

On the Entropy of Models for the Web Graph

Martin Sauerhoff

Krengelstrasse 9, 44869 Bochum, Germany

`martin.sauerhoff@udo.edu`

Abstract. It is a well-known fact that the adjacency relation of the web graph is highly compressible, which is exploited in coding schemes for storing the graph in practice. We analytically investigate how three exemplary stochastic graph models explain this important phenomenon by giving precise estimates of the average entropy per vertex for them. We consider a preferential attachment model, a copying model, and a hierarchical model. While for all three models the average entropy per vertex is asymptotically logarithmic in the number of vertices, the constant factor in these estimates allows to get a more detailed picture. The preferential attachment model turns out to have a factor of 1, thus allowing essentially no compression. The copying model and the hierarchical model have constant factors proportional to the fraction of copied successors per vertex and to the fraction of links with endpoints chosen randomly from the whole graph, resp. Thus, the latter models allow to explain the observed compressibility of web graphs to some extent, depending on the choice of parameters. For the hierarchical model without global links, we even get constant average entropy.

1. Introduction

The web graph can be formally described at any given instant by a directed graph with the web pages (more precisely, URLs) as vertices and hyperlinks as edges. Understanding the statistical properties of this graph is highly practically relevant for carrying out efficient algorithms on it, e. g., algorithms mining data from the adjacency relation like PageRank [10]. Given the enormous size of the web graph (in the order of 10^{12} vertices for the indexable web in 2005 [17]), one tries to understand the structure of the web graph by experiments on snapshots on the one hand and stochastic models on the other.

For detailed overviews over existing literature dealing with stochastic models for the web graph and techniques for analyzing them we refer to [12, 23]. We only briefly mention some major steps of the development that are relevant for what follows and describe the models investigated here in more detail later on. All models we deal with are dynamic and pure-birth, i. e., the random graphs according to the models evolve over a series of discrete time steps where in each step, new vertices and edges are added.

One crucial feature of the web graph that has been observed in various experiments (see, e. g., [11, 14, 15]) is that the indegree of its vertices are approximately distributed according to a

power law, i. e., over wide ranges for d , the fraction of vertices with indegree d is proportional to $1/d^\alpha$ for some specific constant α . For sufficiently large d , this also seems to be true for the outdegree. The two known mechanisms that plausibly recreate this feature in a dynamic model of the web graph are *preferential attachment*, first described in this context by Barabási and Albert [3], and *copying*, introduced by Kumar et al. [20]. In the preferential attachment model of Barabási and Albert, the destinations of the edges leaving a new vertex are chosen with probability linear in the indegree of the destinations. This leads to a power law distribution of the indegree with a fixed exponent of 3, as shown rigorously by Bollobás et al. [9]. Variants that allow to get power laws with arbitrary exponents for in- and outdegree have been described by Aiello et al. [2]. In a copying model, a fraction of the outgoing edges are copied from a randomly chosen prototype vertex, while the rest is chosen uniformly at random. Kumar et al. [20] have shown that this also leads to a power law for the indegree, where the exponent depends on the fraction of copied successors.

The described models capture mechanisms for generating the observed degree distributions and also have some other desirable properties like small diameter (shown for the model of Barabási and Albert in [8]) or a large number of bipartite cliques (shown for the copying models in [20]). Nevertheless, they still seem to disagree quantitatively with experimental findings, e. g., with respect to clustering features and the sizes of connected components [19]. Furthermore, it is unclear whether they are suitable to explain the “self-similarity” of the real web graph observed in experiments, e. g., in [14]. As a remedy, *hierarchical models* try to capture the way web pages are organized on web sites in the real world. Such models have been proposed, e. g., by Laura et al. [21], Ravasz and Barabási [26], Eiron and McCurley [16], and Han et al. [18]. The experiments reported in these papers seem to indicate that the respective models indeed tend to be better than previous models at combining the diverging goals of having power laws for the degree distributions with the right exponents and also clustering features of the real graph like many bipartite cliques.

Given the huge size of the web graph, it is of obvious practical importance to be able to store it in compressed form. It has been experimentally observed that the adjacency relation of snapshots of the web graph indeed allows this, i. e., the respective information can be encoded with considerably less bits than the worst-case $\log n$ bits per edge destination in an n vertex graph [1, 4, 6, 7, 25, 28]. For example, the compression algorithms in the WebGraph package, achieving the best published compression ratios so far, manage to store graphs with about 10^9 nodes using typically 2–3 bits per edge and thus achieve compression ratios of more than 90 % over the trivial encoding (see [6] and the homepage of the authors). The most important empirical properties of the web graph exploited for this compression are the *locality* of links, i. e., many links have their source and destination on the same web site or both are even close to each other in the directory hierarchy of the web site, and the *similarity* of web pages, meaning that pages close to each other share a lot of common links. Furthermore, several of the experimental papers report that using a Huffman code for the link destinations based on the indegree distribution can be used for compression. On the other hand, a back-of-the-envelope calculation by Adler and Mitzenmacher [1] shows that such a scheme uses an expected number of $\Theta(m \log n)$ bits for a graph with n vertices, $m = \Theta(n)$ edges, and a power law distributions for the indegree with exponent larger than 2. Hence, the improvement achievable by Huffman encoding over the trivial encoding can only be in the constant factor hidden in the asymptotic estimate.

Eiron and McCurley [16] have also investigated to which extent stochastic models of the web graph allow compression. For this, they have experimentally determined the average entropy of the adjacency lists over all vertices for snapshots of three stochastic models of the web graph after a fixed number of iterations (they call this *isolated destination entropy*). These models are preferential attachment as described by Barabási and Albert, a version of the linear growth copying model of Kumar et al., and finally a version of their own hierarchical model. The hierarchical model of Eiron and McCurley explicitly models the directory trees (web sites) containing the web pages and the link structure between these pages as individual random graphs, which comes closer to the real world scenario than other models of the same kind, but in its full-blown form is also quite complicated due to a lot of choosable parameters. The entropy values measured for the three models turn out to decrease in the given order of the models, as may be expected, which can serve as an argument for incorporating copying and hierarchical organization in a realistic model of the web graph.

In order to explain the mentioned experiments theoretically and to better understand the feature of compressibility of the web graph, it is clearly important to also have analytical estimates of the entropy of the usual stochastic models of the web graph. Somewhat surprisingly, it seems that no such analyses have been carried out so far. The only theoretical results about the entropy of stochastic network models we are aware of are from the physics literature. Solé and Valverde [27] have calculated the entropy of degree distributions for some simple classes of random graphs and for some static real-world networks and instances of stochastic models for the web graph, where in the latter two cases the entropy is measured with respect to vertices chosen uniformly at random. Park and Newman [24] and Bianconi [5] have described how stochastic network models can be derived by maximizing the entropy under constraints like fixed degree distributions, using methods from statistical mechanics. These approaches are obviously only remotely related to what we are interested in here.

The rest of the paper is organized as follows. In Section 2, we describe the models considered here in detail and present our entropy estimates for them. In Section 3, we introduce some definitions and general tools from information theory needed later on. Finally, Section 4 contains the proofs of the main theorems.

2. Models and Our Results

Similar to the paper of Eiron and McCurley [16], we consider the preferential attachment model of Barabási and Albert [3] and the linear growth copying model of Kumar et al. [20]. Furthermore, we define and analyze a simplified variant of the hierarchical model of Eiron and McCurley. We keep the important feature of two individual processes creating the directory and the link structure, resp., of the web graph. On the other hand, in order to get a model that is accessible to analytical methods, we reduce the parameters to some essential few.

We suggest to investigate the extent of compressibility of the link structure as an additional criterion for rating the plausibility of models for the web graph. In order to get an objective measure of compressibility, we estimate the entropy of the random graphs created by the models and thus complement the experimental findings of Eiron and McCurley by analytical results.

Throughout the paper, we only consider directed graphs. All models describe a random graph by starting from a fixed graph at the beginning and then modifying this through a series of discrete

time steps in which vertices and edges are added. For a random graph G over n vertices, its *entropy* is defined by $H(G) := \sum_g \Pr\{G = g\} \log(1/\Pr\{G = g\})$, where the sum extends over all (directed) graphs with n vertices.

In what follows, we precisely describe the models considered here along with our entropy estimates for them.

2.1. The Barabási-Albert Model

This model has been introduced by Barabási and Albert [3] and has been made more precise by Bollobás et al. [8,9] whose version of the model we rely on here. By $G_t^{(m)}$ we denote the random graph according to this model obtained after t time steps, where in each step a new vertex with $m \in \mathbb{N}$ outgoing edges with random destinations according to the preferential attachment rule is added. In what follows, we consider the case $m = 1$ and apply the results to the general case at the end.

Let $G_t^{(1)} = (V_t, E_t)$ denote the graph after time step $t \in \mathbb{N}$, where $V_t = \{1, \dots, t\}$. The graph is defined by the following process.

- $G_1^{(1)}$ consists of a single vertex with a self-loop.
- $G_t^{(1)}$ for $t \geq 2$ is obtained from $G_{t-1}^{(1)}$ by adding a new vertex t to $G_{t-1}^{(1)}$ with an outgoing edge to $v_t \in \{1, \dots, t\}$ chosen at random according to the distribution

$$\Pr\{v_t = i \mid G_{t-1}^{(1)} = G\} = \frac{d_G(i)}{2t - 1}, \quad 1 \leq i \leq t,$$

where $d_G(i)$ is the (total) degree of vertex i in G and $d_G(t) = 1$. A self-loop yields a contribution of 2 to the degree.

For this model, we derive the following estimate for the average entropy per vertex, which is precise up to less than one bit.

Theorem 1. *For $c := 19/6 - (7/5) \log 3 < 0,948$, we have $\log T - 1/\ln 2 - c - o(1) \leq H(G_T^{(1)})/T \leq \log T - 1/\ln 2 + o(1)$.*

The graph $G_t^{(m)}$ for arbitrary $m \in \mathbb{N}$ is created by starting with a single vertex with m self-loops and in each of $t - 1$ time steps adding a node with m successors. The edges are added one after another with the probability of a destination being the fraction of its degree at this time over the total degree plus one (the latter for the “outgoing half” of the new edge). More conveniently, the same distribution is obtained by taking $G_{mt}^{(1)}$ and identifying each sequence of m consecutive vertices $(i - 1)m + 1, \dots, im$, for $i = 1, \dots, t$. We get the following bounds on the entropy of these graphs.

Theorem 2. *For $c := 19/6 - (7/5) \log 3$, we have $m \log T - m(1/\ln 2 + c + o(1)) \leq H(G_T^{(m)})/T \leq m(\log T - 1/(\ln 2) + o(1))$.*

The trivial binary encoding of a graph with T vertices and outdegree m uses $m \lceil \log T \rceil$ bits per vertex. The model of Barabási and Albert leads to graphs whose entropy is smaller than this trivial bound only by at most an additive constant for constant m , allowing essentially no compression. It is thus clear that preferential attachment in its pure form does not recreate the right distribution of links found in the real web graph.

2.2. The Linear Growth Copying Model

The linear growth copying model has been introduced by Kumar et al. [20]. The model has parameters $\alpha \in (0, 1)$ and $d \in \mathbb{N}$, where α is called the *copy factor* and d is the (fixed) outdegree of the vertices.

The graph G_t after time step $t \in \mathbb{N}$ is defined by the following process.

- G_1 is a single vertex with d self-loops.
- G_t for $t \geq 2$ is obtained from G_{t-1} as follows.
 - Choose a *prototype* vertex $w \in \{1, \dots, t-1\}$ uniformly at random.
 - Add a new vertex $u = t$ to G_{t-1} . For each $i = 1, \dots, d$, add an edge (u, v_i) with v_i defined as follows:
 - With probability α , choose $v_i \in \{1, \dots, t-1\}$ uniformly at random.
 - With probability $1 - \alpha$, let v_i be the i -th successor of w .

For the linear growth copying model, we show:

Theorem 3. For $\alpha \in (0, 1)$ and $d \in \mathbb{N}$,

$$(\alpha d + 1 - c_{\alpha,d}) \log T - O(d) \leq H(G_T)/T \leq (\alpha d + 1 - \alpha^d) \log T + O(d),$$

where

$$c_{\alpha,d} = (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \alpha^2 \frac{1 - \alpha}{1 + \alpha} d \right) \leq 1.$$

If we have non-constant degree growing with T , i. e., $d = d(T) = \omega(1)$ for $T \rightarrow \infty$, this in particular implies that the average entropy per vertex in G_T is of order $\alpha d \log T - o(d \log T)$ and the average entropy per edge is at least $\alpha \log T - o(\log T)$. Compared to the trivial encoding, compression thus allows to save a fraction of bits directly proportional to the copy factor.

According to the results of Kumar et al. [20], the exponent of the power law distribution for the indegrees in the linear growth copying model approaches $(2 - \alpha)/(1 - \alpha)$ for large T . Thus, for achieving the experimentally exponent of about 2.1, we have to set $\alpha \approx 1/11$. This gives a compression ratio which tallies quite well with the best experimentally observed ratios of about 90 % [6]. Nevertheless, it seems that more experimental results on the real web graph are required to get a meaningful upper bound on the entropy as a function of the number nodes that could be compared with the lower bound derived here.

2.3. A Simple Hierarchical Model

We investigate a simplified version of the hierarchical model proposed by Eiron and McCurley [16]. Their model (or rather class of models) has a lot of choosable parameters which we reduce to some essential few needed for insights into how this model may explain the compressibility of the web graph. The parameters of our simplified version are numbers $\alpha_1, \alpha_2, \alpha_3 \in (0, 1)$ and $d \in \mathbb{N}$, the latter being the fixed outdegree of the URLs.

As Eiron and McCurley, we have two individual random graphs at each time step. The first one describes the directory hierarchy found on different web sites and is called *directory graph*. Each vertex represents a directory and is labeled with a list of URLs that the respective directory

contains. Edges represent the containment relation between directories. Second, we have the *URL graph* which describes the link structure of the web. Vertices are URLs and edges represent links between URLs. We now describe how these graphs are generated.

Directory graph: The graph G_t after time step $t \in \mathbb{N}$ is obtained as follows.

- G_1 is a single directory with one URL.
- G_t for $t > 1$ is obtained from G_{t-1} by one of the following actions chosen with the given probabilities.
 - (1) Probability α_1 : Add a new, isolated directory containing one URL.
 - (2) Probability α_2 : Add a new directory containing one URL as a subdirectory of an existing directory, where the parent directory is chosen with probability proportional to its number of URLs.
 - (3) Probability $1 - \alpha_1 - \alpha_2$: Add a URL to an existing directory chosen with probability proportional to its number of URLs.

URL graph: Let U_t be the graph after time step $t \in \mathbb{N}$. Add in- and outlinks in this graph for each new URL as follows:

Inlinks:

- For a URL in a new directory tree, add an inlink from a URL chosen uniformly at random from all existing URLs except the new one.
- For a URL in an existing directory tree, add an inlink from a URL chosen uniformly at random from all existing URLs except the new one in the directory where new URL has been created and its parent directory.

Outlinks: Call the directory tree in which the new URL has been inserted its *insertion tree*. For each $i = 1, \dots, d$ independently create a link from the new URL to an existing URL chosen with probability proportional to the indegree of the latter, from the following sets of eligible URLs:

- With probability α_3 , choose a destination from all URLs outside the insertion tree of the new URL.
- With probability $1 - \alpha_3$, choose a destination from all URLs in the insertion tree of the new URL.

We observe that random URL graph U_t with the link structure we are interested in also contains entropy derived from the underlying, random directory graph G_t . Hence, we investigate the entropy of U_t conditioned on G_t . We show the following upper bound on this entropy:

Theorem 4. $H(U_T | G_T)/T = (\alpha_1 + d\alpha_3) \log T + O(d)$.

It is not hard but tedious to obtain a corresponding lower bound by a similar proof. We refrain from doing this since our main point is to show the savings in compression that this hierarchical model allows compared to the pure preferential attachment and pure copying models. The constant factor of the logarithmic term in the estimate is proportional to the average number of global links of a URL, i. e., links from or to endpoints chosen uniformly at random from the whole web graph. Apart from such global links the average entropy per vertex is constant. Thus, incorporating locality into a web graph model and restricting the use of the (obviously unrealistic) global links makes it easy to achieve realistically large compressibility ratios.

3. Preliminaries

We assume that the reader is familiar with the basics of information theory and refer to [13] for an introduction. We only briefly mention definitions and facts used here. All random variables considered here take values from finite sets without saying so explicitly. Furthermore, all logarithms are base 2.

Let X and Y be random variables taking values in R and S , resp. The *entropy* of X is defined as $H(X) := \sum_{x \in R} \Pr\{X = x\} \log(1/\Pr\{X = x\})$. The *conditional entropy* of X given Y is $H(X|Y) := \sum_{y \in S} H(X|Y = y) \cdot \Pr\{Y = y\}$. For random variables X_1, \dots, X_n , $H(X_1, \dots, X_n)$ denotes the entropy of the joint distribution of X_1, \dots, X_n . We note the following well-known facts (see, e. g., [13]).

Proposition 5.

- (1) For a random variable X taking values in R , $0 \leq H(X) \leq \log |R|$.
- (2) For random variables X, Y , $H(X) \geq H(X|Y) = H(X, Y) - H(Y)$.
- (3) Let X be a random variable taking values in R and let $f: R \rightarrow S$ be a function. Then $H(f(X)) \leq H(X)$. Furthermore, $H(f(X)) = H(X)$ if f is bijective.
- (4) Chain rule of entropy: Let X_1, \dots, X_n be random variables. Then

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1}).$$

As a consequence, $H(X_1, \dots, X_n) \leq H(X_1) + \dots + H(X_n)$ with equality if X_1, \dots, X_n are independent.

- (5) Let X_0, X_1 be random variables with disjoint support and let $B \in \{0, 1\}$ be a random variable with $\Pr\{B = 1\} = \alpha$. This implies that

$$H(X_B) = \alpha H(X_1) + (1 - \alpha)H(X_0) + h_2(\alpha),$$

where $h_2(x) := -(x \log x + (1 - x) \log x)$ for $x \in [0, 1]$ is the binary entropy function.

We also note the following fact about the entropy of rounded integer fractions.

Proposition 6. Let $m \in \mathbb{N}$ and let X be a random variable taking values in a finite subset of the integers. Then $H(\lfloor X/m \rfloor) \geq H(X) - \log m$.

Proof: We observe that $f(x) := (\lfloor x/m \rfloor, x \bmod m)$ defines a bijective mapping from \mathbb{Z} to \mathbb{Z}^2 . Hence,

$$H(X) = H(f(X)) = H(\lfloor X/m \rfloor, X \bmod m) \leq H(\lfloor X/m \rfloor) + \log m,$$

which yields the claim. □

In the next two propositions we collect some estimates of sums in advance that we will need later on in our proofs.

Proposition 7. *Let β be a constant with respect to t . Then*

$$\sum_{z=1}^t \frac{1}{z^\beta} = \begin{cases} O(t^{1-\beta}), & \text{for } 0 < \beta < 1; \\ O(\log t), & \text{for } \beta = 1; \\ O(1), & \text{for } \beta > 1. \end{cases}$$

$$\sum_{z=1}^t \frac{\log z}{z^\beta} = \begin{cases} O(t^{1-\beta} \log t), & \text{for } 0 < \beta < 1; \\ O(\log^2 t), & \text{for } \beta = 1; \\ O(1), & \text{for } \beta > 1. \end{cases}$$

These results follow from standard estimates for the harmonic numbers for the case $\beta = 1$ and by upper bounding the sums by integrals for the remaining cases. (See the Appendix for details.)

Proposition 8.

$$\sum_{z=1}^t \frac{\log z}{(z+1)(z+2)} \leq \frac{25}{12} + \frac{7}{10} \log 3.$$

Proof: By standard arguments from calculus, it follows that the function $x \mapsto (\log x)/(x+1)^2$ is monotonically decreasing for $x \geq 3$. Hence,

$$\begin{aligned} \sum_{z=1}^t \frac{\log z}{(z+1)(z+2)} &\leq \frac{1}{12} + \frac{\log 3}{20} + \sum_{z=4}^t \frac{\log z}{(z+1)^2} \\ &\leq \frac{1}{12} + \frac{\log 3}{20} + \int_3^t \frac{\log z}{(z+1)^2} dz. \end{aligned}$$

Furthermore,

$$\begin{aligned} \int_3^t \frac{\log z}{(z+1)^2} dz &= \left(\frac{z}{z+1} \log z - \log(z+1) \right) \Big|_3^t \\ &= \frac{t}{t+1} \log t - \log(t+1) - \frac{3}{4} \log 3 + 2 < 2 - \frac{3}{4} \log 3. \end{aligned}$$

Substituting this into the previous inequality yields the claim. \square

4. Entropy Estimates

In this section, we prove the main theorems from Section 2.

4.1. The Barabási-Albert Model

For the proof of Theorem 1, we use the asymptotically exact characterization of the number of vertices with fixed degree in the random graphs of the model due to Bollobás et al. [9]. For $1 \leq d \leq t+1$ let $N_{t,d}$ denote the number of vertices in $G_t^{(1)}$ of (total) degree d .

Lemma 9 ([9]). *Let $1 \leq d \leq t^{1/15}$. Then with probability at least $1 - e^{-t/8}$,*

$$|N_{t,d} - \alpha_d \cdot t| \leq \sqrt{t \log t},$$

where

$$\alpha_d := \frac{4}{d(d+1)(d+2)}.$$

Corollary 10. *Let $\varepsilon_t := 2(\log t)^{1/2}/t^{3/10}$. Then with probability at least $1 - e^{-t/16}$, the following is satisfied for all d with $1 \leq d \leq t^{1/15}$:*

$$N_{t,d} \geq \frac{4}{d(d+1)(d+2)}(1 - \varepsilon_t)t.$$

Proof of Theorem 1: The upper bound follows immediately from the fact that G_T is chosen uniformly from $T!$ different values. By Stirling's formula, its entropy is thus

$$\log T! = T \log T - \frac{1}{\ln 2}T + o(T).$$

In what follows, we prove the lower bound. For $t \in \mathbb{N}$ let v_1, \dots, v_t be the random destinations of the edges in G_t in the order of their creation. For estimating the entropy, we may identify G_t with this list of vertices. We prove the theorem for $G_T, T \in \mathbb{N}$.

By the chain rule of entropy, the fact that $H(G_1) = H(v_1) = 0$, and the definition of conditional entropy,

$$H(G_T) = H(v_1, \dots, v_T) = \sum_{t=1}^{T-1} H(v_{t+1} | v_1, \dots, v_t) = \sum_{t=1}^{T-1} H(v_{t+1} | G_t). \quad (1)$$

Hence, it suffices to lower bound the entropy of v_{t+1} for each $t = 1, \dots, T - 1$ conditioned on G_t . We prove the following.

Claim. *Let $c := 25/12 - (7/10) \log 3$. There is a constant β with respect to t , $0 < \beta < 1$, such that for all $t = 1, \dots, T - 1$,*

$$H(v_{t+1} | G_t) \geq (\log(2t + 1) - 2c)(1 - t^{-\beta}).$$

We first show how this claim implies the theorem. By (1) and the claim,

$$H(G_T) \geq \sum_{t=1}^{T-1} (\log(2t + 1) - 2c)(1 - t^{-\beta}).$$

We expand the terms in the sum and estimate the resulting summands separately. First,

$$\begin{aligned} \sum_{t=1}^{T-1} \log(2t + 1) &\geq \sum_{t=1}^{T-1} \log(2t) = \log((T-1)!) + T - 1 \\ &= T \log T + \left(1 - \frac{1}{\ln 2}\right)T - o(T). \end{aligned}$$

Using Proposition 7, it follows that

$$\sum_{t=1}^{T-1} (\log(2t+1) - 2c) \cdot t^{-\beta} = o(T).$$

Together, these results yield

$$H(G_T) \geq T \log T - \left(2c + \frac{1}{\ln 2} - 1\right)T - o(T),$$

where $c = 25/12 - (7/10) \log 3$. This gives the lower bound required for the theorem.

It remains to prove the claim. We use that

$$H(v_{t+1} | G_t) = \sum_G H(v_{t+1} | G_t = G) \cdot \Pr\{G_t = G\}.$$

Thus, we may work under the assumption that $G_t = G$ is fixed if $\Pr\{G_t = G\}$ is sufficiently large. Let $\varepsilon_t := 2(\log t)^{1/2}/t^{3/10}$. We consider a fixed choice G for G_t for which

$$N_{t,d} \geq \frac{4}{d(d+1)(d+2)}(1 - \varepsilon_t)(t - 1)$$

for all $d \in \{1, \dots, \lfloor t^{1/15} \rfloor\}$. Due to Corollary 10, the probability of this event is at least $1 - e^{-t/16}$.

For $d \in \{1, \dots, t+1\}$ let $p_{t,d} = d/(2t+1)$ be the probability that a fixed vertex with degree d is chosen as destination v_t of the edge from vertex t . Then we obtain

$$\begin{aligned} H(v_{t+1} | G_t = G) &= \sum_{d=1}^{t+1} N_{t,d} p_{t,d} \log(1/p_{t,d}) \geq \sum_{d=1}^{\lfloor t^{1/15} \rfloor} N_{t,d} p_{t,d} \log(1/p_{t,d}) \\ &\geq \frac{4(1 - \varepsilon_t)t}{2t+1} \left(\log(2t+1) \sum_{d=1}^b \frac{1}{(d+1)(d+2)} - \sum_{d=1}^b \frac{\log d}{(d+1)(d+2)} \right), \end{aligned}$$

where we use the abbreviation $b := \lfloor t^{1/15} \rfloor$.

For the first of the above two sums we can apply telescope summation:

$$\sum_{d=1}^b \frac{1}{(d+1)(d+2)} = \sum_{d=1}^b \left(\frac{1}{d+1} - \frac{1}{d+2} \right) = \frac{1}{2} - \frac{1}{b+2}.$$

The second sum has already been estimated in Proposition 8:

$$\sum_{d=1}^b \frac{\log d}{(d+1)(d+2)} \leq \frac{1}{12} + \frac{\log 3}{20} + 2 - \frac{3}{4} \log 3 = \frac{25}{12} - \frac{7}{10} \log 3 = c.$$

Putting the results for the two sums together yields

$$H(v_{t+1} | G_t = G) \geq \frac{4t}{2t+1} (1 - \varepsilon_t) \left(\log(2t+1) \left(\frac{1}{2} - \frac{1}{b+2} \right) - c \right).$$

Following the plan outlined at the beginning, we get a lower bound on $H(v_{t+1} | G_t)$ by summing the above inequality over all choices for G , weighted by the appropriate probabilities. The probability of each G for which the above estimate works is at least $1 - e^{-t/16}$. This gives

$$\begin{aligned} H(v_{t+1} | G_t) &\geq (1 - e^{-t/16}) \frac{4t}{2t+1} (1 - \varepsilon_t) \left(\log(2t+1) \left(\frac{1}{2} - \frac{1}{b+2} \right) - c \right) \\ &= (\log(2t+1) - 2c) (1 - e^{-t/16}) \frac{2t}{2t+1} (1 - \varepsilon_t) \left(1 - \frac{2 \log(2t+1)}{(\log(2t+1) - 2c)(b+2)} \right). \end{aligned}$$

Using that $\varepsilon_t = 2(\log t)^{1/2}/t^{3/10}$ and $b = \lfloor t^{1/15} \rfloor$, the product of the four rightmost factors can be upper bounded by $1 - t^{-\beta}$ for a suitable constant β with $0 < \beta < 1$, which yields

$$H(v_{t+1} | G_t) \geq (\log(2t+1) - 2c)(1 - t^{-\beta}).$$

This proves the claim and thus the theorem. \square

It remains to prove the result for arbitrary outdegrees m and $G_t^{(m)}$.

Proof of Theorem 2: The upper bound is again trivial, since there are $(T!)^m$ possible realizations of the random variable $G_T^{(m)}$.

For the lower bound, we use that $G_T^{(m)}$ can be derived from $G_{mT}^{(1)}$ by identifying each block of m consecutive vertices. To simplify notation, we choose $\{0, \dots, T-1\}$ and $\{0, \dots, mT-1\}$ as the sets of vertices of $G_T^{(m)}$ and $G_{mT}^{(1)}$, resp. For $t = 1, \dots, T$, let $X_{t,1}, \dots, X_{t,m}$ be the random destinations of the edges leaving the t -th block of vertices, $m(t-1), \dots, mt-1$, in $G_{mT}^{(1)}$. Let $X'_{t,1}, \dots, X'_{t,m}$ be the corresponding destinations of the edges leaving vertex $t-1$ in $G_T^{(m)}$. Then, for $i = 1, \dots, m$, $X'_{t,i} = \lfloor X_{t,i}/m \rfloor$. By Proposition 6, for $i = 1, \dots, m$,

$$H(X'_{t,i} | G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) \geq H(X_{t,i} | G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) - \log m. \quad (1)$$

For $t = 1, \dots, T$ let $Y_t := (X_{t,1}, \dots, X_{t,m})$ and $Y'_t := (X'_{t,1}, \dots, X'_{t,m})$. Then $G_{mT}^{(1)}$ may be identified with Y_1, \dots, Y_T and $G_T^{(m)}$ with Y'_1, \dots, Y'_T .

By the chain rule of entropy and (1),

$$\begin{aligned} H(G_T^{(m)}) &= \sum_{t=1}^T H(Y'_t | G_{t-1}^{(m)}) \\ &= \sum_{t=1}^T \sum_{i=1}^m H(X'_{t,i} | G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) \\ &\geq \sum_{t=1}^T \sum_{i=1}^m (H(X_{t,i} | G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) - \log m). \end{aligned} \quad (2)$$

We observe that the vector of random variables $(G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1})$ is a function of $(G_{m(t-1)}^{(1)}, X_{t,1}, \dots, X_{t,i-1})$. Hence,

$$H(G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) \leq H(G_{m(t-1)}^{(1)}, X_{t,1}, \dots, X_{t,i-1})$$

and thus

$$H(X_{t,i} | G_{t-1}^{(m)}, X'_{t,1}, \dots, X'_{t,i-1}) \geq H(X_{t,i} | G_{m(t-1)}^{(1)}, X_{t,1}, \dots, X_{t,i-1}).$$

Applying this in (2) gives

$$\begin{aligned} H(G_T^{(m)}) &\geq \sum_{t=1}^{T-1} \sum_{i=1}^m (H(X_i | G_{m(t-1)}^{(1)}, X_{t,1}, \dots, X_{t,i-1}) - \log m) \\ &= H(G_{mT}^{(1)}) - (m \log m)T. \end{aligned}$$

The theorem now follows by applying the bounds on the entropy of $G_{mT}^{(1)}$ obtained from Theorem 1. \square

4.2. The Linear Growth Copying Model

Proof of Theorem 3, Upper Bound: By the chain rule of entropy, it again suffices to bound $H(G_{t+1} | G_t)$. Let v_1, \dots, v_d be the random successors of the vertex $t+1$ added in G_{t+1} , then $H(G_{t+1} | G_t) = H(v_1, \dots, v_d | G_t)$. Let $B_1, \dots, B_d \in \{0, 1\}$ be the independent random variables with $B_i = 1$ if the i -th successor of vertex $t+1$ is copied and $B_i = 0$ if it is chosen uniformly at random. We observe that

$$H(v_1, \dots, v_d | G_t) \leq H(v_1, \dots, v_d | G_t, B_1, \dots, B_d) + d.$$

In what follows, we show that

$$H(v_1, \dots, v_d | G_t, B_1, \dots, B_d) \leq (\alpha d + 1 - \alpha^d) \log t,$$

which implies the upper bound part of the theorem.

For $I \subseteq \{1, \dots, d\}$ let E_I be the event that $B_i = 1$ for $i \in I$ and $B_i = 0$ for $i \notin I$. Then, by the definition of conditional entropy,

$$\begin{aligned} H(v_1, \dots, v_d | G_t, B_1, \dots, B_d) &= \sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\} \\ |I|=k}} (1 - \alpha)^k \alpha^{d-k} H(v_1, \dots, v_d | G_t, E_I). \end{aligned}$$

We consider the entropy term in the above sum. We observe that the random variables in the group of copied successors of vertex $t+1$ on the one hand and in the group of uniformly random successors on the other are independent of each other. Hence, we may add the entropies for these types of successors. For the first type, it is at most $\log t$, since these successors are functions of the random variable describing the prototype. For the $d - k$ uniformly random successors, the entropy is obviously $(d - k) \log t$. Hence,

$$H(v_1, \dots, v_d | G_t, E_I) \begin{cases} \leq (d - k + 1) \log t, & \text{for } |I| = k \geq 1; \\ = d \log t, & \text{for } |I| = k = 0. \end{cases}$$

Substituting this into the above sum gives

$$\begin{aligned}
& H(v_1, \dots, v_d \mid G_t, B_1, \dots, B_d) \\
& \leq \log t \left(\alpha^d d + \sum_{k=1}^d \binom{d}{k} (1-\alpha)^k \alpha^{d-k} (d-k+1) \right) \\
& = \log t (\alpha^d d + \alpha d + 1 - \alpha^d (d+1)) \\
& = \log t (\alpha d + 1 - \alpha^d),
\end{aligned}$$

as desired. \square

Next, we prepare the proof of the lower bound in Theorem 3. An essential ingredient is an estimate of the number $D_{t,I}$ of vertices in G_t that differ in the restriction of their successor arrays to the indices in the set $I \subseteq \{1, \dots, d\}$. We prove the following lower bound on $D_{t,I}$ in advance.

Lemma 11. *Let $I \subseteq \{1, \dots, d\}$ with $|I| = k \geq 1$ and let $0 < \delta < 1/2$ be a constant with respect to t . Then there are constants $c > 0$ and $t_0 \in \mathbb{N}$ with respect to t such that for all $t \geq t_0$, with probability at least $1 - 2e^{-t^{1-2\delta}/2}$,*

$$D_{t,I} \geq \beta_k t - ct^{1-\delta},$$

where $\beta_k = 1 - ((1 + (k-1)\alpha^2)/(1+\alpha)) \cdot (1-\alpha)^{k-1}$. Furthermore, $\beta_1 = \alpha/(1+\alpha) \leq \beta_k \leq 1$.

For the proof of this lemma, we apply the method of bounded differences based on the following fact (see, e. g., [22], Section 4.4.3).

Lemma 12. *Let $(G_t)_{t \in \mathbb{N}}$ be a sequence of random graphs, where G_t for $t \geq 2$ is obtained from G_{t-1} by adding a vertex and some random edges from this vertex to existing vertices. For $t \in \mathbb{N}$ let X_t be a random variable of G_t , i. e., a mapping from the set of possible realizations of G_t to real values, such that, for $t \geq 2$, $|X_t - X_{t-1}| \leq c$. Then for all $\lambda > 0$ and all $t \in \mathbb{N}$,*

$$\Pr\{|X_t - E(X_t)| > \lambda\} \leq 2e^{-\lambda^2/(2c^2t)}.$$

Furthermore, we use an estimate of the number $N_{t,0}$ of vertices in G_t with indegree 0 due to Kumar et al.

Lemma 13 ([20]). *Let $0 < \delta < 1$. For all $t \in \mathbb{N}$, with probability at least $1 - e^{-t^{1-2\delta}/4}$,*

$$N_{t,0} \geq \frac{t + \alpha}{1 + \alpha} - \alpha^2 \ln t - t^{1-\delta}.$$

Proof of Lemma 11: We have $D_{1,I} = 1$ and $D_{u+1,I} = D_{u,I} + X_{u+1,I}$, where $X_{u+1,I} = 1$ if the new vertex added in G_{u+1} in step $u+1$ is such that the restriction of its successor array to the index set I differs from the respective parts of the successor arrays of all previous vertices and $X_{u+1,I} = 0$ otherwise.

Let E_r be the event that exactly r successors of vertex $u + 1$ of those with index in I are chosen uniformly at random from $\{1, \dots, u\}$. For any fixed graph G after step u , we have

$$\Pr\{X_{u+1,I} = 0 \mid G_u = G\} = \sum_{r=0}^k \binom{k}{r} (1-\alpha)^{k-r} \alpha^r \Pr\{X_{u+1,I} = 0 \mid G_u = G, E_r\}.$$

Suppose that G satisfies the lower bound on the number of vertices with indegree 0 from Lemma 13. The probability of obtaining such a G as an instance of G_u is at least $1 - e^{-u^{1-2\delta}/4}$. For the case $r = 1$ in the above sum, i. e., a single random successor of vertex $u + 1$, the probability of choosing a successor that already occurs among the first u vertices is exactly

$$1 - \frac{N_{u,0}}{u} \leq \frac{\alpha}{1+\alpha} + \varepsilon_u,$$

with $\varepsilon_u := u^{-\delta} + \alpha^2(\ln u)/u - (\alpha/(1+\alpha))u^{-1}$.

For the case $r \geq 2$ we observe that the probability that the restriction of the successor array of vertex $u + 1$ to the r random successors agrees with a fixed vector from $\{1, \dots, u\}^r$ is u^{-r} . On the other hand, there are at most u such vectors which occur as part of the successor arrays of the previous vertices $1, \dots, u$. Hence, the probability that the restriction of the successor array to I of the new vertex agrees with the respective part of a successor array of a previous vertex is at most $u^{-(r-1)} \leq 1/u$. Altogether,

$$\begin{aligned} & \Pr\{X_{u+1,I} = 0 \mid G_u = G\} \\ & \leq (1-\alpha)^k + k(1-\alpha)^{k-1}\alpha \left(1 - \frac{N_{u,0}}{u}\right) + \sum_{r=2}^k \binom{k}{r} (1-\alpha)^{k-r} \alpha^r u^{-(r-1)} \\ & \leq \frac{1+(k-1)\alpha^2}{1+\alpha} (1-\alpha)^{k-1} + k(1-\alpha)^{k-1}\alpha\varepsilon_u + \frac{1}{u}. \end{aligned}$$

Maximizing with respect to α yields $k(1-\alpha)^{k-1}\alpha \leq (1-1/k)^{k-1} \leq e^{-1+1/k} \leq 1$, thus

$$\Pr\{X_{u+1,I} = 0 \mid G_u = G\} \leq \frac{1+(k-1)\alpha^2}{1+\alpha} (1-\alpha)^{k-1} + \varepsilon_u + \frac{1}{u}.$$

By summing the above over all appropriate G satisfying the lower bound in Lemma 13, we get

$$\Pr\{X_{u+1,I} = 1\} \geq \beta_k - \varepsilon'_u,$$

with

$$\beta_k = 1 - \frac{1+(k-1)\alpha^2}{1+\alpha} (1-\alpha)^{k-1} \quad \text{and} \quad \varepsilon'_u := \varepsilon_u + \frac{1}{u} + e^{-u^{1-2\delta}/4}.$$

It follows that

$$E(D_{t,I}) = 1 + \sum_{u=1}^{t-1} \Pr\{X_{u+1,I} = 1\} \geq 1 + (t-1)\beta_k - \sum_{u=1}^{t-1} \varepsilon'_u.$$

We estimate the last term in the above lower bound. By Proposition 7, it follows that

$$\sum_{u=1}^{t-1} \varepsilon'_u = \sum_{u=1}^{t-1} \left(\varepsilon_u + \frac{1}{u} + e^{-u^{1-2\delta}/4} \right) = O(t^{1-\delta})$$

and thus, for suitable constants $c_1 > 0$, $t_1 \in \mathbb{N}$ and $t \geq t_1$,

$$E(D_{t,I}) \geq \beta_k t - c_1 t^{1-\delta}.$$

Since obviously $|X_{t,I}| \leq 1$ for all t , we can apply Lemma 12. For $\lambda := t^{1-\delta}$ this yields that there are constants $c_2 > 0$, $t_2 \in \mathbb{N}$ such that for all $t \geq t_2$ with probability at least $1 - 2e^{-t^{1-2\delta}/2}$,

$$D_{t,I} \geq E(D_{t,I}) - c_2 t^{1-\delta} \geq \beta_k t - (c_1 + c_2) t^{1-\delta}.$$

We thus obtain the result claimed in the lemma if we set $c := c_1 + c_2$ and $t_0 := \max\{t_1, t_2\}$. It is obvious from its definition and the fact that $\alpha \in (0, 1)$ that $\beta_k \leq 1$. For the lower bound on β_k , it suffices to observe that β_k is monotonously increasing in k , which can be shown by standard arguments from calculus (see the Appendix). \square

We are now ready to complete the proof of the main result of this section.

Proof of Theorem 3, Lower Bound: We will prove the following claim.

Claim. Let $0 < \delta < 1/2$. For suitable constants $c > 0$, $t_0 \in \mathbb{N}$ with respect to t and for all $t \geq t_0$,

$$H(G_{t+1} | G_t) \geq (\alpha d + 1 - c_{\alpha,d}) \log t - c(\log t)/t^\delta.$$

where

$$c_{\alpha,d} = (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \alpha^2 \frac{1-\alpha}{1+\alpha} d \right) \leq 1.$$

First, we show how this implies the lower bound in the theorem. By the chain rule of entropy and the claim,

$$H(G_T) = \sum_{t=1}^{T-1} H(G_{t+1} | G_t) \geq \sum_{t=t_0}^{T-1} (c' \log t - c(\log t)/t^\delta),$$

where $c' := \alpha d + 1 - c_{\alpha,d}$. Thus, applying Proposition 7,

$$\begin{aligned} H(G_T) &\geq c' \sum_{t=1}^{T-1} \log t - c' \sum_{t=1}^{t_0-1} \log t - o(T) \\ &= c' \log((T-1)!) - c' \log((t_0-1)!) - o(T) \\ &= c' T \log T - (c'/\ln 2) \cdot T - o(T). \end{aligned}$$

This implies the desired lower bound since $c'/\ln 2 = O(d)$.

We now prove the above claim. The proof essentially follows along the same lines as the proof of the upper bound of the theorem. We consider the random choice of the successors v_1, \dots, v_d of the new vertex $t + 1$ in G_{t+1} . Let again $B_1, \dots, B_d \in \{0, 1\}$ be the independent random variables where $B_i = 1$ if the i -th successors is copied from the prototype and $B_i = 0$ if it is chosen uniformly at random. For $I \subseteq \{1, \dots, d\}$ let E_I be the event that $B_i = 1$ for $i \in I$ and $B_i = 0$ for $i \notin I$. Then we obtain

$$\begin{aligned} H(G_{t+1} \mid G_t) &\geq H(G_{t+1} \mid G_t, B_1, \dots, B_d) \\ &= \sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\}, \\ |I|=k}} (1 - \alpha)^k \alpha^{d-k} H(G_{t+1} \mid G_t, E_I) \end{aligned}$$

We handle the contributions to the above sum by the random successors and by the copied successors of vertex $t + 1$ separately. We may add the results since the choices of these successors are independent of each other. We observe that the choice of the random successors of vertex $t + 1$ does not depend on the random choices for G_t . It is therefore easy to see that their contribution to the above sum is $\alpha d \log t$. It remains to lower bound the contribution of the copied successors.

For this, we investigate

$$S := \sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\}, \\ |I|=k}} (1 - \alpha)^k \alpha^{d-k} H(v_i, i \in I \mid G_t, E_I).$$

This may be expanded to

$$\sum_{k=0}^d \sum_{\substack{I \subseteq \{1, \dots, d\}, \\ |I|=k}} (1 - \alpha)^k \alpha^{d-k} \sum_G H(v_i, i \in I \mid G_t = G, E_I) \cdot \Pr\{G_t = G\}. \quad (1)$$

Fix a set $I \subseteq \{1, \dots, d\}$ with $|I| = k \geq 1$ and let $0 < \delta < 1/2$ be any constant. By Lemma 11, we get constants $c_1 > 0$ and $t_1 \in \mathbb{N}$ such that, for $t \geq t_1$, with probability at least $1 - 2e^{-t^{1-2\delta}/2}$ there is a set U of $u := |U| \geq \beta_k t - c_1 t^{1-\delta}$ vertices of G_t that pairwise differ in the restrictions of their successor arrays to the indices in I . We consider the entropy in the innermost sum in (1) for a graph G of the described kind. Since the prototype w is chosen uniformly at random from $\{1, \dots, t\}$, it follows that

$$\begin{aligned} H(v_i, i \in I \mid G_t = G, E_I) &\geq H(v_i, i \in I \mid G_t = G, E_I, w \in U) \cdot \Pr\{w \in U\} \\ &\geq (\log u)(\beta_k - c_1 t^{-\delta}). \end{aligned}$$

Now we lower bound $\log u$. Using that $\beta_k \geq \alpha/(1 + \alpha)$ (by Lemma 11) and $\log(1 - x) \geq -2x$ for all $x \leq 1/2$, we get

$$\log u \geq \log(\beta_k t - c_1 t^{1-\delta}) = \log t + \log(\beta_k - c_1 t^{-\delta}) = \log t - O(t^{-\delta}).$$

Thus, for suitable constants $c_2 > 0$, $t_2 \in \mathbb{N}$ and $t \geq t_2$,

$$H(v_i, i \in I \mid G_t = G, E_I) \geq \log t - c_2 t^{-\delta}.$$

We may assume that the constants occurring in the above estimates for different I are independent of I by taking their maximum. Substituting the obtained result into the above sum (1) for the terms with $k \geq 1$, we thus get that

$$\begin{aligned} S &\geq (1 - 2e^{-t^{1-2\delta}/2}) \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} (\log t - c_2 t^{-\delta}) (\beta_k - c_1 t^{-\delta}) \\ &= (1 - 2e^{-t^{1-2\delta}/2}) (\log t) \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \beta_k - O((\log t)/t^\delta). \end{aligned}$$

This yields (see the Appendix for some details of the calculations):

$$\begin{aligned} \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \beta_k &= \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \left(1 - \frac{1 + (k-1)\alpha^2}{1 + \alpha} (1 - \alpha)^{k-1}\right) \\ &= 1 - \alpha^d - \frac{1}{1 - \alpha^2} \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^{2k} \alpha^{d-k} (1 + (k-1)\alpha^2) \\ &= 1 - (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \alpha^2 \frac{1 - \alpha}{1 + \alpha} d\right). \end{aligned}$$

Since $\beta_k \leq 1$, the above term is upper bounded by 1. Using this, it follows that

$$2e^{-t^{1-2\delta}/2} (\log t) \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \beta_k = O((\log t) e^{-t^{1-2\delta}/2}).$$

Hence, for suitable constants $c > 0$, $t_0 \in \mathbb{N}$ and $t \geq t_0$,

$$S \geq (1 - c_{a,d}) \log t - c(\log t)/t^\delta,$$

where

$$c_{a,d} = (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \alpha^2 \frac{1 - \alpha}{1 + \alpha} d\right).$$

Putting the estimates together yields

$$H(G_{t+1} | G_t) \geq (\alpha d + 1 - c_{a,d}) \log t - c(\log t)/t^\delta.$$

Since we have already noticed that $c_{a,d} \leq 1$, this completes the proof of the claim. \square

4.3. A Simple Hierarchical Model

In what follows, we prepare the proof of Theorem 4 by some lemmas. First, we investigate the entropy of an object chosen uniformly from a set whose size is itself distributed by a power law.

Lemma 14. *Let $t \in \mathbb{N}$ and let $\beta > 2$ and $c, c' > 0$ be constants with respect to t . Let $x_1, \dots, x_t \in \{0, \dots, t\}$ with $\sum_{i=1}^t x_i = t$ and $c/x^\beta \leq |\{i \mid x_i = x\}|/t \leq c'/x^\beta$ for $x \in \{1, \dots, t\}$. Let Y be a random variable with $\Pr\{Y = i\} = x_i/t$ for $i \in \{1, \dots, t\}$. Let $Z \in \{1, \dots, x_Y\}$ be chosen uniformly at random. Then $H(Z)$ is upper bounded by a constant with respect to t that depends only on β , c , and c' .*

Proof: Let $N_x := |\{i \mid x_i = x\}|$ for $x \in \{1, \dots, t\}$. Then $\Pr\{x_Y = x\} = (x/t) \cdot N_x$ and it follows by the assumptions that, for all $x \in \{1, \dots, t\}$,

$$c/x^{\beta-1} \leq \Pr\{x_Y = x\} \leq c'/x^{\beta-1}.$$

Using this, we can approximate the distribution of Z as follows.

Claim. For $z \in \{1, \dots, t\}$,

$$\frac{c}{\beta-1} \cdot (z^{-(\beta-1)} - (t+1)^{-(\beta-1)}) \leq \Pr\{Z = z\} \leq \frac{\beta c'}{\beta-1} \cdot z^{-(\beta-1)}.$$

Proof of the Claim: By the definitions,

$$\Pr\{Z = z\} = \sum_{x=z}^t \frac{1}{x} \cdot \Pr\{x_Y = x\}.$$

Using the approximation of the distribution of x_Y , it follows that

$$\sum_{x=z}^t \frac{c}{x^\beta} \leq \Pr\{Z = z\} \leq \sum_{x=z}^t \frac{c'}{x^\beta}.$$

We estimate the above sums using integrals. First, we can lower bound the left hand sum by

$$\int_z^{t+1} \frac{c}{x^\beta} dx = \frac{c}{\beta-1} \left(z^{-(\beta-1)} - (t+1)^{-(\beta-1)} \right),$$

which gives the lower bound in the claim. Using that $z \geq 1$, we get the following upper bound on the right hand sum, giving the upper bound in the claim:

$$\begin{aligned} \frac{c'}{z^\beta} + \int_z^t \frac{c'}{x^\beta} dx &= \frac{c'}{z^\beta} + \frac{c'}{\beta-1} \left(z^{-(\beta-1)} - t^{-(\beta-1)} \right) \leq c' \cdot z^{-(\beta-1)} + \frac{c'}{\beta-1} \cdot z^{-(\beta-1)} \\ &= \frac{\beta c'}{\beta-1} \cdot z^{-(\beta-1)}. \end{aligned}$$

□

Now we are ready to estimate the entropy of Z . By the above estimates for $\Pr\{Z = z\}$ and with $d := c/(\beta-1)$ and $d' := (\beta c')/(\beta-1)$,

$$\begin{aligned} H(Z) &= - \sum_{z=1}^t \Pr\{Z = z\} \log \Pr\{Z = z\} \\ &\leq - \sum_{z=1}^t d' z^{-(\beta-1)} \left(\log(z^{-(\beta-1)} - (t+1)^{-(\beta-1)}) + \log d \right). \end{aligned}$$

We estimate the first logarithmic term in advance: Using that $t \geq z$ and that $x \mapsto -\log x$ is monotonously decreasing, we get

$$\begin{aligned} -\log(z^{-(\beta-1)} - (t+1)^{-(\beta-1)}) &\leq -\log(z^{-(\beta-1)} - (z+1)^{-(\beta-1)}) \\ &= (\beta-1)\log z - \log\left(1 - \left(\frac{z}{z+1}\right)^{-(\beta-1)}\right). \end{aligned}$$

Since $\beta > 2$, the last term above can be upper bounded by $-\log(1 - z/(z+1)) = \log(z+1)$, which gives

$$-\log(z^{-(\beta-1)} - (t+1)^{-(\beta-1)}) \leq (\beta-1)\log z + \log(z+1).$$

Hence,

$$H(Z) \leq d'(\beta-1) \sum_{z=1}^t \frac{\log z}{z^{\beta-1}} + d' \sum_{z=1}^t \frac{\log(z+1)}{z^{\beta-1}} - d' \log d \sum_{z=1}^t \frac{1}{z^{\beta-1}}.$$

By the assumptions, β is a constant with $\beta > 2$. Hence, Proposition 7 implies that the above three sums can each be upper bounded by constants, which gives the claimed result. \square

For $t \in \mathbb{N}$ and $u \in \{1, \dots, t\}$, let $N_{t,u}$ be the number of directory trees in G_t with u URLs. The stochastic process generating the numbers of URLs in the directory trees here is easily seen to be the same as that generating the indegrees (or outdegrees) of the vertices in the random graphs according to “model A” of Aiello et al. [2]. Their results immediately yield that the distribution of $N_{t,u}$ asymptotically follows a power law as described below. A detailed proof of the fact is given in [12].

Lemma 15 ([2, 12]). *For $t \in \mathbb{N}$ and $u \in \{1, \dots, t\}$,*

$$\Pr\{|N_{t,u} - \beta_u \cdot t| > \lambda\sqrt{t} + 2\} \leq \exp(-\lambda^2/2),$$

where

$$\beta_u := \frac{(1/\alpha - 1)(u-1)!}{\prod_{j=1}^u (1/\alpha + j)} = \frac{(1/\alpha - 1)\Gamma(1/\alpha - 1)}{u^{1/\alpha+1}}(1 + \varepsilon_{\alpha,u}),$$

with $\alpha := 1 - \alpha_1$ and $|\varepsilon_{\alpha,i}| \rightarrow 0$ for $i \rightarrow \infty$.

We are now ready to prove the main theorem of this section.

Proof of Theorem 4: We observe that the total number of URLs in the URL graph U_t is t . For each URL $u \in \{1, \dots, t\}$ in U_t let E_u be the set of its in- and outlinks. We then may identify U_t with (E_1, \dots, E_t) . Following the pattern of the proofs in the previous sections, we bound the entropy added by one iteration of the process creating the URL graph, here conditioned on G_T . We show:

Claim. There is a constant $c_{\alpha_1, \alpha_3} > 0$ with respect to t depending only on α_1, α_3 and a constant $t_0 \in \mathbb{N}$ such that, for $t \geq t_0$,

$$H(E_{t+1} | U_t, G_T) \leq (\alpha_1 + d\alpha_3) \log t + c_{\alpha_1, \alpha_3}.$$

As always, we first use the claim to prove the theorem. By the chain rule of entropy and the trivial bound $H(E_t) \leq (d+1)(\log t)$ for all $t \in \{1, \dots, T\}$,

$$\begin{aligned} H(U_T | G_T) &= \sum_{t=1}^{T-1} H(E_{t+1} | U_t, G_T) \\ &\leq \sum_{t=1}^{t_0-1} (d+1) \log t + \sum_{t=t_0}^{T-1} ((\alpha_1 + d\alpha_3) \log t + c_{\alpha_1, \alpha_3}) \\ &\leq (\alpha_1 + d\alpha_3)T \log T + (c_{\alpha_1, \alpha_3} + (\alpha_1 + d\alpha_3)/\ln 2)T + o(T) + O(d). \end{aligned}$$

This implies the upper bound in the theorem.

It remains to prove the claim. First, we observe that G_T is composed of G_t and some additional random variable R accounting for the directories and URLs added in steps $t+1, \dots, T$. Hence,

$$H(E_{t+1} | U_t, G_T) = H(E_{t+1} | U_t, G_t, R) \leq H(E_{t+1} | U_t, G_t).$$

We estimate the latter entropy conditioned on $U_t = U$ and $G_t = G$ where U and G are suitably chosen. Choosing $\lambda := t^{1/2-\delta}$ for an arbitrary constant $0 < \delta < 1/2$ in Lemma 15, we get constants $c, c' > 0$ and a constant $t_1 \in \mathbb{N}$ such that for all $t \geq t_1$, with probability at least $1 - e^{-(1/4)t^{1-2\delta}}$, the following is satisfied for all $u \in \{1, \dots, t\}$:

$$c/u^\beta \leq N_{t,u}/t \leq c'/u^\beta,$$

where $\beta := 1 - 1/(1 - \alpha_1)$. Fix instances U of U_t and G of G_t where G_t satisfies the above property.

Since all individual links are chosen independently of each other, adding the contributions of each of these links yields the total entropy. First, we consider the inlinks. Due to the definition of the model, we either have, with probability α_1 , an inlink from an URL v_{global} chosen uniformly at random from all existing URLs, or, with the remaining probability, an inlink from an existing URL v_{local} in the directory where the new URL has been inserted or its parent directory. By Proposition 5, part (5), the total entropy of the inlink is thus

$$\alpha_1 \cdot H(v_{\text{global}}) + (1 - \alpha_1) \cdot H(v_{\text{local}}) + h_2(\alpha_1).$$

Trivially, $H(v_{\text{global}}) \leq \log t$. Furthermore, we can only overestimate $H(v_{\text{local}})$ by assuming that v_{local} is chosen uniformly at random from all URLs in the insertion tree of the new URL. Then the choice of v_{local} fits the assumptions of Lemma 14 where we set $x_u := N_{t,u}$ for $u \in \{1, \dots, t\}$ (recall that we work under the condition $G_t = G$ implying that the $N_{t,u}$ are fixed numbers). The lemma yields that $H(v_{\text{local}})$ is upper bounded by a constant depending only on the constants in the power law distribution and thus, in the application here, a constant c_{α_1} depending on α_1 . Hence, the contribution of the inlinks to the total entropy can be upper bounded by

$$\alpha_1 \log t + (1 - \alpha_1)c_{\alpha_1} + h_2(\alpha_1).$$

It remains to consider the outlinks. We have d independent contributions of links of the following kind. With probability α_3 , the destination is a URL v_{global} outside the insertion tree,

and with the remaining probability, the destination is a URL v_{local} in the insertion tree of the new URL. In both cases we can only overestimate the entropy by assuming that the destinations are chosen uniformly at random instead of with probability proportional to the indegree. Then, again, $H(v_{\text{global}}) \leq \log t$ and we may again upper bound $H(v_{\text{local}})$ by the constant c_{α_1} obtained from Lemma 14. Hence, the contribution of the outlinks is at most

$$d(\alpha_3 \log t + (1 - \alpha_3)c_{\alpha_1} + h_2(\alpha_3)).$$

Adding the contributions of in- and outlinks yields

$$\begin{aligned} H(E_{t+1} | U_t = U, G_t = G) &\leq \alpha_1 \log t + (1 - \alpha_1)c_{\alpha_1} + h_2(\alpha_1) \\ &\quad + d(\alpha_3 \log t + (1 - \alpha_3)c_{\alpha_1} + h_2(\alpha_3)) \\ &\leq (\alpha_1 + d\alpha_3) \log t + c'_{\alpha_1, \alpha_3}, \end{aligned}$$

where $c'_{\alpha_1, \alpha_3} > 0$ is a constant depending on α_1 and α_3 . Thus, by summing over all U and G of the considered type and the law of total probability, we get for all $t \geq t_1$ that

$$H(E_{t+1} | U_t, G_t) \leq (1 - e^{-(1/4)t^{1-2\delta}})((\alpha_1 + d\alpha_3) \log t + c'_{\alpha_1, \alpha_3}).$$

For a suitable constant $c_{\alpha_1, \alpha_3} > 0$ depending on α_1, α_3 and $t \geq t_0 \geq t_1, t_0 \in \mathbb{N}$ a suitable constant, this implies

$$H(E_{t+1} | U_t, G_t) \leq (\alpha_1 + d\alpha_3) \log t + c_{\alpha_1, \alpha_3}.$$

By the remarks at the beginning, this proves the claim. □

Acknowledgment

Thanks to André Gronemeier for proofreading and helpful discussions and for simplifying the proof of Proposition 6.

References

- [1] M. Adler and M. Mitzenmacher. Towards compressing web graphs. In *Proc. of the Data Compression Conference (DCC)*, 203–212, 2001.
- [2] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. In *Proc. of 42nd FOCS*, 510–519, 2001.
- [3] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [4] K. Bharat, A. Broder, M. Henzinger, P. Kumar, and S. Venkatasubramanian. The Connectivity Server: fast access to linkage information on the Web. *Computer Networks and ISDN Systems*, 30(1-7):469–477, 1998.

- [5] G. Bianconi. The entropy of randomized network ensembles. *EPL*, 2008. ID 28005.
- [6] P. Boldi, M. Santini, and S. Vigna. A large time aware Web graph. *SIGIR Forum*, 42(1), 2008.
- [7] P. Boldi and S. Vigna. The WebGraph Framework I: Compression techniques. In *Proc. of 13th World Wide Web Conference (WWW)*, 595–602, 2004.
- [8] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.
- [9] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.
- [10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. of 7th World Wide Web Conference (WWW)*, 107–117, 1998.
- [11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, and R. Stata. Graph structure in the web. In *Proc. of 9th World Wide Web Conference (WWW)*, 309–320, 2000.
- [12] A. Cami and N. Deo. Techniques for analyzing dynamic random graph models of web-like networks: An overview. *Networks*, 51(4):211–255, 2008.
- [13] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
- [14] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins. Self-similarity in the web. *ACM Transactions on Internet Technology (TOIT)*, 2(3):205–223, 2002.
- [15] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the Webgraph. *The European Physical Journal B*, 38:239–243, 2004.
- [16] N. Eiron and K. S. McCurley. Link structure of hierarchical information networks. In *Proc. of the 3rd Workshop on Algorithms and Models for the Web Graph*, 143–155, 2004.
- [17] A. Gulli and A. Signorini. The indexable Web is more than 11.5 billion pages. In *Proc. of 14th World Wide Web Conference (WWW)*, 902–903, 2005.
- [18] J. Han, Y. Yu, L. Chenxi, H. Dingyi, and G.-R. Xue. A hierarchical model of web graph. In *Proc. of 2nd Conference on Advanced Data Mining and Applications (ADMA)*, 790–797, 2006.
- [19] A. Kogias and D. Anagnostopoulos. A methodology for the evaluation of web graph models and a test case. In *Proc. of 38th Conference on Winter Simulation*, 2202–2209, 2006.
- [20] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. S. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. of 41st FOCS*, 57–65, 2000.
- [21] L. Laura, S. Leonardi, G. Caldarelli, and P. De Los Rios. A multi-layer model for the web graph. In *Proc. of 2nd International Workshop on Web Dynamics*, 2002. Online proceedings available on the Web.

- [22] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
- [23] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
- [24] J. Park and M. E. J. Newman. Statistical mechanics of networks. *Physical Review E*, 70, 2004. ID 066117.
- [25] S. Raghavan and H. Garcia-Molina. Representing web graphs. In *Proc. of 19th International Conference on Data Engineering (ICDE)*, 405–416, 2003.
- [26] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67, 2003. ID 026112.
- [27] R. V. Solé and S. Valverde. Information theory of complex networks: On evolution and architectural constraints. In *Complex Networks, Lecture Notes in Physics 650*, 189–207. Springer, 2004.
- [28] T. Suel and J. Yuan. Compressing the graph structure of the web. In *Proc. of the Data Compression Conference (DCC)*, 213–222, 2001.

Appendix

Proof of Corollary 10: By Lemma 9, we can lower bound the probability that the event

$$N_{t,d} \geq \frac{4}{d(d+1)(d+2)}t - \sqrt{t \log t}$$

occurs for all d with $1 \leq d \leq t^{1/15}$ by

$$1 - e^{-t/8} \cdot t^{1/15} \geq 1 - e^{-t/16}.$$

Suppose the lower bound for $N_{t,d}$ is valid for d . Then, by using that $d \leq t^{1/15}$, it follows that

$$\begin{aligned} N_{t,d} &\geq \frac{4 - \sqrt{(\log t)/t} \cdot (2d)^3}{d(d+1)(d+2)} \cdot t \geq \frac{4 - 8\sqrt{(\log t)/t} \cdot t^{1/5}}{d(d+1)(d+2)} \cdot t \\ &= \frac{4(1 - 2(\log t)^{1/2}/t^{3/10})}{d(d+1)(d+2)} \cdot t. \end{aligned}$$

This gives the desired result. □

Proof of Proposition 7: It only remains to handle the cases where $\beta \neq 1$. For this, we estimate the sums by integrals as follows.

$$\begin{aligned} \sum_{z=1}^t \frac{1}{z^\beta} &\leq 1 + \int_1^t \frac{1}{z^\beta} dz = 1 + \left(\frac{z^{-\beta+1}}{-\beta+1} \right) \Big|_1^t \quad \text{and} \\ \sum_{z=1}^t \frac{\log z}{z^\beta} &\leq \int_1^{t+1} \frac{\log z}{z^\beta} dz = z^{-\beta+1} \left(\frac{1}{-\beta+1} \log z - \frac{1}{(-\beta+1)^2 \ln 2} \right) \Big|_1^{t+1}. \end{aligned}$$

The estimates in the proposition now follow by using the appropriate bounds on β . □

Proof of Lemma 11, lower bound on β_k : We claim that

$$\beta_k = 1 - \frac{1 + (k-1)\alpha^2}{1 + \alpha} (1 - \alpha)^{k-1}$$

is monotonously increasing as a function of k . For this, it suffices to show that

$$(1 + (k-1)\alpha^2)(1 - \alpha)^{k-1}$$

is monotonously decreasing in k . This follows by investigating the derivative with respect to k , which is

$$\begin{aligned} (1 - \alpha)^{k-1} (\ln(1 - \alpha)(1 + (k-1)\alpha^2) + \alpha^2) &\leq (1 - \alpha)^{k-1} ((-\alpha)(1 + (k-1)\alpha^2) + \alpha^2) \\ &\leq -(1 - \alpha)^{k-1} (k-1)\alpha^3 < 0. \end{aligned}$$

□

Proof of Theorem 3, calculation details: We have

$$\begin{aligned} \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \beta_k &= \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^k \alpha^{d-k} \left(1 - \frac{1 + (k-1)\alpha^2}{1 + \alpha} (1 - \alpha)^{k-1}\right) \\ &= 1 - \alpha^d - \frac{1}{1 - \alpha^2} \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^{2k} \alpha^{d-k} (1 + (k-1)\alpha^2) =: S. \end{aligned}$$

First,

$$\sum_{k=1}^d \binom{d}{k} (1 - \alpha)^{2k} \alpha^{d-k} = ((1 - \alpha)^2 + \alpha)^d - \alpha^d = (1 - \alpha + \alpha^2)^d - \alpha^d =: A.$$

Furthermore,

$$\begin{aligned} \sum_{k=1}^d \binom{d}{k} (1 - \alpha)^{2k} \alpha^{d-k} k &= d(1 - \alpha)^2 \sum_{k=0}^{d-1} \binom{d-1}{k} (1 - \alpha)^{2k} \alpha^{d-k-1} \\ &= d(1 - \alpha)^2 (1 - \alpha + \alpha^2)^{d-1} =: B. \end{aligned}$$

Thus,

$$\begin{aligned} S &= 1 - \alpha^d - \frac{1}{1 - \alpha^2} (A + \alpha^2(B - A)) = 1 - \alpha^d - A - \frac{\alpha^2}{1 - \alpha^2} B \\ &= 1 - (1 - \alpha + \alpha^2)^d - \frac{\alpha^2}{1 - \alpha^2} d(1 - \alpha)^2 (1 - \alpha + \alpha^2)^{d-1} \\ &= 1 - (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \frac{\alpha^2}{1 - \alpha^2} d(1 - \alpha)^2\right) \\ &= 1 - (1 - \alpha + \alpha^2)^{d-1} \left(1 - \alpha + \alpha^2 + \frac{\alpha^2(1 - \alpha)}{1 + \alpha} d\right). \end{aligned}$$

□