Empirical Validation of the Buckley–Osthus Model for the Web Host Graph: Degree and Edge Distributions

Maxim Zhukovskiy Dmitry Vinogradov Yuri Pritykin Liudmila Ostroumova
Evgeniy Grechnikov Gleb Gusev Pavel Serdyukov Andrei Raigorodskii

Yandex
16 Leo Tolstoy St., Moscow, 119021 Russia

{zhukmax, dimavin, pritykin, ostroumova-la, grechnik,
gleb57, pavser, raigorodsky}@yandex-team.ru

August 14, 2012

Abstract

There has been a lot of research on random graph models for large real-world networks such as those formed by hyperlinks between web pages in the world wide web. Though largely successful qualitatively in capturing their key properties, such models may lack important quantitative characteristics of Internet graphs. While preferential attachment random graph models were shown to be capable of reflecting the degree distribution of the webgraph, their ability to reflect certain aspects of the edge distribution was not yet well studied.

In this paper, we consider the Buckley–Osthus implementation of preferential attachment and its ability to model the web host graph in two aspects. One is the degree distribution that we observe to follow the power law, as often being the case for real-world graphs. Another one is the two-dimensional edge distribution, the number of edges between vertices of given degrees. We fit a single "initial attractiveness" parameter a of the model, first with respect to the degree distribution of the web host graph, and then, absolutely independently, with respect to the edge distribution. Surprisingly, the values of a we obtain turn out to be nearly the same. Therefore the same model with the same value of the parameter a fits very well the two independent and basic aspects of the web host graph. In addition, we demonstrate that other models completely lack the asymptotic behavior of the edge distribution of the web host graph, even when accurately capturing the degree distribution.

To the best of our knowledge, this is the first attempt for a real graph of Internet to describe the distribution of edges between vertices with respect to their degrees.

Keywords: web host graph, preferential attachment, random graph models, Buckley–Osthus random graphs, power law degree distribution, assortative mixing, edge distribution with respect to vertex degrees

1 Introduction

The study of the web as a hyperlink graph yields a valuable insight into web algorithms for crawling, searching, and community discovery [11, 23, 33]. Valid random graph models of the web

*Short version of this paper will be published in the proceedings of CIKM 2012 (see http://www.cikm2012.org/).

provide methods of generating WWW-like graphs that are significantly smaller and simpler than real WWW graphs, but yet preserve certain key properties of the hyperlink structure of the web. Such artificial graphs could serve as a convenient experimental platform, where new approaches to search, indexing, compression can be evaluated.

Vertices of the *webgraph* correspond to web pages and edges represent hyperlinks between them. Webgraphs have been extensively studied with respect to many quantitative aspects such as degree distribution, diameter, number of connected components, macroscopic structure, and assortative mixing (e.g., see [3, 6, 12, 19, 24, 33, 38]).

In this paper, we consider the *web host graph*. Vertices of this graph are web hosts and edges correspond to hyperlinks between their pages. The web host graph is much smaller than the webgraph, but is still a very useful resource and an abstraction of the web. For a lot of purposes, modern search engines consider hosts (and web sites associated with them) rather than web pages as the smallest possible entities in the web. Particularly, the smaller size of the web host graph allows for simpler and more efficient link analysis useful for web search related tasks.

We study the web host graph from two perspectives.

First, we look at the distribution of degrees of the web host graph vertices. It was shown that degrees of the webgraph vertices, much like in many other real world networks [22], obey the power law [3, 6, 12, 24]. Albert et al. were the first to find the power law in the degree distribution of the web pages in the domain *.nd.edu [3]. Not surprisingly, we observe that the degree distribution in the web host graph also follows the power law.

Second, we study how edges in the web host graph are distributed between vertices depending on degrees of these vertices. To the best of our knowledge, this is the first study of this property for graphs of Internet. However, in a reduced form, this notion was previously studied under the name of *degree correlation* or *assortativity* (e.g., see [6]). A convenient way of capturing the degree correlation is by examining the properties of $d_{nn}(d)$, the expected average degree of neighbors of a random vertex with degree d. In real-word networks, one often observes $d_{nn}(d) \sim d^{\delta}$ with some δ, which is negative for the webgraph (disassortativity) [38] and usually positive for social networks (assortativity) [36]. We study a more general property of *the distribution of edges* between vertices with respect to their degrees, that is, the total number of edges $X(d_1, d_2)$ between pairs of vertices with degrees d_1, d_2. In fact, one can obviously derive both degree distribution and $d_{nn}(d)$ from the edge distribution: $\#(d) = 1/d \sum_{d_1} X(d_1, d)$, $d_{nn}(d) = \frac{\sum_{d_1} d_1 X(d,d_1)}{\sum_{d_1} X(d,d_1)}$. On the other hand, the degree distribution of a real-world graph does not determine its edge distribution, and therefore the latter may be considered as an additional more general aspect of the graph. In fact, there are assortative and disassortative real-world graphs with close power-law degree distributions, and therefore their edge distributions do differ as well.

There are a number of important random graph models whose features (such as degree distribution, diameter, etc.) are supposed to be close to those of the real Internet graphs and social networks. Barabási and Albert proposed the most well-known approach [4] that was realized in various preferential attachment models. Several of them have precise mathematical definitions — the Bollobás and Riordan model [10], the Buckley and Osthus model [14, 20, 21], the Copying Models [30, 32], Directed Scale-Free Graphs [7] and the general model of Cooper and Frieze [18] (see also [32]). An extensive review of these models can be found elsewhere (e.g., see [6, 8]).

We focus on random graph models that allow for mathematical analysis of their properties and thus have to be relatively simple. Bollobás and Riordan were the first to propose a precisely defined preferential attachment model, and proved the power law for degree distribution in this model with

mathematical rigor. It was shown that the number of vertices with degree d decreases as $d^{-\gamma}$ with $\gamma = 3$ [9]. On the other hand, Barabási and Albert empirically estimated this exponent in the real webgraph to be approximately 2.1 ± 0.1 [4]. As we show in this paper, the parameter γ for the web host graph is also far from 3. Therefore the model of Bollobás and Riordan is not realistic for both the webgraph and the web host graph. The random graph model of Buckley and Osthus, which is a generalization of the Bollobás–Riordan model, solves this problem. Namely, the Buckley–Osthus model depends on a parameter a called *initial attractiveness*. For an integer value of a, Buckley and Osthus proved theoretically that the degree distribution in this model also follows the power law with the exponent $-2 - a$ [14].

Recently Grechnikov obtained two results concerning the Buckley–Osthus model [26]. First, he extended the result from [14] about the degree distribution to an arbitrary positive a, not necessarily integer. Second, he obtained an accurate asymptotic estimate for the edge distribution for growing degrees as an explicit formula depending on a as a parameter.

We expect the Buckley–Osthus model to be a good approximation of the web host graph to a certain extent. Relying on the both aforementioned theoretical results concerning two different aspects of the Buckley–Osthus model, we find the best fit of the initial attractiveness parameter a for the real web host graph assuming it is generated in the model. First, we choose the value of the parameter a so that the exponent in the power law for the real web host graph is close to $-2 - a$. In a second approach, completely independent from the first one, we estimate a using the best fit of the formula from [26] for the edge distribution in the Buckley–Osthus model to the really observed edge distribution in the web host graph. Surprisingly, in both cases we find out that the model agrees very well with the real graph with the same value of $a \approx 0.3$. In other words, this very same model with $a \approx 0.3$ accurately approximates two completely different and a priori independent basic aspects of the web host graph, degree and edge distributions (and therefore also assortativity). This is especially impressive as the model itself is very simple and has only a single degree of freedom, namely the initial attractiveness parameter a. Note that we were able to describe the distribution of edges in a real web host graph only using the model for this graph and properties of this model obtained theoretically. Without the model, to come up with the asymptotics of the edge distribution would be really hard, if not impossible.

We compare the Buckley–Osthus model and its ability to describe the real web host graph with other random graph models. We focus on the models that generate graphs with the degree distribution following the power law. We demonstrate that degrees do not immediately relate to edges between vertices given their degrees. Even after being fit to the web host graph with respect to degree distribution, the models are not able to capture the edge distribution and in fact completely lack the asymptotic behavior of this edge distribution observed in the web host graph and in the Buckley–Osthus model.

To sum up, contributions of this paper are the following:

- We have studied the distribution of degrees and the distribution of edges between vertices given their degrees for the web host graph.
- Relying on previous theoretical study, we demonstrated that the web host graph corresponds very well to the Buckley–Osthus random graph model with the initial attractiveness parameter $a \approx 0.3$ with respect to the degree distribution, the edge distribution and (consequently) the assortativity. We obtained the same value of the parameter two times by fitting the model with respect to the both independent quantitative aspects: the degree and the edge distributions.

- We generated graphs in the Buckley–Osthus model and empirically examined their theoretically proved properties in practice. We also showed that some other random graph models fail to capture the edge distribution of the web host graph though may successfully capture the degree distribution.

To the best of our knowledge, this is the first attempt to empirically validate the Buckley–Osthus model on the real web data. Moreover, this is the first attempt to rigorously study the distribution of edges between vertices with respect to their degrees for graphs of Internet.

The remainder of the paper is organized as follows. Section 2 is a short review on random graph models. In particular, in Section 2.2 we describe the theoretical properties of the Buckley–Osthus model critical for our experiments with the web host graph. We describe the experimental part of the work in Sections 3 and 4. In Section 3, we report the results of the approximation of the web host graph by the Buckley–Osthus model with respect to degree and edge distributions. In Section 4, we compare it with approximations by other models. In Section 5 we discuss potential applications and future work.

2 Random Webgraph Models

One possible theoretical approach to what the model of a webgraph might be is the mathematical concept of random graphs. The essence of this approach is in the idea of a webgraph developing stochastically. Once the rules or the parameters of this stochastic process are precisely specified, a random graph may obtain (sometimes unexpected) stable properties, in spite of the stochastic nature of its formation. Some of such properties may reflect those of the real webgraph rather accurately.

There have been a lot of attempts to model the hyperlink graph of the web as a random graph. Probably the simplest (even regardless of the webgraph) are random graphs of the Erdős–Rényi model, where a graph is constructed by creating a fixed number of vertices and a fixed number of edges drawn independently uniformly at random over pairs of vertices. However, this model is not suitable for the webgraph (as well as the web host graph) as it lacks scalability, that is, does not have a power law degree distribution.

In 1999, Barabási and Albert [4] observed that the degree distribution of the real webgraph follows the power law with the exponent approximately equal to -2.1. They proposed a concept of preferential attachment that explained the phenomenon. The basic underlying idea is the following. A graph is constructed with a random process. At each step of the process, a new vertex is added and a fixed number of edges are added from the new vertex to randomly chosen already existing vertices. Vertices with higher degree acquire new edges with higher probability that linearly depends on their degree ("rich get richer").

2.1 Preferential attachment models

The general idea of preferential attachment obtained a precise mathematical formulation in the model of Bollobás and Riordan [9] defined in the following way. We construct a series of graphs (Markov chain) $G_m^n, n = 1, 2, \ldots$, with n vertices and mn edges, where $m \in \mathbb{Z}$ is a fixed number. Let us consider the case $m = 1$ first. Let G_1^1 be a graph consisting of one vertex with a self-loop. A graph G_1^t is obtained from G_1^{t-1} by adding a vertex t and an edge from t to a vertex i, where i

is chosen randomly within the existing vertices with the following probability distribution:

$$\mathrm{P}(i=s)=\begin{cases} d_{G_1^{t-1}}(s)/(2t-1) & \text{if } 1\leqslant s\leqslant t-1,\\ 1/(2t-1) & \text{if } s=t, \end{cases}$$

where $d_{G_1^t}(s)$ denotes the degree of the vertex s in G_1^t. A graph G_m^n is constructed from G_1^{mn} by merging the vertices $1,\ldots,m$ into the vertex 1 of the new graph, merging the vertices $m+1,\ldots,2m$ into the vertex 2 of the new graph, etc. Note that one can consider the variant of the model with directed graphs: in this case, an edge between the vertices i and j goes from i to j if $i>j$.

The Bollobás–Riordan model accurately captures some of the key properties of different real-world graphs. For instance, "small-world phenomenon" of many real-world networks, i.e., a surprisingly small diameter, is also observed in the model. Bollobás and Riordan proved that indeed the diameter of G_m^n is about $\frac{\log n}{\log\log n}$ for large n [10]. They also showed that the degree distribution of G_m^n obeys the power law: the number of vertices with degree d in the model is well approximated by $d^{-\gamma}$, with $\gamma=3$ [9]. However, this disagrees with the webgraph where the estimate $\gamma_{WWW}=2.1\pm0.1$ was observed [4]. This means that even though the Bollobás–Riordan model is similar in some aspects to real graphs of Internet qualitatively, it needs to be refined to better capture the reality quantitatively. In this work, we estimate the value of γ_{Host} in the power law for the web host graph to be approximately 2.276 ± 0.001 (see Section 3.4).

A possible approach for such a refinement is the model proposed independently by two groups of researchers [20, 21]. They proposed to extend the model with a parameter called *initial attractiveness* of a vertex, a positive constant that does not depend on degree. Later Buckley and Osthus gave an explicit construction of this model [14]. The degree distribution of a Buckley–Osthus graph also obeys the power law, but now varying the value of a in the definition of the model, one can tune the exponent γ in the power law of the resulting graph.

More specifically, the model generates a series of graphs $H_{a,m}^n, n=1,2,\ldots$, with n vertices and mn edges, where $m\in\mathbb{Z}$ is a fixed number. The definition of $H_{a,1}^n$ recapitulates the definition of G_m^n, and the only difference is that the probability of a newly added edge in $H_{a,1}^n$ equals

$$\mathrm{P}(i=s)=\begin{cases} \frac{d_{H_{a,1}^{t-1}}(s)+a-1}{(a+1)t-1} & \text{if } 1\leqslant s\leqslant t-1,\\ \frac{a}{(a+1)t-1} & \text{if } s=t. \end{cases}$$

A graph $H_{a,m}^n$ is obtained from $H_{a,1}^{mn}$ in the same manner as G_m^n from G_1^{mn}. Note that for $a=1$, we obtain the initial Bollobás–Riordan model G_m^n. For an integer a, Buckley and Osthus proved [14] that the degree distribution of a random graph in the model follows the power law with $\gamma=2+a$.

Previously, the Buckley–Osthus model has not been compared with real graphs. In Section 2.2, we present further properties of the Buckley–Osthus model obtained recently and then use them in Section 3 for comparison of this model with the real web host graph.

2.2 Properties of the Buckley–Osthus Model

In this section, we present recent theoretical results on degree and edge distributions of the Buckley–Osthus random graph model [26]. Our experiments with real graphs (Section 3) are based on the results from this section.

2.2.1 Degree distribution

The following theorems show the dependence of the degree distribution in the Buckley–Osthus model on the initial attractiveness parameter a and generalize the results of [14].

For a given pair of functions f, g, we say that $f(n) = O(g(n))$ if there exists a positive real number C such that $|f(n)| \leqslant Cg(n)$ for sufficiently large n. We also say that $f(n) = o(g(n))$ if f is dominated by g asymptotically and $f(n) = \omega(g(n))$ if f dominates g asymptotically.

Let $\#_a(d, n)$ be the number of vertices with degree d in the model $H^n_{a,m}$. We denote by $\mathrm{E}X$ the expectation of a random variable X .

Theorem 1 ([26]) *For $d \geqslant m$ and for every fixed positive a,*

$$\mathrm{E}\left(\#_a(d, n)\right) = \frac{\mathrm{B}(d - m + ma, a + 2)}{\mathrm{B}(ma, a + 1)} n + O\left(\frac{1}{d}\right).$$

Here $\mathrm{B}(x, y)$ is the beta function. Note that

$$\frac{\mathrm{B}(d - m + ma, a + 2)}{\mathrm{B}(ma, a + 1)} \sim Cd^{-2-a}$$

as $d \to \infty$ with $C = (a + 1)\frac{\Gamma(ma+a+1)}{\Gamma(ma)}$, where $\Gamma(x)$ is the gamma function (an extension of the factorial function).

We say that a certain property holds **whp** (with high probability) if the probability of this property tends to 1 as $n \to \infty$. The following concentration result shows that the degree distribution obeys the power law with $\gamma = 2 + a$.

Theorem 2 ([26]) *Consider $d \geqslant m$ to be the value of a function of n and $\psi(n)$ to be a function tending to infinity arbitrarily slowly. Then **whp** we have*

$$\left|\#_a(d, n) - \frac{\mathrm{B}(d - m + ma, a + 2)}{\mathrm{B}(ma, a + 1)} n\right| \leqslant$$

$$\leqslant \left(\sqrt{d^{-a-2}n} + d^{-1}\right) \psi(n).$$

In contrast with the result of [14], a is not necessarily integer here. Roughly speaking, Theorems 1 and 2 imply that for large d, we have

$$\#_a(d, n) \sim b_1 d^{-2-a} n \tag{1}$$

with some constant b_1 in an appropriate range of degrees d.

2.2.2 Edges between vertices of given degrees

In this subsection, we report the results capturing the behaviour of $X_a(d_1, d_2, n)$, the total number of edges between vertices of degree d_1 and vertices of degree d_2 in a Buckley–Osthus graph. For $d_1 = d_2$, we count every edge twice, but we do not count self-loops. We use this function in Section 3 comparing the web host graph with some random graph models including the Buckley–Osthus model.

The number of edges between vertices of given degrees in the Buckley–Osthus model can be estimated in the following way.

Theorem 3 ([26]) *Consider d_1, d_2 to be the values of two functions of n tending to infinity as n grows. Then*

$$\mathrm{E}X_a(d_1, d_2, n) = c_a(d_1, d_2)n + O(1),$$

where

$$c_a(d_1, d_2) = ma(a+1)\frac{\Gamma(ma+a+1)}{\Gamma(ma)}\frac{(d_1+d_2)^{1-a}}{d_1^2 d_2^2}\times$$
$$\times\left(1 + O\left(\frac{1}{d_1} + \frac{1}{d_2} + \frac{d_1 d_2}{(d_1+d_2)^2}\right)\right).$$

The following theorem is a concentration result.

Theorem 4 ([26]) *Let $c > 0$. Then*

$$\mathrm{P}\left(|X_a(d_1, d_2, n) - \mathrm{E}X_a(d_1, d_2, n)| \geqslant c(d_1+d_2)\sqrt{mn}\right) \leqslant$$
$$\leqslant 2\exp\left(-\frac{c^2}{8}\right).$$

In particular, for an arbitrary function $c(n)$ tending to infinity as n grows we have ***whp*** $|X - \mathrm{E}X| < c(n)(d_1+d_2)\sqrt{mn}$.

Thus it follows that $\mathrm{E}X_a(d_1, d_2, n)$ behaves as

$$(d_1+d_2)^{1-a} d_1^{-2} d_2^{-2} n \tag{2}$$

if the ratio $\max(d_1, d_2)/\min(d_1, d_2)$ is sufficiently large (otherwise this formula does not capture the asymptotic behavior).

In Section 3, we also use the fact that the number of loops and multiple edges in the Buckley–Osthus random graph is considerably smaller than the total number of edges. To be more precise, the following statement holds.

Proposition 1 *For every $0 < a < 1$ we have*

$$\mathrm{E}N(\text{loops in } H_{a,m}^n) = O(\ln n),$$

$$\mathrm{E}N(\text{multiple edges in } H_{a,m}^n) = O\left(n^{1-a}\right).$$

Proposition 1 is proved in Appendix. Here we denote by $N(\text{loops in } H_{a,m}^n)$ the number of loops in $H_{a,m}^n$ and denote by $N(\text{multiple edges in } H_{a,m}^n)$ the number of multiple edges in the random graph. Recall that the total number of edges in $H_{a,m}^n$ is mn and dominates both n^{1-a} and $\ln n$.

2.3 Other results related to edge distribution

The distribution of edges between vertices with respect to their degrees is closely related to an interesting quantitative characteristic of graphs called *assortativity*, or *degree correlation* [6, 19, 36, 37, 38]. Informally, a graph is *assortative* (has a positive degree correlation) if vertices of high degree tend to connect with vertices of high degree. On the other hand, a graph is *disassortative* (has a negative degree correlation) if vertices of high degree tend to connect with vertices of low degree. A convenient way of capturing the degree correlation is by examining the properties of $d_{nn}(d)$, the average degree of neighbors of a vertex with degree d (first average over neighbors of a vertex, and then over vertices of degree d). In real-word networks, often $d_{nn}(d) \sim d^{\delta}$ with some δ, which is negative for the webgraph (disassortativity) [38] and usually positive for social networks (assortativity) [36]. Disassortativity of protein networks was studied in [34].

The function $X(d_1, d_2)$ of edge distribution, the number of edges between vertices of degrees d_1, d_2, may be considered as a generalization of $d_{nn}(d)$. In fact, the latter can be restored from the former:

$$d_{nn}(d) = \frac{\sum_{d_1} d_1 X(d, d_1)}{\sum_{d_1} X(d, d_1)}.$$

As mentioned in [36] and [38], networks in the Barabási–Albert model [2] have $d_{nn}(d) \approx \text{const}$ and thus do not demonstrate assortative mixing. However, it can be shown experimentally that a graph in the Buckley–Osthus model, that is a generalization of the Barabási–Albert model, may demonstrate assortativity (for $a > 1$) or disassortativity (for $a < 1$). In particular, we compare the web host graph and the Buckley–Osthus model with $a \approx 0.3$ with respect to the function $d_{nn}(d)$ in Section 4 and find them to be close to each other (see Fig. 6).

It was claimed in [35] that the negative degree correlation may be explained by the model where a graph is chosen uniformly at random from the set of all graphs with a prescribed power-law degree distribution without multiple edges. The authors stated that the resulting graph will have a negative degree correlation with high probability, for vertices of high degree are forced to connect each other rarely, or otherwise multiple edges will be more likely to appear. The authors of [37] obtain some theoretical results for a similar model. They also argue that the graph of Internet is disassortative. In [15] the assortative co-authorship graph is modeled. The proposed model is based on preferential attachment, with an additional idea of adding new links between already existing vertices chosen based on their degrees. This idea can be utilized for modeling both assortative and disassortative graphs. However, in contrast with the Buckley–Osthus model, these models are not based on any natural rules that would explain the underlying laws of graph formation. They are rather specific and thus may be suspected in "overfitting" when approximating real graphs.

3 Experiments on the web host graph

3.1 Preliminaries

Let us consider a random graph in the Buckley–Osthus model $H^n_{a,m}$. For simplicity, we ignore edge directions, merge multiple edges and remove loops. Due to Proposition 1, the difference between the obtained graph H and the initial one is not important for us and Theorems 1, 2, 3, 4 are still applicable to H.

In what follows, we denote by $\#(d)$ the number of vertices of degree d and by $X(d_1, d_2)$ the total number of edges between all vertices of degree d_1 and all vertices of degree d_2. The following

two properties follow from equations (1), (2).

1) The function $\#(d)$ is approximated well by

$$b_1 d^{-2-a} \tag{3}$$

for some constant b_1 in an appropriate range of degrees.

2) The number $X(d_1, d_2)$ of edges between vertices of degrees d_1, d_2 is approximated well by the function

$$b_2 (d_1 + d_2)^{1-a} d_1^{-2} d_2^{-2} \tag{4}$$

for some constant b_2 in an appropriate range of degrees.

For the web host graph (undirected, without loops or multiple edges, see Section 3.2 for details) we define $\#_{\mathrm{Host}}(d)$ and $X_{\mathrm{Host}}(d_1, d_2)$ in exactly the same manner as for a random graph H. Each of these two functions can be considered as an empiric density function of some distribution. Indeed, let ξ be the degree of a random vertex and ψ be the ordered pair of degrees of vertices adjacent to a random edge (here the order of the vertices is also chosen randomly). Then the function $\#_{\mathrm{Host}}$ is the empirical density function of the random quantity ξ and X_{Host} is that of the random vector ψ.

It is known that as d grows, the variation of the function $\#_{\mathrm{Host}}(d)$ may dominate its mean [17], see figures in [12]. The same might by true for the function $X_{\mathrm{Host}}(d_1, d_2)$ as d_1 and d_2 grow. Therefore it is more convenient, in particular less vulnerable to fluctuations of the data, to study the corresponding distribution functions instead of the density functions.

To that end, we consider the following *cumulative* functions:

$$\begin{aligned}
\widetilde{\#}_{\mathrm{Host}}(d) &= \sum_{j>d} \#_{\mathrm{Host}}(j), \\
\widetilde{X}_{\mathrm{Host}}(d_1, d_2) &= \sum_{j_1 \geqslant j_2, j_1 > d_{\max}, j_2 > d_{\min}} X_{\mathrm{Host}}(j_1, j_2), \\
\widetilde{\rho}_{\mathrm{Host}}(d_1, d_2) &= \frac{\widetilde{X}_{\mathrm{Host}}(d_1, d_2)}{\widetilde{\#}_{\mathrm{Host}}(d_1) \widetilde{\#}_{\mathrm{Host}}(d_2)},
\end{aligned} \tag{5}$$

where $d_{\min} = \min\{d_1, d_2\}$, $d_{\max} = \max\{d_1, d_2\}$.

The main assumption that we make in our experiments is the following: we assume that the web host graph is obtained using a Buckley–Osthus graph model, such as the graph H described above. Under this assumption, one can show that the cumulative characteristics of the web host graph that we just defined have the following properties.

1) The function $\widetilde{\#}_{\mathrm{Host}}(d)$ is approximated well by

$$f_{a_1, b_1}(d) = b_1 d^{-1-a_1} \tag{6}$$

for some constants a_1, b_1 in an appropriate range of degrees. Note that the exponent in the power law is reduced by 1 after the integration.

2) The function $\widetilde{\rho}_{\mathrm{Host}}(d_1, d_2)$ is approximated well by

$$g_{a_2,b_2}(d) = b_2(d_1 + d_2)^{1-a_2} d_1^{a_2} d_2^{a_2} \tag{7}$$

for some constants a_2, b_2 in an appropriate range of degrees. Note that the approximating function does not change after the integration (see Appendix).

Recall that under our assumption, we actually have $a_1 = a_2$. However, it is worth mentioning that just the fact that some graph satisfies both properties 1) and 2), does not automatically imply the equality $a_1 = a_2$. We explain it in detail in Section 4.

In our experiments, we estimate the constants b_1, b_2, a_1, a_2, with the following results.

1) The values of the approximating functions $f_{a_1,b_1}(d)$ and $g_{a_2,b_2}(d_1, d_2)$ are close enough to $\widetilde{\#}_{\mathrm{Host}}(d)$ and $\widetilde{\rho}_{\mathrm{Host}}(d_1, d_2)$, respectively, for a sufficiently large range of degrees d, d_1, d_2.

2) The estimated values of a_1 and a_2 are very close to each other, with relative difference only about 0.5%.

These facts make us believe that our main assumption about the realization of the Buckley–Osthus random graph with respect to the two quantitative aspects is reasonable. The results are described in detail in Section 3.4.

We describe and justify our method for estimation of the parameters a_1, a_2 in Section 3.3. The experiments with simulated graphs confirm the validity of this method (Section 4).

3.2 Data

All experiments are performed with the web host graph crawled in November 2011 by the major Russian search engine `yandex.ru`. The robot is constantly crawling the web, collecting and updating web pages and links between them. From this data, cleaned from spam and duplicates, a web host graph can be constructed in the following way. Vertices of this graph correspond to owners. *An owner* roughly corresponds to all pages downloaded by the robot at least once that belong to the same second level domain. (In some cases a second level domain is subdivided into several owners. Sometimes different second level domains are merged into a single owner.) An edge between two vertices-owners is drawn if there is a link from a page of one owner to a page of another owner. For the purposes of our work, we further simplify the graph, making it undirected and removing duplicate edges and self-loops. The web host graph constructed in this manner consists of 86.8 million vertices and 1.33 billion edges[1]. We do not suspect any bias in the way this data is collected that may substantially affect our results.

3.3 Framework for Parameter Estimation

In Section 3.1, we already explained that the functions f and g defined by Equations (6) and (7) approximate $\widetilde{\#}_{\mathrm{Host}}$ and $\widetilde{\rho}_{\mathrm{Host}}$ for some appropriate values of the parameters a_1, b_1 and a_2, b_2, respectively. In this section we describe the method we use to optimize these parameters in order to obtain the best possible approximations.

Let $\Delta = \{[\alpha^k] : k \in \mathbb{N}\}$, where $\alpha = 1.01$, and $[\cdot]$ denotes the integer part of a number.

[1]To obtain the graph, please see `http://events.yandex.ru/events/publications/` or contact the authors.

We use non-linear least squares method to minimize the overall deviation between empirical and theoretical functions over points from Δ. Most authors fit the power law distribution to empirical data using a plain linear regression in log-log scale. Problems with this approach and reasons why it is not appropriate for fitting the degree distribution of a real graph have already been discussed extensively [17]. In our case, we see the following additional reasons not to use this method.

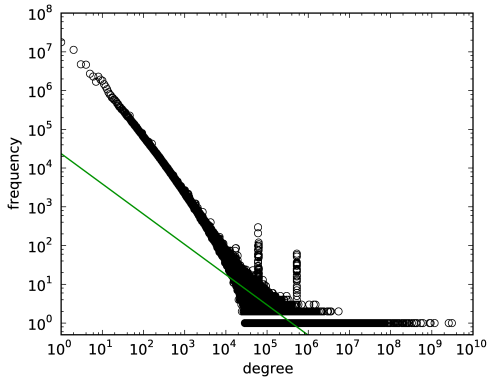
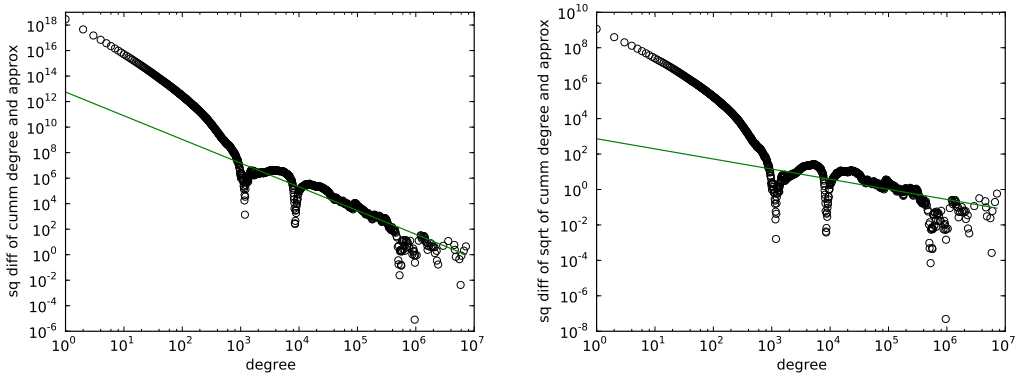


Figure 1: Degree distribution of the web host graph (for each degree, the number of vertices having this degree is represented with a black circle) and approximation using linear regression (green) in logarithmic scale.



(a) (b)

Figure 2: (a) Squared deviation between cumulative degree distribution and approximation using our method. (b) Squared deviation between *square roots* of cumulative degree distribution and approximation using our method. In both cases, linear regression for the range $[10^{2.9}, 10^{5.9}]$ is shown for convenience.

1) Empirical argument: Fig. 1 illustrates that the linear regression for $\log(\#_{\mathrm{Host}}(d))$ is a pretty bad approximation.

2) It can be shown that (assuming the Buckley–Osthus model) the variance and the mean of $\widetilde{\#}_{\mathrm{Host}}(d)$ have the same order of growth as d grows, and therefore the variance of $\sqrt{\widetilde{\#}_{\mathrm{Host}}(d)}$

can be bounded by a constant. This means that the following objective function

$$\frac{1}{|D_1|}\sum_{d\in D_1}\left(\sqrt{\widetilde{\#}_{\mathrm{Host}}(d)}-\sqrt{f_{a_1,b_1}(d)}\right)^2, \tag{8}$$

where $D_1 \subset \Delta$ is a certain range of degrees, is much more appropriate for optimization, as in this case contributions of different summands are better calibrated. To illustrate the validity of this argument empirically, we plot the distribution of $\widetilde{\#}_{\mathrm{Host}}(d) - f_{a_1,b_1}(d)$ (Fig. 2a) and $\sqrt{\widetilde{\#}_{\mathrm{Host}}(d)} - \sqrt{f_{a_1,b_1}(d)}$ (Fig. 2b) for parameters of a_1, b_1 estimated in Section 3.4.

3) The linear regression is not applicable to estimation of parameters a_2, b_2, for function g can not be represented in a linear form. Moreover, we are again able to show that the variance and the mean of the empirical probability of an edge are of the same order.

In accordance with 3), we estimate parameters a_2, b_2 minimizing the following objective function

$$\frac{1}{|D_2|}\sum_{(i,j)\in D_2}\left(\sqrt{\widetilde{\rho}_{\mathrm{Host}}(i,j)}-\sqrt{g_{a_2,b_2}(i,j)}\right)^2, \tag{9}$$

where $D_2 = \{(d_1, d_2) \in D_1^2 : d_1/d_2 > 10\}$, and D_1 is the degree range chosen for estimating a_1, b_1 (to be determined later). Note that we introduce the restriction on d_1/d_2 in accordance with Theorems 3 and 4 that give a good estimation for $\widetilde{\rho}$ only for sufficiently large d_1/d_2 (see Section 2.2.2). The value $C = 10$ was chosen manually.

We minimize the objective functions (8) and (9) using the Gauss–Newton algorithm for a non-linear least squares optimization problem (e.g., see [27]). Varying the degree range D_1, we examined the product of the resulting optimized objectives (8), (9) and chose $D_1 = [10^{2.9}, 10^{5,9}]$ as the range of length 3 (in the logarithmic scale) with the minimal value of the product. The choice of the degree range for our approximations is further justified empirically in our observations on the deviations (Fig. 2). For ranges of larger lengths, the optimized product of objectives starts to grow substantially.

In the next subsection we describe the results of our experiments.

3.4 Estimation for Empirical Cumulative Degree and Edge Distributions

In this section we discuss the results of the two estimation methods for the parameter a described in Section 3.3.

Table 1 shows the estimate of a_2 we obtained deriving the best fit of g_{a_2,b_2} to the empiric conditional probability $\widetilde{\rho}_{\mathrm{Host}}(d_1, d_2)$ that a pair of vertices v_1, v_2 forms an edge in the web host graph given that $\max(\deg(v_1), \deg(v_2)) > \max(d_1, d_2)$ and $\min(\deg(v_1), \deg(v_2)) > \min(d_1, d_2)$ (see Equation (5) for the definition).

We measure the accuracy of the estimation of a_2 employing bootstrapping in the following way. We sample the set of edges of the same size as originally, choosing each edge uniformly at random from the collection of all edges, with replacement. For each sample, we substitute the empirical function $X_{\mathrm{Host}}(d_1, d_2)$ with that for the sampled set, refresh $\widetilde{\rho}_{\mathrm{Host}}(d_1, d_2)$ according to (5) and apply the estimating method described in Section 3.3. Applying the described procedure 1000 times independently, we obtain one estimate for a_2 for each edge sampling. The normalized sum of

parameter	degree distribution	edge distribution
a	$a_1 = 0.2762$	$a_2 = 0.2774$
σ	2.631	0.0599
σ_s	0.005666	$8.518 \cdot 10^{-6}$

Table 1: Results of the approximation of the cumulative distribution for degrees from the interval $D_1 = [10^{2.9}, 10^{5.9}]$ and for edges between vertices with degrees from the domain D_2 for the web host graph (see Sections 3.1 and 3.3 for details).

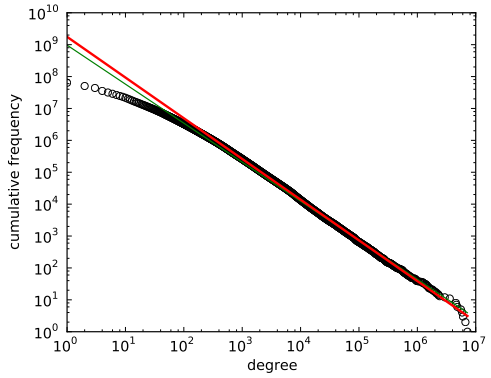


Figure 3: Cumulative degree distribution (black) and approximation using our method (red) in logarithmic scale. For comparison, the result of approximation using linear regression in log-log coordinates is shown (green).

squared deviations between these 1000 estimates and the one obtained from the initial dataset is shown in Table 1 as σ_s^2. We denote the normalized sum of squared deviations between $g_{a_2,b_2}(d_1, d_2)$ and $\widetilde{\rho}_{\text{Host}}(d_1, d_2)$ in the domain D_2 by σ^2.

For the chosen range of degrees $D_1 = [10^{2.9}, 10^{5.9}]$, we obtain $a_2 = 0.2774$ and $b_2 = 8.331 \cdot 10^{-4}$. The results of this approximation are shown on Fig. 4. We observe a very good fit of approximation with the data. Note that due to the term $(d_1 + d_2)^{1-a}$ predicted theoretically, the approximation was even able to capture a concave area around the diagonal $d_1 = d_2$. This would not be possible with a simpler approximation of the form $d_1^a d_2^a$.

The result of estimation of a_1 that we obtained approximating $\widetilde{\#}_{\text{Host}}$ by the function f_{a_1,b_1} in the range of degrees D_1 is also shown in Table 1. We also measure the estimation accuracy using bootstrapping, sampling with replacement 1000 sets of vertices, applying our method to the corresponding degree distribution and obtaining 1000 values of estimates. The normalized sum of squared deviations between $f_{a_1,b_1}(d_1, d_2)$ and $\widetilde{\#}_{\text{Host}}(d_1, d_2)$ is denoted by σ^2 as well.

We want to stress that surprisingly we obtained the same value $a \approx 0.27$ of the parameter approximating independently the degree and the edge distributions. This is a double evidence that the Buckley–Osthus model is good for the web host graph. In the next section, we further support this claim, comparing it with other models.

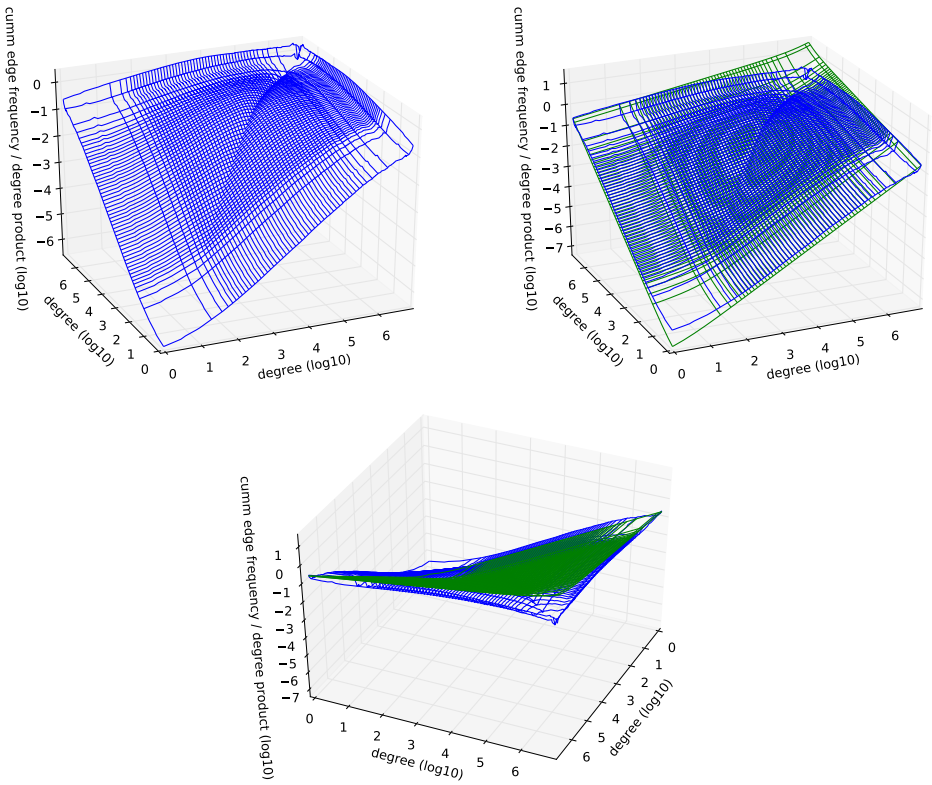


Figure 4: Cumulative edge distribution (blue) and approximation using our method (green) in logarithmic scale (axes are labeled with $\log_{10}$ of the values, pictures differ only in the view angle).

4 Experiments on Simulated Graphs

Here we describe the results of our experiments with graphs artificially generated in various random graph models. We have two goals: to demonstrate that for a random graph with the power law degree distribution the probability of an edge between vertices of given degrees is not determined by the exponent in the power law, and to show that the Buckley–Osthus model has the best approximation to the web host graph as compared with other models.

First of all, we generate ten samples of the Buckley–Osthus (BO) random graphs with 86.8M vertices with $a = 0.276$ and $m = 12$ (close to the ratio of the number of edges and the number of vertices observed in the actual web host graph). The cumulative degree and edge distributions of one resulting graph are shown on Fig. 7 and 5, respectively, in comparison with those for the web host graph. In both cases, we observe a strong fit, recapitulating the results from Section 3.4 (compare with Fig. 4).

Fig. 6 compares the function d_{nn}, average degree of a neighbor, for the web host graph and a sample generated in the Buckley–Osthus model with $a = 0.276$ that corresponds to the best approximation by the model. As expected, the two distributions are very close to each other. Interestingly, even fluctuations of the two are very similar.

In addition to the Buckley–Osthus model, we consider two other random graph models: the configuration model (GDS) and the Holme–Kim model (HK).

The first model chooses from all graphs with a specified fixed degree sequence uniformly at random [5]. For our experiment, we generate a sequence of 86.8M numbers following the power law distribution with the exponent -2.276 and use this distribution as a degree sequence in the model. Then we generate five samples of random graphs in this model with 86.8M vertices and 128M edges using a simple simulation algorithm [5]. The degree distribution of the resulting graph follows the power law by construction.

The second model is based on the idea of preferential attachment with triad formation steps in the graph construction process [28]. We generate nine samples of random graphs with 86.8M vertices and 1B edges. Degree distribution of the resulting graph follows the power law with the exponent -3.

The degree and the edge distributions for a single sample from both models in comparison with those for the web host graph are shown on Fig. 7 and 8, respectively.

For each of the simulated graphs, we apply exactly the same two approximation procedures as described in Section 3.3 and previously applied to the web host graph. Table 2 shows the results: v and e are the number of vertices and edges in the sample graphs, a_1 and a_2 are the parameters of the best fit for degree and edge distributions, respectively. Note that the algorithm diverges for edge distribution approximation of the HK model, and the value of a_2 is not defined in this case. We also show the standard deviation of the obtained estimates of a_1 and a_2 over the several samples of the model. The GDS model has a fixed degree distribution that results in always the same estimate of a_1.

Not surprisingly, the approximation algorithm extracts the parameters a_1 and a_2 planted in the sample of the BO model with high accuracy, as it is the underlying assumption of this algorithm that the graph is modeled by the Buckley–Osthus model.

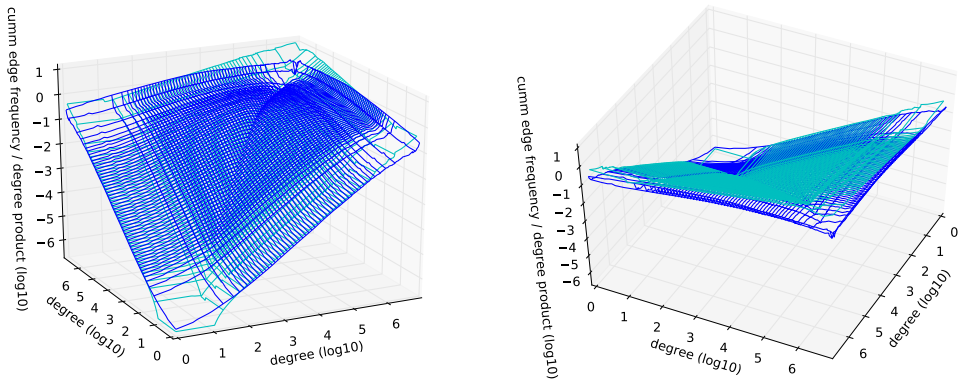


Figure 5: Cumulative edge distributions for the web host graph (blue) and for the Buckley–Osthus simulated graph (cyan) in logarithmic scale (axes are labeled with $\log_{10}$ of the values, pictures differ only in the view angle).

Although all generated graphs have the power law degree distribution, only the Buckley–Osthus graph has the edge distribution close to that observed in the real web host graph.

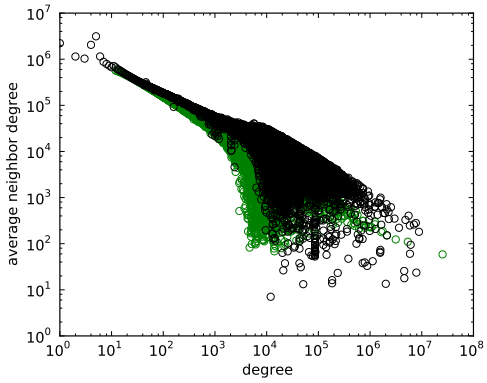


Figure 6: Average degree of a neighbor of a vertex depending on the degree of this vertex for the real web host graph (black) and a sample generated by the Buckley–Osthus model (green) with $a = 0.276$ (corresponding to the best approximation).

model	parameters	estimates
BO	$v = 8.68 \cdot 10^7$, $e = 1.04 \cdot 10^9$	$a_1 = 0.289 \pm 0.0033$, $a_2 = 0.274 \pm 0.0038$
GDS	$v = 8.68 \cdot 10^7$, $e = 1.26 \cdot 10^8$	$a_1 = 0.29 \pm 0$, $a_2 = 1.053 \pm 0.00048$
HK	$v = 8.68 \cdot 10^7$, $e = 1.04 \cdot 10^9$	$a_1 = 1.06 \pm 0.0088$, $a_2 = \text{n/a}$

Table 2: Results of the approximation of the cumulative distributions of degrees from the interval $D_1 = [10^{2.9}, 10^{5.9}]$ and edges between vertices with degrees from the interval D_1 for generated graphs (see Sections 3.1 and 3.3 for details). Number of vertices and edges in graphs are shown as v and e, respectively. Results of the approximation using the method described in Section 3.3, are shown as a_1 and a_2.

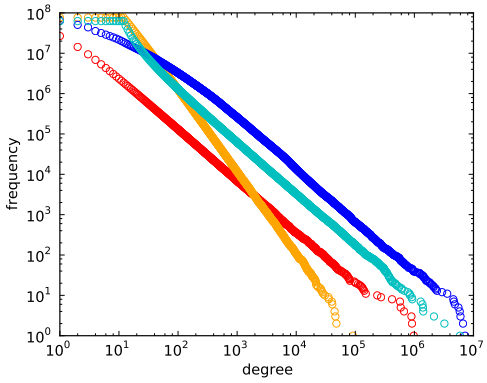


Figure 7: Cumulative degree distributions for the web host graph (blue), the BO simulated graph (cyan), the GDS simulated graph (red), and the HK simulated graph (orange) in logarithmic scale.

5 Conclusion

In this paper we study the degree and edge distributions of the web host graph. We compare it with the Buckley–Osthus model of random graphs and find that the model agrees with the real

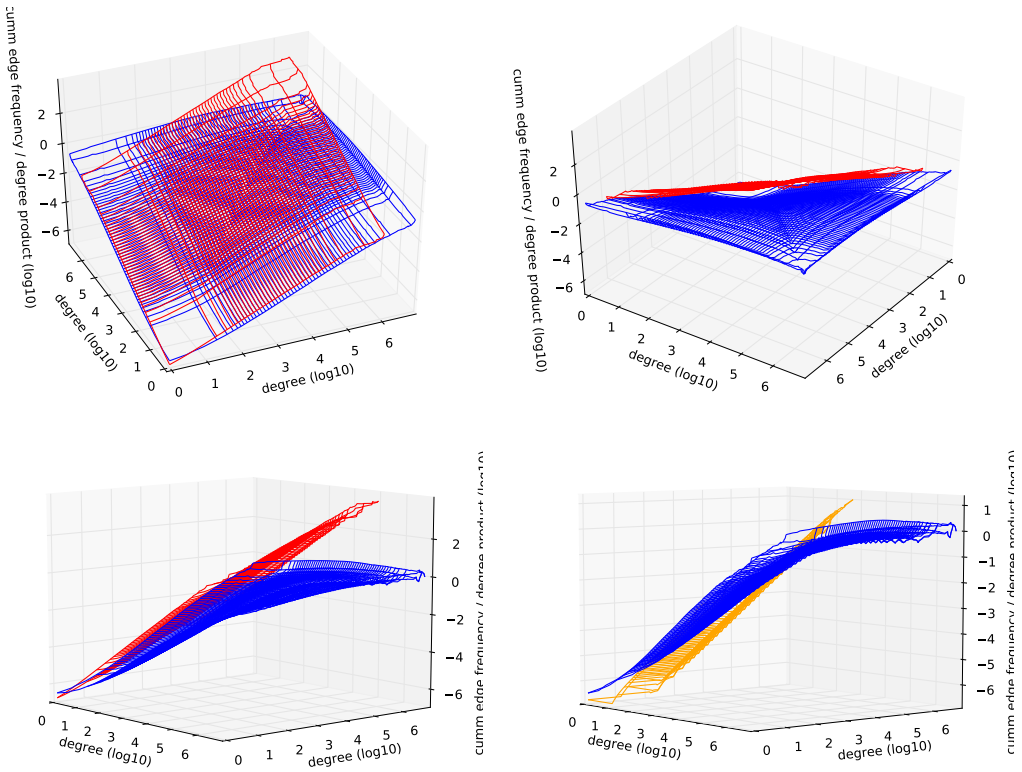


Figure 8: Cumulative edge distributions for the web host graph (blue), the GDS simulated graph (red), and the HK simulated graph (orange) in logarithmic scale. Pictures for the GDS model differ only in the view angle.

data. More precisely, we use two different approaches to estimate the initial attractiveness parameter a assuming the web host graph is generated in the Buckley–Osthus model. In two different independent attempts, we compare the distribution of the number of edges between vertices with respect to their degrees and the degree distribution in the real graph with theoretical predictions for the Buckley–Osthus model. The values of a obtained with two methods are very close to each other, and therefore we conclude that the web host graph is very similar to the Buckley–Osthus random graph with this particular value of a.

Besides our results being interesting on their own, we believe they may potentially be related with real world problems of practical interest.

One example of such a relation may be the work of Y. Lu et al. [39] that made use of the power law degree distribution in the webgraph and proposed the algorithm PowerRank, an improvement over PageRank. We may expect that further empirical and theoretical studies of graphs representing the Internet may help progress in other tasks related with search and in particular with ranking and crawling.

It has been argued that the web contains many communities, sets of pages or hosts that are in particular characterized by abnormally high density of links between them [11, 25, 29]. In this respect, understanding how edges are distributed in the graph may potentially be useful for algorithms detecting and testing such communities, providing a better description of expected

background that prospective communities may be compared against. We expect that theoretical and empirical results in the direction presented in this paper may prove useful for these problems.

One can imagine a lot of directions for future work related with our results, both theoretical and practical.

It would be interesting to continue to study the Buckley–Osthus random graph model, as well as other models, and extend theoretical knowledge of their properties. For the first time we described the distribution of edges between vertices given their degrees in a real Internet graph. Now it is interesting to compare different models with respect to this property, and our techniques may be useful.

Even though we showed a good correlation of the model with real data, we had to simplify the data in certain important aspects. It would be interesting to generalize existing random graph models or probably to develop new ones that could model graphs closer to the reality: with multiple edges, directed, hierarchical, dynamically evolving with time. In particular, the clustering coefficient of a Buckley–Osthus graph still significantly differs from the one in the reality. However, some of the aspects of the Buckley–Osthus model may be promising.

It would definitely be interesting to develop and test the aforementioned and similar ideas of applications to ranking, crawling, and community detection. We strongly believe that deeper and broader theoretical results on models of Internet graphs coupled with empirical observations of certain characteristics of real such graphs may lead to practical applications and insights.

References

[1] M. Abramowitz, I. A. Stegun (editors), *Handbook of mathematical functions with formulas, graphs and mathematical tables*, Dover, tenth GPO printing, 1964.

[2] R. Albert, A.-L. Barabási, Topology of Evolving Networks: Local Events and Universality, Phys. Rev. Lett., vol. 85, pp. 5234–5237, 2000.

[3] R. Albert, H. Jeong, A.-L. Barabási, *The Diameter of the WWW*, Nature 401. 130, 1999.

[4] A.-L. Barabási, R. Albert, *Emergence of scaling in random networks*, Science, 286, 509–512, 1999.

[5] E.A. Bender, E.R. Canfield, *The asymptotic number of labeled graphs with given degree sequences*, J. Combin. Theory A 24, 296–307, 1978.

[6] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, *Complex networks: Structure and dynamics*, Physics Reports 424, 175–308, 2006.

[7] B. Bollobás, Ch. Borgs, J. Chayes, O. Riordan, *Directed scale-free graphs*, Proc. SODA'03, 2003.

[8] B. Bollobás, O. M. Riordan. *Mathematical results on scale-free random graphs*, Handbook of graphs and networks, Wiley-VCH, Weinheim, 2003.

[9] B. Bollobás, O. M. Riordan, J. Spencer, G. Tusnády. *The degree sequence of a scale-free random graph process*, Random Structures and Algorithms, 18:3, 279–290, 2001.

[10] B. Bollobás, O. M. Riordan, *The diameter of a scale-free random graph*, Combinatorica, 24:1, 5–34, 2004.

[11] A. Bonato, *A Survey of Models of the Web Graph*, In: A. López–Ortiz and A. Hamel (Eds.): CAAN 2004, LNCS 3405, pp. 159–172, 2005.

[12] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, *Graph structure in the web*, Proc. 9th WWW, 2000.

[13] A. Broder, R. Lempel, F. Maghoul, and J. Pedersen, *Efficient pagerank approximation via graph aggregation*, Proc. 13th WWW conf., pages 484–485, 2004.

[14] P. G. Buckley, D. Osthus, *Popularity based random graph models leading to a scale-free degree sequence*, Discrete Math. 282 (2004), 53–68.

[15] M. Catanzaro G. Caldarelli, L. Pietronero, *Assortative model for social networks*, Phys. Rev., E 70, 037101. 2004.

[16] F. Chung, L. Lu, *The average distances in random graphs with given expected degrees*, Proc. Natl. Acad. Sci. USA 99, 15879–15882, 2002.

[17] A. Clauset, C. R. Shalizi, M. E. J. Newman. *Power-Law Distributions in Empirical Data*, SIAM review, 51(4), pp. 661–703, 2009.

[18] C. Cooper, A. Frieze. *A general model of web graphs*, J. Random Structures and Algorithms 22(3), 2003.

[19] L. da F. Costa, F. A. Rodrigues, G. Travies, P. R. Villas Boas *Characterization of complex networks: A survey of measurements*, Advances in Physics, 56(1), 167–242, 2007.

[20] S. N. Dorogovtsev, J. F. F. Mendes, A. N. Samukhin. *Structure of growing networks with preferential linking*, Phys. Rev. Lett. 85 (2000), 4633.

[21] E. Drinea, M. Enachescu, M. Mitzenmacher, *Variations on random graph models for the web*, technical report, Harvard University, Department of Computer Science, 2001.

[22] D. Easley, J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[23] K. Efe, V. Raghavan, C. H. Chu, A. L. Broadwater, L. Bolelli, S. Ertekin, *The Shape of the Web and Its Implications for Searching the Web*, Proc. International Conference on the Advances in Infrastructure for Electronic Business, Science, and Education on the Internet, 2000.

[24] M. Faloutsos, P. Faloutsos, Ch. Faloutsos, *On power-law relationships of the Internet topology*, Proc. SIGCOMM'99, 1999.

[25] D. Gibson, R. Kumar, A. Tomkins, *Discovering Large Dense Subgraphs in Massive Graphs*, Proc. 31th VLDB, pp. 721–732, 2005.

[26] E.A. Grechnikov, *The degree distribution and the number of edges between vertices of given degrees in the Buckley-Osthus model of a random web graph*, Journal of Internet Mathematics, accepted for publication. arXiv:1108.4054v1, 2011.

[27] Greene, William H. *Econometric Analysis*, Prentice Hall, 4th ed, 1999.

[28] P. Holme, B.J. Kim, *Growing scale-free networks with tunable clustering*, Phys. Rev. E, vol. 65(2), 026107, 2002.

[29] J. Kleinberg, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. *The web as a graph: Measurements, models and methods*, In: Proc. ICCC, LNCS 1627, 1999.

[30] A. Kogias, D. Anagnostopoulos, *A simulation-based evaluation of an exponential growth copying model for the web-graph*, Proc. 12th WWW, 2003.

[31] P. L. Krapivsky and S. Redner, *Organization of Growing Random Networks*, Phys. Rev. E 63, 066123, 2001.

[32] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, *Stochastic models for the Web graph*, Proc. FOCS'00, 2000.

[33] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, *Web as a graph*, Proc. PODS 2000, pp. 1-10, 2000.

[34] S. Maslov, K. Sneppen, *Specificity and stability in topology of protein networks*, Science, 296, 5569, pp. 910–913, 2002.

[35] S. Maslov, K. Sneppen, A. Zaliznyak, *Detection of Topological Patterns in Complex Networks: Correlation Profile of the Internet*, Physica A 333, pp. 529–540, 2004.

[36] M.E.J. Newman, *Assortative mixing in networks*, Phys. Rev. Lett., vol. 89(20), 208701, 2002.

[37] J. Park, M.E.J. Newman, *The origin of degree correlations in the Internet and other networks*, Phys. Rev. E, vol. 68, 026112 (2003)

[38] R. Pastor-Satorras, A. Vázquez, A. Vespignani, *Dynamical and Correlation Properties of the Internet*, Phys. Rev. Lett., vol. 87(25), 258701, 2001.

[39] Y. Lu, B. Zhang, W. Xi, Z. Chen, Y. Liu, M. R. Lyu, W.-Y. Ma, *The PowerRank Web Link Analysis Algorithm*, Proc. 13th WWW, 2004.

A Proof of Proposition 1

We can estimate the expectation of the number of loops in the following way:

$$EN(\text{loops in } H^n_{a,m}) = O\left(\sum_{i=1}^{n} \frac{1}{i}\right) = O\left(\ln n\right).$$

To estimate the number of multiple edges we should take into account that we have no vertices of degrees greater than $2mn$ in $H^n_{a,m}$. Also (using the same ideas as in the proof of Theorem 3) it can be shown that $E\#_a(d,i) = O\left(\frac{i}{d^{2+a}}\right)$. Therefore

$$EN(\text{multiple edges in } H^n_{a,m}) =$$

$$= O\left(\sum_{i=1}^{n}\sum_{d=1}^{2mi} E\#_a(d,n)\left(\frac{d-1+a}{(a+1)i}\right)^2\right) =$$

$$= O\left(\sum_{i=1}^{n}\sum_{d=1}^{2mi} \frac{i}{d^{2+a}}\frac{d^2}{i^2}\right) = O\left(n^{1-a}\right).$$

B Proof of the theoretical approximation in Equation (7)

Here we prove the theoretical approximation from Equation (7) for the empiric conditional probability $\widetilde{\rho}_{\text{Host}}(d_1,d_2)$.

First, for sufficiently large d_1/d_2, we obtain the following approximate formula using the estimations (3) and (4):

$$\widetilde{\rho}_{\text{Host}}(d_1,d_2) \approx \frac{b_2\sum_{i\geqslant j,\, i>d_1, j>d_2}(i+j)^{1-a_2}(ij)^{-2}}{b_1^2\sum_{i>d_1} i^{-2-a_2}\sum_{j>d_2} j^{-2-a_2}}. \tag{10}$$

For d_1/d_2 large enough, the numerator of the right-hand side of (10) equals

$$\sum_{i>d_1\geqslant j>d_2} b_2(i+j)^{1-a_2}(ij)^{-2} + \sum_{i\geqslant j>d_1} b_2(i+j)^{1-a_2}(ij)^{-2} \approx$$

$$c_1(d_1+d_2)^{1-a}(d_1d_2)^{-1} + c_2(d_1)^{-1-a} \approx c_1(d_1+d_2)^{1-a}(d_1d_2)^{-1}$$

for some constants c_1, c_2. Estimating the denominator of the right-hand side of (10) by $c(d_1d_2)^{-1-a_2}$, we get $\widetilde{\rho}_{\text{Host}}(d_1,d_2) \approx g_{a_2,b_2}(d_1,d_2)$, where g_{a_2,b_2} is defined by (7).