Computing and Informatics, Vol. , 2010, ??–??, V 2010-Aug-26

POSITIONAL WORD WEB AND ITS NUMERICAL AND ANALYTICAL STUDIES

Peter Náther, Mária Markošová

Dept. of Applied Informatics,
Faculty of Mathematics, Physics and Informatics
Comenius University, Bratislava, Slovakia

e-mail: nather@ii.fmph.uniba.sk, markosova@fmph.uniba.sk

Abstract. *We present numerical studies of the positional word web [1], based on an ancient text (The Bible). Adopting methods from the graph theory we show that such a web has small world and scale free properties [6, 11]. Degree distribution and hierarchy of nodes in various word webs are also studied. Complete solution of the mathematical model, which explains the differences between the data and the previous model of Dorogovtsev and Mendes [4] is presented.*

PACS 87.10.+e, 89.20.-a,89.75.Da, 89.75.Hc

Keywords: small world, scale-free, word web language

1 INTRODUCTION

Everywhere, where interactions between certain amount of similar entities is present, a network is probably a useful tool to model such a system [1, 2, 4, 5, 7, 8]. A network consists of nodes connected by edges and its mathematical representation is a graph. Networks are called **binary**, if an edge between two nodes either exists or not, and if it doesn't posses any additional numerical quantity. To characterize a network, some local and global measures are used.

Local network structure is reflected in the clustering coefficient C

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i. \tag{1}$$

Here N is the number of nodes and C_i measures the ratio of connected neighbours of the node i:

$$C_i = \frac{2E_i}{k_i(k_i-1)}, \tag{2}$$

where E_i is a number of existing edges between neighbours of the node i and k_i is its degree.

Separation l, calculated as an average shortest distance between randomly chosen pairs of nodes, is a global quantity characterizing how far, in the number of traveled edges, nodes are from each other.

If the network combines high clustering with low node separation, it is called a **small world network**. Many networks in Nature are of this type [6].

Apart from the above mentioned properties, the final structure of real networks strongly depends on their development in time. If the network grows, it develops a typical structure, depending on the type of the node addition and other processes on network [9, 10, 11, 12]. The node addition might be accompanied by a deletion, but the ratio of deleted nodes is often negligible.

Barabási and Albert have shown analytically [9] that if the network grows by preferential node attachment, the final structure is **scale free** (3), possessing a huge number of nodes with few neighbours, and few nodes having many neighbours. The concept "preferential" means that the attachment probability of a newcomer is proportional to the degree of the older node choosen for linking.

Network dynamics is reflected in the degree distribution [10, 11]. In the Barabási and Albert network (BA model) [9] this function has a power law character

$$P(k) \propto k^{-\gamma_{BA}}. \tag{3}$$

with scaling exponent $\gamma_{BA} = 3$. In such scale free networks [9] there is no typical node having typical degree. Another variants of the preferential linking lead to the scaling exponents different from $\gamma_{BA} = 3$ [10, 13].

Theory of growing networks has many interesting applications [1, 14, 19]. It is very popular in computer science as well. Because for example Internet and

WWW network are also complex networks. To understand how the structure of network infulences its performance and allows to develop efficient algorithm is very important [21, 22]. In this paper we present one of them. We study **positional word web** [1, 4], consisting of words as nodes and word interactions as an edges. In this web the interaction is defined by the neighbourhood in a sentence. Similar network, developed from texts in the English national corpus, has been investigated by Cancho and Solé (CS word web) [1]. The authors have shown, that the degree distribution indicates preferential addition of nodes. But there are two different scaling regimes, one for well connected kernel lexicon words and the other for less connected ones. They guess, that the kernel vocabulary has different dynamics, then the non kernel one.

The results of Cancho and Solé were revised by Dorogovtsev and Mendes and by us [4, 13]. They proposed a mathematical model explaining two scaling regimes. In this model they add a process of a line creation between old nodes in the network. We extended their model in such a way, that exponents of scaling regimes of our model are the same as were measured. To show this we created a word web of the ancient text, The Bible, which did not change its vocabulary for several hundreds of years. Therefore we believe that it consists of kernel lexicon words only. We have measured the degree distribution of The Bible positional word web (BWW), together with its small world properties. To show, that the appearance of the two scaling regimes depends rather on the network size then on other properties, we generated several artificial networks, simulating processes included in our model [13].

2 WORD WEB

Lexicon of the human language is composed of hundred thousands words. Let us have, for example an English language. English national corpus consists of about 500000 words [1]. Not all of them are used by all members of the English population. There are words, which are frequent and understandable for all, independently of the age, education, etc. This set is called **kernel lexicon** and includes about ten thousand words.

What is the structure of the lexicon in our brain? This is a good question, if we have in mind, how great the word database is, and how quickly our brain retrieves in it. Several studies have been made to find the answer [1, 2, 6]. All of them show, that its structure is scale free and small world like.

Small worlds are networks optimizing local structure preservation and good node communication [6]. They are partly ordered and partly disordered - as an example can serve lattice graph in which some amount of randomly chosen edges is rewired, creating random shortcuts. Clustering coefficient (1,2) of such net is usually rather high, but the node separation is low due to the shortcuts.

Degree distribution of the positional word web indicates scale free structure (3), but with two different scaling regimes [4]. For well-connected kernel words with great degree, the scaling exponent is close to γ_{BA}. Less connected words scale with

$\gamma^1_{DM} = 1.5$ (we marked the γ exponent with 1 for the first of two scaling regimes of the distribution and DM for the Dorogovtsev-Mendés model). These results of Cancho and Solé [1] were explained by Dorogovtsev and Mendes [4].

Dorogovtsev and Mendes do not agree, that the different scaling means different dynamics in the kernel and the non kernel vocabulary. They rather reason, that the two scaling regimes occur due to some additional processes running together with the preferential node addition (DM model).

DM model of the positional word web includes these processes:

- One starts with a small initial net. Each time unit a node comes and links itself preferentially by m new edges to the old nodes. Each newcomer is labeled according its birth time s. This is the same as in the Barabasi Albert model (BA model) [9].
- Simultaneously ct new edges (that means $2ct$ edge ends, $c << 1$) are created and connect the old nodes with preference.

What all of this means in the word web terminology? The new nodes are in fact new words which time to time appear in the vocabulary. They are included in the context of the old ones. But, simultaneously, several old words (nodes) enrich their own context. They are used with some words, with which they have not been used before. Let us mention one example. To tell the sentence "Portable computer has a new design" in sixties had no sence. Word "computer" was not used together with "portable", because computers were big devices. Now it is quite a common phrase.

Mathematically the DM model can be written as:

$$\frac{\partial k(s,t)}{\partial t} = (m + 2ct)\frac{k(s,t)}{\int_0^t k(s,t)ds} \tag{4}$$

where $k(s,t)$ is an average degree of node with a birth time s seen at time t and the integral gives the sum of all node degrees

$$\int_0^t k(s,t)ds = 2mt + ct^2, \tag{5}$$

and thus $\frac{k(s,t)}{2mt+ct^2}$ is the node attachement kernel expression. With a help of (5) the solution of (4) is [4]:

$$k(s,t) = m\left(\frac{ct}{cs}\right)^{\frac{1}{2}}\left(\frac{2m+ct}{2m+cs}\right)^{\frac{3}{2}}. \tag{6}$$

Here $k(s,t)$ scales with s as $k(s,t) \propto s^{-\beta}$, where β is another scaling exponent, which express the scaling of $k(s,t)$ with s. It has been shown, that the scaling exponent γ of the distribution (3) is related to β as [10]

$$\gamma = 1 + \frac{1}{\beta}. \tag{7}$$

For $s << t$ (well connected words) $\beta_{DM} = \frac{1}{2}$ and $\gamma^2_{DM} = 3$ (again we marked the γ exponent with 2 for the second scaling regime of the degree distribution and DM for the DM - model), and for $s \sim t$ (less connected words) $\beta_{DM} = \frac{1}{2} + \frac{3}{2}$ and $\gamma^1_{DM} = 1.5$ [4]. Therefore distribution changes at a certain point called k_{cross}. This point can be estimated from the equation (6) by following reasoning. Second fraction in (6) for fixed time t becomes important if cs is of the same order as the constant $2m$. Because m is of the oder 1, to get k_{cross} we put cs equal 1. Thus from equation (6) k_{cross} is given by

$$k_{cross} \approx m(ct)^{\frac{1}{2}}(2m + ct)^{\frac{3}{2}}. \tag{8}$$

The DM model therefore explains two scaling regimes in the degree distribution of the CS word web. But, as has been measured by Cancho and Solé [1], the scaling exponent of the steeper part of the real word web is not $\gamma_{BA} = 3$, but somewhat lower ($\gamma = 2.7$). The difference between measured and theoretical values indicates, that there might be another processes, not included in the DM model, which influence the distribution.

3 WORD WEB MODEL

What are the other processes, which should be considered? Let us reason a little. New words are created and added to the vocabulary all the time, but not only this. The meaning of old words also develops in time. Some of them appear in a new context. Sometimes they loose some of their previous meanings, and get another.

In the network terminology the addition of a new context means the appearance of new edges between old words. This has been encountered in the DM model, leading to the two different scaling regimes. Loosing a meaning and getting a new one means, that some ends of old edges are rewired. Edge rewiring can be preferential, random, or a combination of both.

To fit the measured data in the CS word web [1, 5] a minimal model inspired by the DM model and by [10] was suggested [13]. Here are the processes included in our word web model:

- Each time unit a node is added and preferentially linked by m edges to the older nodes.
- Simultaneously ct new edges are created and linked preferentially among old nodes.
- In the same time m_r old nodes are randomly selected and one edge end of each of them is rewired preferentially.

If these processes run a long time, continuum approach [12] is good for their description. In this approximation $k(s,t)$ is a continuous variable. The dynamical equation describing above mentioned processes is as follows:

$$\frac{\partial k(s,t)}{\partial t} = (m + 2ct + m_r)\frac{k(s,t)}{\int_0^t k(s,t)ds} - \frac{m_r}{t} \tag{9}$$

The first addend in (9) defines preferential linking. The second one represents random selection of edge ends to be rewired. In the first addend the term $m\frac{k(s,t)}{\int_0^t k(s,t)ds}$ represents preferential linking of m new edges, $2ct\frac{k(s,t)}{\int_0^t k(s,t)ds}$ describes preferential linking od $2ct$ new edge ends among old nodes and $m_r\frac{k(s,t)}{\int_0^t k(s,t)ds}$ tells that m_r edge ends are rewired preferentially.

To solve this equation, the integral $\int_0^t k(s,t)ds$, giving the sum of all degrees in the net, needs to be specified. This sum is influenced only by the new link creation; rewiring left it unaffected. Because the only edge creation processes are the same as in the DM model, the integral is given by (5).
Substituting (5) into (9) the equation (9) is reformulated:

$$\frac{\partial k(s,t)}{\partial t} = (m + 2ct + m_r)\frac{k(s,t)}{2mt + ct^2} - \frac{m_r}{t} \tag{10}$$

This is a simple linear first order differential equation of the type

$$\frac{\partial y}{\partial x} = -f_1(x)y - f_2(x) \tag{11}$$

with the solution

$$y = e^{-\int f_1(x)dx}\Big[\phi - \int f_2 e^{\int f_1(x)}dx\Big]. \tag{12}$$

Using (12) we get

$$k(s,t) = \left(\frac{t}{s}\right)^A \left(\frac{2m+ct}{2m+cs}\right)^{2-A} g(s,t) \tag{13}$$

where $A = \frac{m+m_r}{2m}$ and

$$g(s,t) = \frac{1}{m^2 - m_r^2}\Big[m + \frac{m_r}{m^2 - m_r^2}[M_1 + M_2\left(\frac{s}{t}\right)^A \left(\frac{2m+cs}{2m+ct}\right)^{2-A}]\Big], \tag{14}$$

where $M_1 = (2m + cs)(m - m_r + cs)$ and $M_2 = (2m + ct)(m - m_r + ct)$.
Then the leading term of (13) is $\left(\frac{t}{s}\right)^{\frac{m+m_r}{2m}} \left(\frac{2m+ct}{2m+cs}\right)^{2-\frac{m+m_r}{2m}}$. If $m \neq m_r$, $g(s,t)$ (14) doesn't influence the solution too much. From (13) it is clear, that

-if $s << t$, $\beta = \frac{m+m_r}{2m}$ and $\gamma = 2 + \frac{m-m_r}{m+m_r}$,

-but if $s \sim t$, $\beta = 2 - \frac{m+m_r}{2m} + \frac{m+m_r}{2m} = 2$ and thus $\gamma = 1.5$.

In the model (9) scaling exponent γ is lower then the value $\gamma_{BA} = 3$ in the region of great k, but maintains the value 1.5 in the region of small k-s. This is exactly what has been measured by Cancho and Solé [1]. One of us has recently used this model to explain their results [13].

4 NUMERICAL STUDIES OF THE WORD WEB

Our next goal is to test if the model (9) fits the distribution of the real word web of the special kind. Namely, we want to know, whether the word web based on an ancient text, which does not change a long time, has the same two-modal scaling, as the web based on the modern English vocabulary. Our expectation was that such an ancient text consist mostly only of the words from the kernel lexicon and thus the one scaling regime would support the hypothesis of Cancho and Solé that the kernel lexicon is responsible for the steeper regime of the degree distribution in their experiment [1].

To do this, we created a positional word web on the basis of several versions of an English translations of The Bible (BWW net). First the small world properties of each BWW were measured. All parameters are collected in the $Tab.1$. As shown, all word webs combine high clustering with low node separation, which is typical for the small world networks [6].

[h]

version	N	$\overline{c}$	ℓ	$\overline{k}$
kjv	11592	0.771	2.18	47
drv	11379	0.772	2.18	47
asv	10077	0.778	2.18	47
nrsv	14717	0.718	2.24	50
bev	4942	0.774	2.12	70
prg	21104	0.700	2.27	49

Table 1. Properties of positional BWW. Here N is the number of distinct words in the text, $\overline{c}$ is the BWW clustering coefficient, ℓ denotes the node separation, $\overline{k}$ the average node degree. We have used several versions of The Bible [20]. Some of them, such as King James version (kjv), Douay Rheims version (drv) are old (kjv has been issued in the year 1711, drv is even older, 1582), the others (American Standard version, asv, 1901; Basic English version, bev, 1941; New Revisited Standard version, nrsv, 1989) are relatively modern. bev is special, because its text has been artificially simplified. It is reflected in slightly different parameters in the table. prg is the word web created from the selected books found in the Project Gutenberg web page [16]. This has been added for comparison of the parameters of the ancient and the modern text.

Because we did not find two scaling regimes in any BWW network $Fig.$ 1, there are two possibilities how to explain this fact. First - the hypothesis of Cancho and Solé about different dynamics of the kernel and non-kernel lexicon is correct (which contradicts the DM model), or there is another reason. We have therefore generated artificial networks, programming the processes proposed in our model, which is introduced in the previous section (9). The parameters of the model were taken from the kjv and the prg word webs. In the experiment of Cancho and Solé [1], there are no multiple edges considered. The analytical equations (4,9,10), however, does not exclude them.

To test the role of multiple edges, artifical networks (having $N = 20000$ nodes) were developed, once with allowed and also disallowed multiple edges. Degree distributions of all networks are depicted on $Fig.$ 2, together with corresponding average scaling exponents. All of the networks have the power law degree distribution indicating scale free structure, but there are no two scaling regimes present.

The most probable reason of this discrepancy between the theory (9) and the data is, that our data sets are too small. In the DM model, there is a crossover point [4] between the two scaling regimes. To get the estimate of the k_{cross} parameter for our model, we can use the same reasoning (8) as for the DM model. Thus the second regime appears when the second fraction of equation (13) becomes important. This happens if cs is of the same order as the constant $2m$. Thus we get estimation of the k_{cross} parameter for our model

$$k_{cross} \approx m(ct)^{\frac{m+m_r}{2m}}(2m+ct)^{2-\frac{m+m_r}{2m}}. \tag{15}$$

Calculating k_{cross} for our networks with $N = t = 20000$ nodes and kjv and prg parameters ($m = 4, c = 0.003$ for kjv and 0.002 for prg , $m_r = 2$) we get k_{cross} values out of the k - range of the $Fig.$ 3. To test this hypothesis, we decreased parameter c in (15) to get lower k_{cross}. Now, as seen on $Fig.$ 4, crossover point is clearly visible. It is therefore true, that the huge amount of nodes or small parameter c is important to get the visible crossover point.

5 CONCLUSION

In conclusion, we present a model of growing network, that explains the difference between the exponent of the steeper part of the degree distribution predicted by the DM model [4] and the value measured by Cancho and Solé [1] in their positional word web. Our model includes additional event, such as preferential edge rewiring (9). In the word web terminology this process means, that certain word looses one of its meanings, or contexts, and gains a different one. Although our model of growing network was inspired by the language data, it has a general relevance to all networks growing by the included processes.

To verify the validity of our model for another real word webs, we developed positional word web for several English translations of The Bible. Using our network creating processes (9), we also generated several networks in which we first disallowed and then allowed multiple edges between nodes. For all networks degree distribution has been measured, showing no two scaling regimes $Fig.$ 2. We have shown, that the reason lies in too small network sizes, or too big c parameter (8) (see. $Fig.$ 4). Because k_{cross} in $Fig.$ 2 is out of the k - range, we have to compare scaling epxonents to $\gamma_{DM}^{1} = 1.5$. This result is in an agreement with Dorogovtsev and Mendes [4], as they have stated that two scaling regimes is a result of network dynamics and not the difference between the kernel and non-kernel lexicon as it was speculated by Cancho and Solé [1]. For the BWW network the scaling exponent is close to 1.7, which is different from the predicted value. We do not know the reason of this discrepancy. It is possible, that BWW - s are too small to get a good and long enough linear part in the degree distribution, and thus more accurate γ measurement. The Bible network can not be larger, as the text of The Bible contains only limited vocabulary as it is possible to see in Tab. 1.

To retrieve as many information about the structure of the word webs as possible, we have also examined distribution of clustering coefficients. More precisely, we have measured the distribution $c(k)$, where $c(k)$ is an average clustering coefficient of all nodes with a degree k. This reflects the modularity of the network [14]. Power law $c(k)$ distribution is an indication of the hierarchical structure of nodes [15, 18]. We have measured the $c(k)$ distribution of the BWW and also of generated networks Fig. 3. In all cases power law tail is present in the distribution, which means that only last points of the distribution have the character of a power law distribution. Moreover, the scaling exponent is quite low (≈ 0.6), indicating only a weak hierarchy for the well connected nodes.

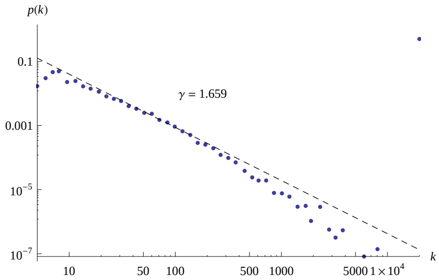
In majority of networks in Nature, node hierarchy is accompanied by the scale free property (3). Natural question therefore arises. What are the processes leading to the strongly hierarchical scale free network structure? This has been partly studied in [17, 18], but the question is still not completely answered.

Acknowledgment: We acknowledge discussions with prof. Pavol Brunovský, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia and with prof. Boris Rudolf, Faculty of Electrotechnics and Electrical Engineering, Bratislava Slovakia. This work has been supported by VEGA grants NO 1/2045/05, 2/4026/04, 1/3612/06, 1/0361/08 and 2/7087/27.

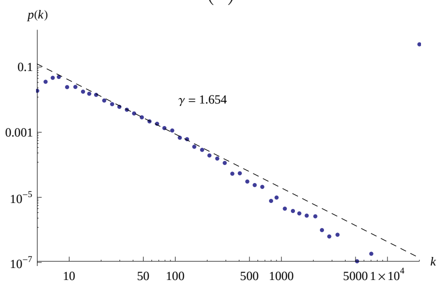
REFERENCES

- [1] Cancho, R. F., Solé, R. , The Small World of Human Language, Proc. Royal Soc. London B **268** (2001) 2261 -2265
- [2] Motter, A. E., de Moura, P. S., Ying - Cheng Lai, Dasgupta, P.,Topology of the Conceptual Network of Language, Phys. Rev. **E65** (2002) R 065102
- [3] Steyvers, M., Tenenbaum, J. B., The Large - Scale Structure of Semantic Networks: Statistical Analysis and a Model of Semantic Growth, Cognitive Science **29** (2005) 41 - 78
- [4] Dorogovtsev, S. N., Mendes, J. F. F., Language as an Evolving Word Web, Proc. Royal Soc. London B **268** (2001) 2603 - 2606
- [5] Markošová, M., Náther. P., Language as a Graph (in Slovak), in Mind, Intelligence and Life, eds. Kvasnička, V., Trebatický, P., Pospíchal, J., Kelemen, STU Bratislava (2007) 298 - 307
- [6] Watts, D. J., Small Worlds, Princeton University Press, Princeton (2004)
- [7] Zanette, D. H., Dynamics of Rumour Propagation on Small World Networks, Phys. Rev. **E65** (2002) 41908 - 41917
- [8] Sen, P. and others, Small World Properties of the Indian Railway Network, Phys. Rev. **E67** (2003) 36106 - 36111, cond - mat/0208535 v1 (2002)
- [9] Barabási, A. L., Albert, R., Emergence of Scaling in Random Network, Science **286** (1999) 509 - 512
- [10] Dorogovtsev, S. N., Mendes, J. F. F., Evolution of Networks, Adv. Phys. **51** (2002) 1079 - 1145
- [11] Albert, R., Barabási, A. L., Statistical Mechanics of Complex Networks, Rev. Mod. Phys **74** (2002) 47 - 101, cond - mat/010696 v1 (2001)
- [12] Dorogovtsev, S. N., Mendes, J. F. F., Scaling Properties of Scale - Free Evolving Networks: Continuous Approach, Phys. Rev. **E63** (2001) 56125 - 56146, cond - mat/001009 v1 (2000)
- [13] Markošová, M., Network model of human language, Physica A **387** (2008) 661 - 666
- [14] Almaas, E., Barabási, A. L., Power laws in Biological Networks, q-bio/0401010 (2004)
- [15] Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barab?si, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. Science **297** (2002) 1551-1555
- [16] www.gutenberg.org We have used texts from four, randomly chosen, books (Donald Grant by George MacDonald, Dora Thorne by Charlote Braeme and Education of Henry Adam by Henry Adam).
- [17] Vásquez, A., Growing networks with local rules, Phys. Rev. E **67** (2003) 056104
- [18] Markošová, M., Náther, P., Rudol, B., Hierarchy in the growing scale - free networks with local rules, to appear in Physica A.
- [19] Chialvo, D. R., Critical brain networks, Physica A **340** (2004) 756
- [20] http://unbound.biola.edu.
- [21] Albert, R. and Jeong, H. and Barabási, A. L., The diameter of the world wide web, Nature **130–131**, (1999) 401

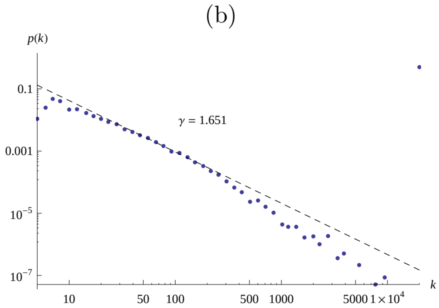
[22] KLEINBERG, J. THE SMALL-WORLD PHENOMENON: an algorithm perspective, STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing, **163–170** (2000) ACM



(a)



(b)



(c)

Fig. 1. Degree distributions of the word webs, with their γ exponents. (a) kjv bible version network, (b) American Standard version and (c) new revisited standard version. This plot shows lack of two scaling regimes in degree distributions of various BWW networks.

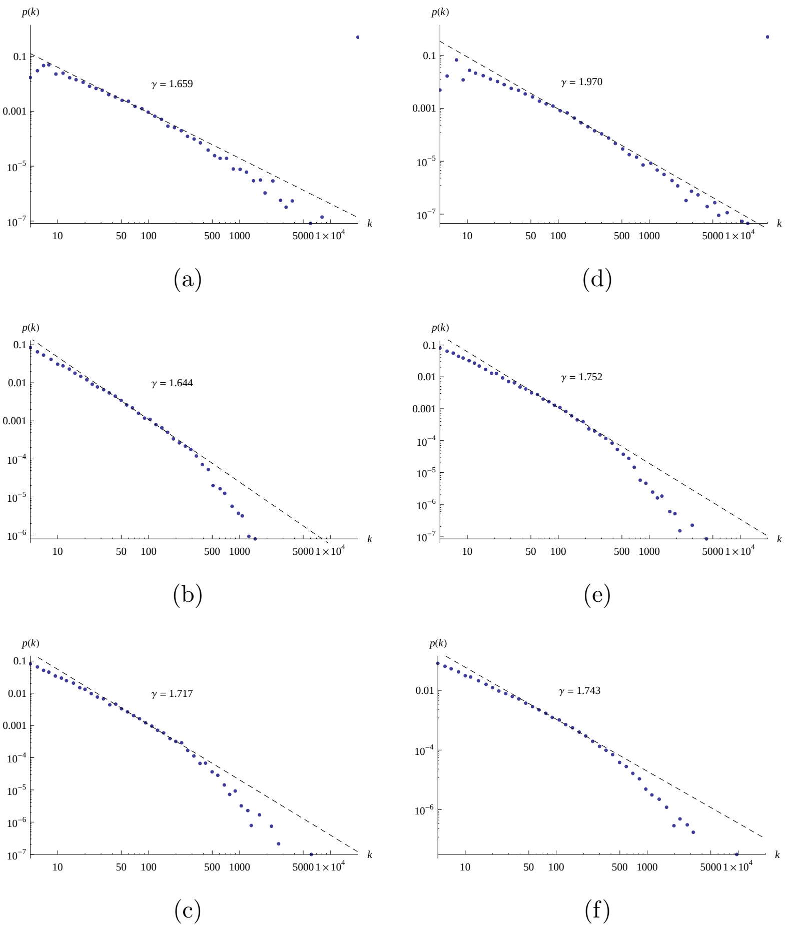


Fig. 2. Degree distributions of the word webs, with their γ exponents. (a) kjv bible version network, (b) generated network, without multiple edges allowed (parameters were following: $t = 11592$, $m = 4$, $c = 0.003$, $m_r = 2$) (c) network generated with the same parameters, but multiple edges allowed. (d) prg network. (e) generated network, without multiple edges allowed (parameters were following: $t = 21104$, $m = 4$, $c = 0.002$, $m_r = 2$) (f) generated with the same parameters, but multiple edges allowed. Parameters of the generated networks were chosen to generate networks with the characteristics similar to the kjv and prg networks

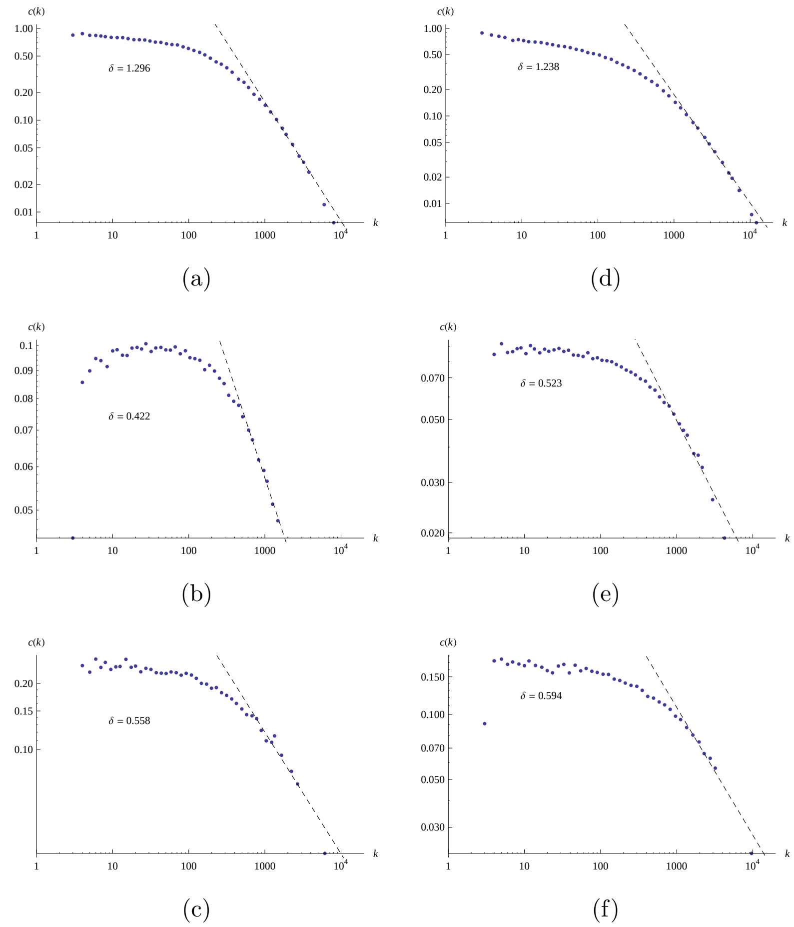


Fig. 3. $c(k)$ distributions of the word webs. (a) kjv bible version network, (b) generated network, without multiple edges allowed (parameters were following: $t = 11592$, $m = 4$, $c = 0.003$, $m_r = 2$) (c) network generated with the same parameters, but multiple edges allowed. (d) prg network. (e) generated network, without multiple edges allowed (parameters were following: $t = 21104$, $m = 4$, $c = 0.002$, $m_r = 2$) (f) generated with the same parameters, but multiple edges allowed. Parameters of the generated networks were chosen to generate networks with the characteristics similar to the kjv and prg networks. All distribution have power-law regime with the exponent ≈ 0.6 for large k.

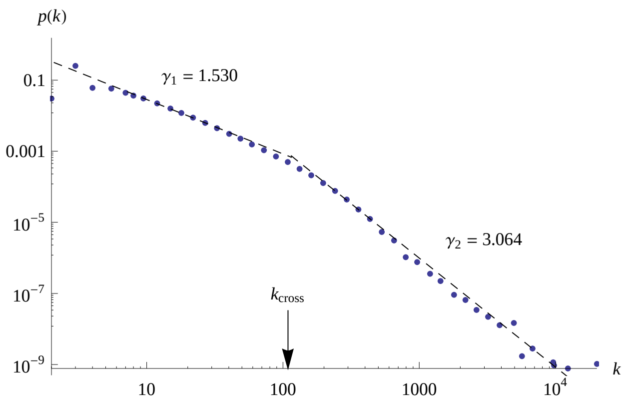


Fig. 4. Degree distributions of the generated network with no multiple edges allowed. Parameters were: $t = 500000$, $m = 3$, $c = 0.000033$, $m_r = 3$. With a low parameter c and large size, the crossoverpoint k_{cross} is present in the distribution.