Comparative analysis of collaboration networks

Tatiana Progulova and Bahruz Gadjiev

International University for Nature, Society and Man,
19 Universitetskaya Street, Dubna, 141980 Russia

Abstract. In this paper we carry out a comparative analysis of the word network as the collaboration network based on the novel by M. Bulgakov "Master and Margarita", the synonym network of the Russian language as well as the Russian movie actor network. We have constructed one-mode projections of these networks, defined degree distributions for them and have calculated main characteristics. In the paper a generation algorithm of collaboration networks has been offered which allows one to generate networks statistically equivalent to the studied ones. It lets us reveal a structural correlation between word network, synonym network and movie actor network. We show that the degree distributions of all analyzable networks are described by the distribution of *q*-type.

Keywords: Maximum entropy principle, Bipartite graphs, Collaboration networks, Word networks.
PACS: 05.045.-a, 87.23.Ge, 89.75.He

INTRODUCTION

The general system theory starts with a definition of "a system" as a complex of interacting components that form an organized entity. Such a definition naturally leads to efficient system simulation in the language of the graph theory [1, 2].

At present the networks of the real world, such as communication networks (WWW and Internet), networks of collaboration, biological networks (neural networks, networks of metabolic reaction, genome and protein networks), spatial networks (networks of airlines and road networks, power grid) and others are well enough studied. However, in spite of the fact that all these networks are growing, and grow basically according to a principle of preferential attachment, for them the combined mechanisms of growth are characteristic. The combined mechanisms of growth (nodes ageing and removal, local action of preferential attachment, accelerated growth, etc.) lead to peculiarities of real world networks topology [1, 2].

In this work we study peculiarities of collaboration networks. Collaborations necessarily imply the presence of two constituents — collaborators and acts of collaboration. So, the set of collaborations can be represented by the bipartite graph. A bipartite graph is a graph whose set of vertices can be divided into non-intersecting subsets V_1 and V_2 in such a way that each edge connects vertices from different subsets.

In this paper we represent the unified algorithm for generation of various collaboration networks and, using a maximum entropy principle, we derive the

distribution function which allows describing topology properties of collaboration networks in the framework of a simple formalism.

PECULIARITIES OF COLLABORATION NETWORKS TOPOLOGY FORMATION

The topology of growing networks is formed through local events, such as new vertices and links addition or interconnection of an edge of one vertex to another. Networks with different topologies – from exponential to power-law ones – can be formed depending on the contribution of these events. In principle, it is described by the extended Barabasi model [3]. In this model the network is developing in the following way. The growth process starts with m_0 isolated vertices, and at each time step one of the following operations is realized:

(i) With probability p between the existing nodes m of new connections is added. One of the vertices for each edge is chosen randomly, the other – in accordance with the preferential attachment principle. Thus,

$$\left(\frac{\partial k_i}{\partial t}\right)_{(i)} = pA\frac{1}{N} + pA\frac{k_i + 1}{\sum_j \left(k_j + 1\right)} \qquad (1)$$

where N is number of nodes. As the change of the total degree after a regular step is $\Delta k = 2m$, $A = m$.

(ii) With probability q interconnection of m edges takes place. The equation to change the degree is of the form:

$$\left(\frac{\partial k_i}{\partial t}\right)_{(ii)} = -qB\frac{1}{N} + qB\frac{k_i + 1}{\sum_j \left(k_j + 1\right)} \qquad (2)$$

Herein, the first member describes the decrease of the degree of the vertex from which the edge was disconnected; the second member describes the increase of the degree of the vertex to which the edge was connected. This second vertex is chosen in accordance with the preferential attachment principle. The total degree does not change and it is possible to show that $B = m$.

(iii) With probability $1 - p - q$ a new vertex comes into the network that connects by m edges with various nodes already existing in the network according to the preferential attachment principle. Thus

$$\left(\frac{\partial k_i}{\partial t}\right)_{(iii)} = \left(1 - p - q\right)C\frac{k_i + 1}{\sum_j \left(k_j + 1\right)}. \qquad (3)$$

The number of connections that join a new node to those that are present in the system is m, and, thus, $C = m$.

Summarizing the contribution of processes (1), (2) and (3), we get

$$\frac{\partial k_i}{\partial t}=\left(\frac{\partial k_i}{\partial t}\right)_{(i)}+\left(\frac{\partial k_i}{\partial t}\right)_{(ii)}+\left(\frac{\partial k_i}{\partial t}\right)_{(iii)}=(p-q)m\frac{1}{N}+(1-p-q)\frac{k_i+1}{\sum_j(k_j+1)} \quad (4)$$

As the system dimension N and the total degree $\sum_j k_j$ change with time as $N(t)=m_0+(1-p-q)t$ and $\sum_j k_j=(1-q)2mt$, for large t we can neglect m_0 in comparison with members that grow linearly with time. Then $N(t)=(1-p-q)t$ and $\sum_j(k_j+1)=(1-q)2mt+N=(1-q)2mt+(1-q-p)t$, and equation (4) takes the form:

$$\frac{\partial k_i}{\partial t}=(p-q)m\frac{1}{(1-p-q)t}+(1-p-q)m\frac{k_i+1}{(1-q)2mt+(1-q-p)t}. \quad (5)$$

Let us introduce notations

$A(p,q,m)=(p-q)\left(\frac{2m(1-q)}{1-p-q}+1\right)$ and $B(p,q,m)=\frac{2m(1-q)+(1-p-q)}{m}$.

Then equation (5) can be written in the form

$$\frac{\partial k_i}{\partial t}=[A(p,q,m)+1+k_i]\frac{1}{B(p,q,m)t} \quad (6)$$

The solution of equation (6) has the form

$$k_i(t)=[A(p,q,m)+1+m]\left(\frac{t}{t_i}\right)^{\frac{1}{B(p,q,m)}}-A(p,q,m)-1. \quad (7)$$

The probability that the node has the degree $k_i(t)$ smaller than k, $P[k_i(t)<k]$, can be written as $P[k_i(t)<k]=P[t_i>C(p,q,m)t]$, where the notation is introduced

$$C(p,q,m)=\frac{(A(p,q,m)+m+1)^{B(p,q,m)}}{(k+m+1)^{B(p,q,m)}}. \quad (8)$$

As t_i must meet the requirement $0\le t_i\le t$, we can distinguish the following cases:

(i) If $C(p,q,m)>1$, $P[k_i(t)<k]=0$.

(ii) If $0<C(p,q,m)<1$, the degree distribution $P(k)$ can be defined analytically.

If connections addition, interconnection and growth occur homogeneously in time $P_i(t_i)=\frac{1}{m_0+t}$, and therefore

$$P[k_i(t)<k]=P[t_i>C(p,q,m)t]=1-P[k_i(t)>k]=1-\frac{C(p,q,m)t}{m_0+t}, \quad (9)$$

Wherein, using the definition $P(k,t)=\frac{\partial P[k_i(t)<k]}{\partial k}$, we obtain for the degree distribution [3]

$$P(k,t)=D(p,q,m)\frac{t}{m_0+t}(k+A(p,q,m)+1)^{-1-B(p,q,m)}, \quad (10)$$

where $D(p,q,m) = B(p,q,m)\left[m + A(p,q,m+1)\right]^{B(p,q,m)}$. Therefore, the degree distribution has a generalized power-law form. The degree distribution is the main topological characteristic of the network that determines the probability that a random chosen vertex in the network at the time moment t has the degree k.

In the limit $t \to \infty$ we obtain a stationary degree distribution:

$$P_{st}(k) = D(p,q,m)\left(k + A(p,q,m) + 1\right)^{-1-B(p,q,m)}. \quad (11)$$

This result is obtained in [3]. However, it can be clearly seen that distribution (11) is the Zipf-Mandelbrot distribution.

We will show further that the Zipf-Mandelbrot distribution can be written in the form of the Tsallis distribution. Introducing the notations $\kappa(p,q,m) = A(p,q,m) + 1$ and $\gamma(p,q,m) = B(p,q,m) + 1$, we get

$$P_{st}(k) = D(p,q,m)\left(\kappa(p,q,m) + k\right)^{-\gamma(p,q,m)}. \quad (12)$$

The latter expression can be written in the form

$$P_{st}(k) = D(p,q,m)\left(\kappa(p,q,m)\right)^{-\gamma(p,q,m)}\left(1 + \frac{1}{\kappa(p,q,m)}k\right)^{-\gamma(p,q,m)}. \quad (13)$$

Let $\frac{1}{\kappa(p,q,m)} = -(1-q)\frac{1}{k_0}$ and $\frac{1}{1-q} = -\gamma(p,q,m)$. Then $k_0 = \frac{\kappa(p,q,m)}{\gamma(p,q,m)}$ and

$$P_{st}(k) = D(p,q,m)\left(\kappa(p,q,m)\right)^{-\gamma(p,q,m)}\left(1 - (1-q)\frac{k}{k_0}\right)^{\frac{1}{1-q}}. \quad (14)$$

Using the probability normalization condition $P_{st}(k)$, it can be written

$$P_{st}(k) = \frac{1}{Z}\left(1 - (1-q)\frac{k}{k_0}\right)^{\frac{1}{1-q}}, \quad (15)$$

where

$$Z = \int_0^{\infty}\left(1 - (1-q)\frac{k}{k_0}\right)^{\frac{1}{1-q}}dk. \quad (16)$$

Thus, we get the Tsallis distribution [4] that has the form of an exponential distribution at $q \to 1$ while at large k it has the form of power-law degree distribution. In log-log scale this distribution has a plateau at $q \neq 1$ and small k where it reaches its maximum value.

Below we will show that the given distribution in such a form does not always describe the topology of collaboration networks.

COLLABORATION NETWORKS

Collaboration networks are a bipartite graph. We have constructed one-mode projections of these networks, defined degree distributions for them and have calculated main characteristics. Bright examples of collaboration networks are

collaboration networks of movie actors, synonyms and words. Hereafter, we will discuss some examples of these networks in detail.

Collaboration Network of Movie Actors

To construct a collaboration network of movie actors we used data on the official sites of the Lenfilm cinema (www.lenfilm.ru) and Mosfilm cinema (www.mosfilm.ru) studios. In the period from 1918 to 1992 these cinema studios – the largest producers of movies in Russia – issues 2910 fiction movies. It should be noted that the data in the studios sites contain lists of only well-known actors of the first and second rank.

We constructed the collaboration network of movie actors in the following way: vertices are actors; two vertices are connected with an edge if the corresponding actors acted at least once during their careers in one and the same film. The constructed network contains 6997 vertices (actors) and 759490 edges. Degree distribution was constructed and given in Fig. 1. The distribution maximum corresponds to $k = 6$. With $k > 6$ the distribution is well described by the degree law: $p(k) \sim k^{-\gamma}$. We defined the value using the maximum likelihood method: $\gamma = 2.14$.

It should be noted that the Tsallis distribution in the form (15) does not describe the actors network distribution.

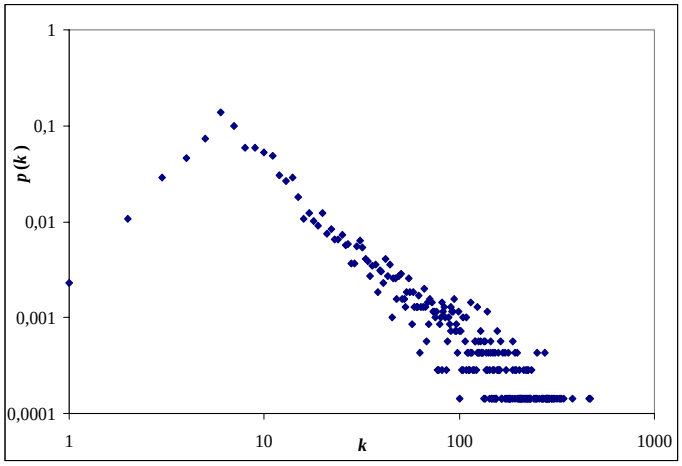

FIGURE 1. Degree distribution for collaboration network of actors

Synonyms Network

The synonyms network was constructed on the “Dictionary of Russian Synonyms and Expressions with Similar Meaning” by N. Abramov. In this network two vertices (words) are connected with an edge in case they are synonyms of one and the same word. In the constructed network there are 2001 vertices and 4871 edges. Degree distribution of the synonyms network is given in Fig. 2. This distribution is described by the Tsallis distribution (15).

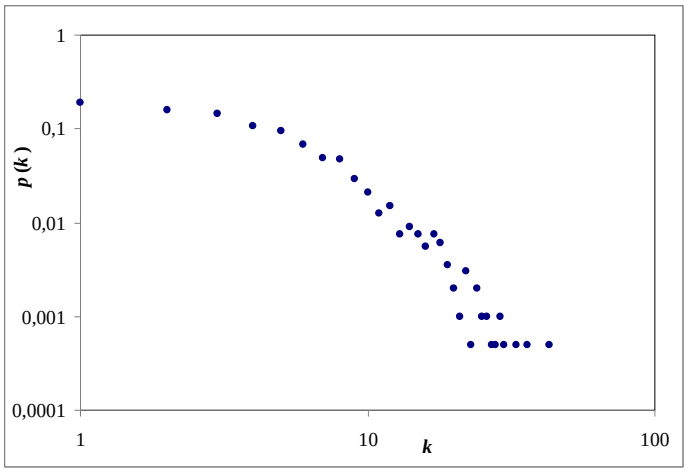

FIGURE 2. Degree distribution for the network of synonyms of the Russian language

Collaboration Network of Words

The collaboration network of words was constructed on the basis of the M. Bulgakov's novel "The Master and Margarita". Here, two words were connected with an edge in case they are both contained in one sentence. This novel can be arbitrarily divided into two parts: "Yeshua's Story" and "The Master's Story". We constructed word networks for each part and for the whole novel. Besides, the word network was constructed on the translation of the novel into English. Degree distribution for the whole novel in the original is shown in Fig. 3. The maximum of this distribution corresponds to $k \approx 30$. The form of distribution and the maximum position are the same for the indicated parts of the novel. The degree distribution of the word network constructed for the translation into English has a similar distribution form but the distribution maximum is at $k \approx 40$. At large k distributions can be described by the power law. We defined the values of the degree exponents. For the whole novel in Russian "Yeshua's Story" and "The Master's Story" $\gamma = 2{,}17;\ 2{,}15;\ 2{,}30$, correspondingly, and for the analogous parts of the English translation $\gamma = 2{,}07;\ 2{,}05;\ 2{,}17$, correspondingly. It means that in translation of the novel's text from Russian into English ordering of the distribution parameters remains.

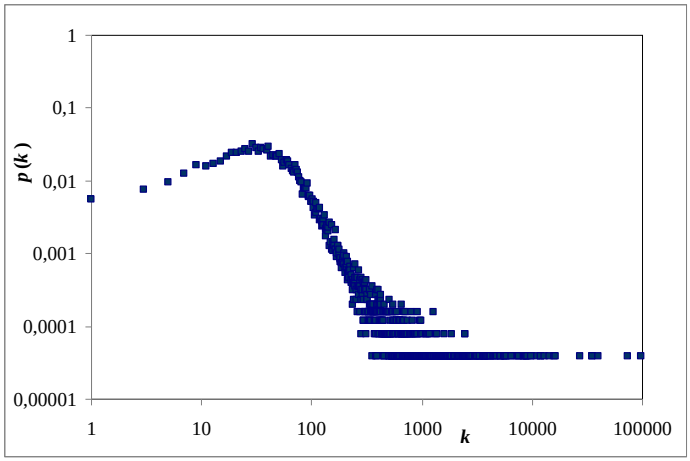

FIGURE 3. Degree distribution for the word network constructed on the novel by M. Bulgakov "The Master and Margarita"

It follows from the above said that the Tsallis distribution in the form (15) does not allow one to describe the distribution family characteristic for collaboration networks. It means that actually forming of collaboration networks includes additional growth mechanisms.

SIMPLE ALGORITHM FOR GENERATION OF GROWING COLLABORATION NETWORKS

Below, we present an algorithm that allows one to describe the family of distributions characteristic for collaboration networks.

We start with m_0 isolated vertices; at each time step with probability p we add m $(m \le m_0)$ new edges or with probability $1-p$ we add a new vertex with m edges. The connection with the existing vertices in the network occurs in accordance with the preferential attachment principle [1]. Value m is randomly taken and is distributed homogenously on a certain interval. In this case the numerical analysis shows that instead of a plateau a characteristic maximum occurs. Thus, with constant m we get degree distribution in the form of the Tsallis distribution while other distributions can be described assuming that m is randomly taken.

MAXIMUM ENTROPY PRINCIPLE AND RESULTS OF FITTING

For the classification of the topology of collaboration networks we use the complex networks theory, namely the nonextensive information entropy theory. The generalized entropy is given by the expression

$$S_q = \frac{1 - \int_0^\infty p^q(k)dk}{q-1}, \quad (17)$$

where q — entropy index that is a measure of the system complexity. We consider the following restrictions: $\int_0^\infty p(k)dk = 1$, that is a normalization condition and $\int_0^\infty |k - k_0|^b p^q(k)dk = const$. From the maximum entropy principle with the given restrictions after using the Lagrange method, we obtain:

$$p(k) = \frac{c}{\left(1 + a|k - k_0|^b\right)^{1/(q-1)}}. \quad (18)$$

The distribution has the form of the Mandelbrot distribution and can be written in the form:

$$p(k) = Z^{-1}\left(\exp_q\left(-\varsigma(k-\eta)^\alpha\right)\right)^q, \quad (19)$$

where

$$Z = \int_0^\infty \left(\exp_q\left(-\varsigma(\kappa-\eta)^\alpha\right)\right)^q dk. \quad (20)$$

Herein, $\exp_q(x) = \left[1 + (1-q)x\right]^{\frac{1}{1-q}}$ — q-exponential function.

The obtained distribution (19) allows one to describe the family of distributions characteristic for collaboration networks, in particular, at $\alpha = 1$ we obtain the Tsallis distribution while $\alpha = 2$ corresponds to q-exponential Gaussian distribution.

The algorithm presented here can be regarded as a “microscopic” mechanism that generates distribution of this type. Using distribution (19) we did the fitting and defined the parameters of the networks under the analysis: q = 3.4 and $\alpha = 1.2$ for the movie actor collaboration network, q = 1.05 and $\alpha = 1.05$ for the synonym network and q = 4.0 and $\alpha = 1.7$ for the world network.

REFERENCES

1. R. Albert and A.-L. Barabasi, *Rev. Mod. Phys.* **47,** 43–97 (2002).
2. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, *Phys. Rep.* **424**, 175–308 (2006).
3. R. Albert and A.-L. Barabasi, *Phys. Rev. Lett.* **85** (24), 5234-5237 (2000).
4. C. Tsallis *Braz. J. Phys.* **29**, 1-35 (1999).