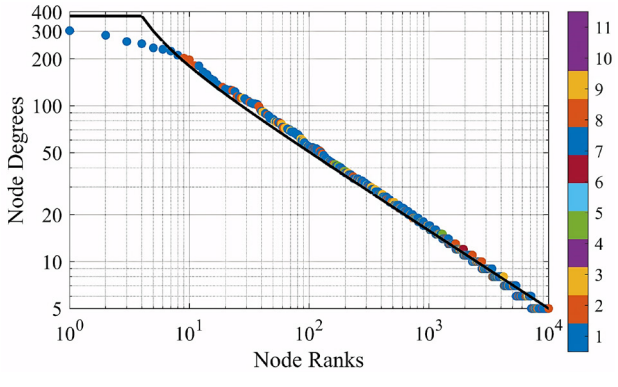

Fig. 6. Rank distribution of node degrees of homophilic network

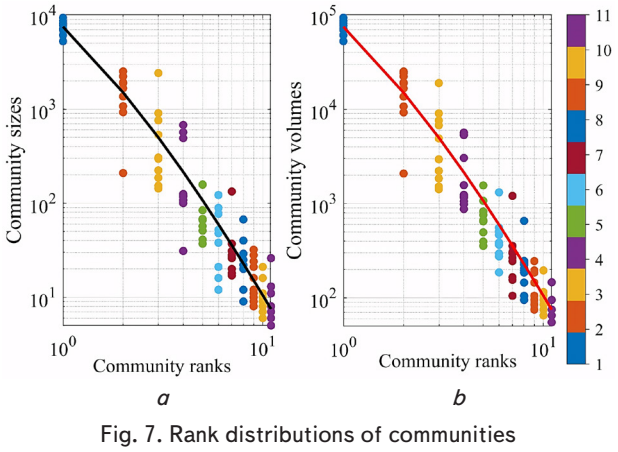

Fig. 7. Rank distributions of communities of a scale-free homophilic network: *a* – by community size; *b* – by community volume

Fig. 8 shows the rank distribution of the averaged community size of the generated scale-free homophilic networks depending on the given scaling index α (for $n=10000$, $m=5$, $\alpha=4$, $\delta=0$).

It is easy to see that the distributions shown in Fig. 7, 8 are very close to the predicted (30) discrete power laws. The modularity of the generated model networks was also estimated. It ranges from 0.4 to 0.8 and fully corresponds to the numerical estimates given in Fig. 5.

For demonstration purposes, a network with $n=500$ nodes ($m=5$, $\delta=0$, $\alpha=3$) was generated. Its visualization was performed using Gephi and is shown in Fig. 9. The generated network has 7 colored communities (the largest of them, red, contains about 320 nodes, the smallest – green and black – 5 nodes each). The size of the nodes in Fig. 9 is proportional to their degree (which varies from 5 to 79).

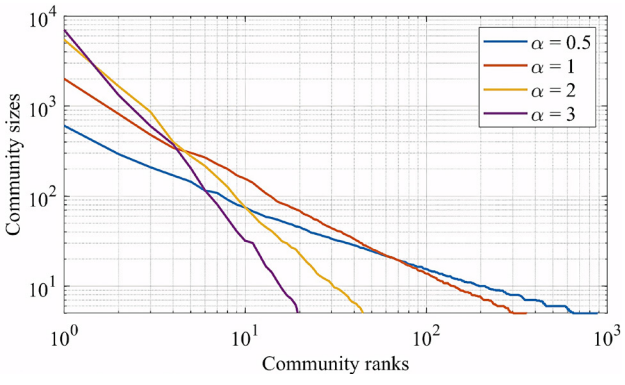

Fig. 8. Rank distribution of the average community size of scale-free homophilic networks when varying the scaling index

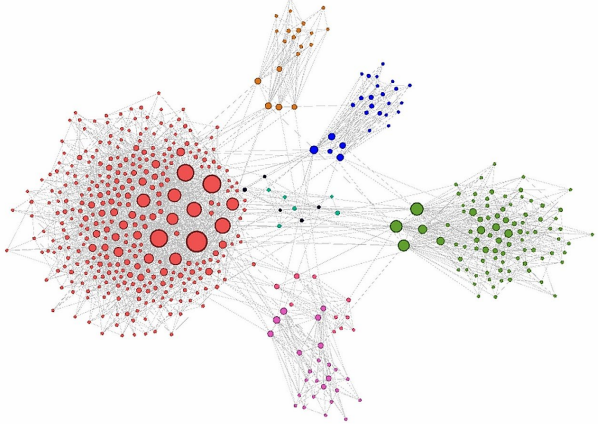

Fig. 9. Example of a homophilic scale-free network with 500 nodes

The drawings above (Fig. 6–9) demonstrate the main properties of the proposed network model, namely homophily and scale-free property with respect to node degrees and community sizes.

6. Discussion of results based on the construction of a generative model of homophilic scale-free networks

The proposed algorithm (Fig. 1) allows for a step-by-step (node-by-node) generation of a growing scale-free network, which at the same time has the property of homophilicity. Although this model is based on the principle of preferential attachment (6), it fundamentally differs from other PA models [7] precisely in that it allows generating homophilic networks. This is achieved due to the two-stage application

of the PA rule. First, this rule in the form of (14) is used to choose the color of a new node, and then, according to (12), it is used to determine the nodes (of the same color as the new one) with which this new node connects.

The proposed model is absolutely workable, i.e., it allows to generate a truly homophilic network (Fig. 9), while the distribution of node degrees is identical (Fig. 6) to the corresponding distribution for scale-free networks that do not have internal structure, i.e., communities.

In the course of our research, a requirement was put forward, which is quite natural, that the scale-free property should apply not only to the distribution of node degrees but also to the distribution of community sizes/volumes. It was proved that the specified requirement will be satisfied if the moments of emergence of new communities t_k are determined according to (25). This corresponds to the values (27) for the mathematical expectation of the time intervals between the emergence of successive communities. It is also shown that depending on the given scaling index (α) of the distribution of community sizes, the specified intervals asymptotically approach the power dependence (28) on the size of the communities, or to constant values. Our results differ significantly from [1], in which the given asymptotic expressions for the specified time intervals are different, and the scale-free property of the size of the communities is declared but not proven.

A certain limiting factor in the use of the model built is the cumbersomeness of dependences (27). The experimental results give reason to believe that in many cases, during the practical application of the model, it is advisable to replace these exact expressions with much simpler asymptotic ones (28). However, this requires additional mathematical substantiation, which is beyond the scope of the current study.

The relationship between the model parameters (scaling of community volumes, time shift), its scale (number of nodes and communities), and the characteristic of the model's homophily – modularity (31), (35) – was determined. The specified dependence was studied numerically (Fig. 4, 5); however, it does not have an expression in terms of elementary mathematical functions, which is a drawback of the study. Accordingly, the problem of determining the specified dependence arises at least for asymptotically large networks. In addition, it would be advisable to study other (in addition to modularity) indicators of network homophily: conductivity, entropy ratio. Although the experiments conducted in [10] indicate the synchronicity of these indicators (in the sense of equal significance) for asymptotically large networks, this fact has not been proven for networks of finite size.

Another area for further research is the generalization of the proposed approach to implementing the homophily property for scale-free network models from the Barabási-Albert model [7] to scale-free networks with arbitrary scaling and to elastic networks [13, 14].

Thus, the improvement of the model created, as well as the corresponding theoretical studies on the combination of scale-free property and homophily of networks, remain a relevant and practically significant task for further research.

7. Conclusions

1. An algorithm for generating a model of homophilic networks has been developed. This result is important since the proposed algorithm provides the possibility of step-by-step generation of a homophilic network of any given size and number of communities.

2. It has been proven that in order to simultaneously maintain homophilicity and preserve the scale-free property of the network (with respect to the node degrees), it is necessary to assign a color to a new node according to the PA rule with respect to the volumes of communities. Thus, the attachment of a new node to the existing ones occurs by a two-step application of the PA rule. It has been theoretically proven and experimentally confirmed that according to the specified rule for coloring nodes, the distribution of their degrees is identical to the distribution of node degrees of ordinary scale-free networks that do not contain communities.

3. The probabilities of the appearance of new colors have been determined, complying with which renders the generated network scale-free in terms of the volume and size of communities. The specified probabilities correspond to the expected time interval between the appearance of communities, which is asymptotically power (relative to the current number of communities), logarithmic, or constant. Both asymptotic and exact estimates of the dependence of length of the specified intervals on the structural parameters of the model that are set have been obtained.

4. The dependence of the main measure of network homophily – modularity – on the model parameters has been established. The specified dependence has no expression in terms of elementary functions but has been investigated experimentally.

5. The model has been implemented in software; the compliance of the properties of the generated networks with the specified requirements was checked. It has been determined that the generated networks are homophilic and scale-free with respect to the node degrees and the volumes of communities.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available, either in numerical or graphical form, in the main text of the manuscript.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Acknowledgments

The author team expresses gratitude to the participants of the scientific seminar at the AI Department of NURE "Models and methods of artificial intelligence in scientific research" for their support and useful advice.

References

1. Li, A., Li, J., Pan, Y. (2013). Homophyly Networks – A Structural Theory of Networks. arXiv. https://doi.org/10.48550/arXiv.1310.8295
2. Dorogovtsev, S. N., Mendes, J. F. F., Samukhin, A. N. (2000). Structure of Growing Networks with Preferential Linking. Physical Review Letters, 85 (21), 4633–4636. https://doi.org/10.1103/physrevlett.85.4633
3. Newman, M. E. J. (2003). Mixing patterns in networks. Physical Review E, 67 (2). https://doi.org/10.1103/physreve.67.026126
4. Mele, A. (2017). A Structural Model of Homophily and Clustering in Social Networks. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3031489
5. Choromański, K., Matuszak, M., Miękisz, J. (2013). Scale-Free Graph with Preferential Attachment and Evolving Internal Vertex Structure. Journal of Statistical Physics, 151 (6), 1175–1183. https://doi.org/10.1007/s10955-013-0749-1
6. Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics, 46 (5), 323–351. https://doi.org/10.1080/00107510500052444
7. Albert, R., Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics, 74 (1), 47–97. https://doi.org/10.1103/revmodphys.74.47
8. Noldus, R., Van Mieghem, P. (2015). Assortativity in complex networks. Journal of Complex Networks, 3 (4), 507–542. https://doi.org/10.1093/comnet/cnv005
9. Shergin, V., Udovenko, S., Chala, L. (2020). Assortativity Properties of Barabási-Albert Networks. Data-Centric Business and Applications, 55–66. https://doi.org/10.1007/978-3-030-43070-2_4
10. Li, A., Li, J., Pan, Y. (2013). Community Structures Are Definable in Networks, and Universal in Real World. arXiv. https://doi.org/10.48550/arXiv.1310.8294
11. Girvan, M., Newman, M. E. J. (2002). Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99 (12), 7821–7826. https://doi.org/10.1073/pnas.122653799
12. Shergin, V., Grinyov, S., Chala, L., Udovenko, S. (2024). Network community detection using modified modularity criterion. Eastern-European Journal of Enterprise Technologies, 6 (4 (132)), 6–13. https://doi.org/10.15587/1729-4061.2024.318452
13. Shergin, V., Chala, L., Udovenko, S., Pogurskaya, M. (2020). Elastic Scale-Free Networks Model Based on the Mediaton-Driven Attachment Rule. 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 291–295. https://doi.org/10.1109/dsmp47368.2020.9204207
14. Raskin, L., Sira, O. (2021). Devising methods for planning a multifactorial multilevel experiment with high dimensionality. Eastern-European Journal of Enterprise Technologies, 5 (4 (113)), 64–72. https://doi.org/10.15587/1729-4061.2021.242304