Models for the Compressible Web

Flavio Chierichetti*, Ravi Kumar†, Silvio Lattanzi*, Alessandro Panconesi*, Prabhakar Raghavan†
* Dipartimento di Informatica, Sapienza University of Rome, Italy.
Email: {chierichetti, lattanzi, ale}@di.uniroma1.it
† Yahoo! Research, 701 First Ave., Sunnyvale, CA 94089, USA.
Email: {ravikumar, pragh}@yahoo-inc.com

***Abstract*—Graphs resulting from human behavior (the web graph, friendship graphs, etc.) have hitherto been viewed as a monolithic class of graphs with similar characteristics; for instance, their degree distributions are markedly heavy-tailed. In this paper we take our understanding of behavioral graphs a step further by showing that an intriguing empirical property of web graphs — their compressibility — cannot be exhibited by well-known graph models for the web and for social networks. We then develop a more nuanced model for web graphs and show that it does exhibit compressibility, in addition to previously modeled web graph properties.**

I. Overview

There are three main reasons for modeling and analyzing graphs arising from the Web and from social networks: (i) they model social and behavioral phenomena whose graph-theoretic analysis has led to significant societal impact (witness the role of link analysis in web search); (ii) from an empirical standpoint, these networks are several orders of magnitude larger than those studied hitherto (search companies are now working on crawls of 100 billion pages and beyond); (iii) from a theoretical standpoint, stochastic processes built from independent random events — the classical basis of the design and analysis of computing artifacts — are no longer appropriate. The characteristics of such *behavioral* graphs (viz., graphs arising from human behavior) demand the design and analysis of new stochastic processes in which elementary events are highly dependent. This in turn demands new analysis and insights that are likely to be of utility in many other applications of probability and statistics.

In such analysis, there has been a tendency to lump together behavioral graphs arising from a variety of contexts, to be studied using a common set of models and tools. It has been observed [3], [9], [22] for instance that the directed graphs arising from such diverse phenomena as the web graph (pages are nodes and hyperlinks are edges), citation graphs, friendship graphs, and email traffic graphs all exhibit *power laws* in their degree distributions: the fraction of nodes with in-degree $k > 0$ is proportional to $1/k^{\alpha}$ typically for some $\alpha > 1$; random graphs generated by classic Erdös–Rényi models cannot exhibit such power laws. To explain the power law degree distributions seen in behavioral graphs, several models have been developed [2], [3], [7], [8], [11], [17], [21], [25] for generating random graphs in which dependent events combine to deliver the observed power laws.

While the degree distribution is a fundamental but local property of such graphs, an important global property is their compressibility — the number of bits needed to store each edge in the graph. Compressibility determines the ability to efficiently store and manipulate these massive graphs [18], [31], [35]. An intriguing set of papers by Boldi, Santini, and Vigna [4]–[6] shows that the web graph is highly compressible: it can be stored such that each edge requires only a small constant number — between one and three — of bits on average; a more recent experimental study confirms these findings [10]. These empirical results suggest the intriguing possibility that the Web can be described with only $O(1)$ bits per edge on average. Two properties are at the heart of the compression algorithm of Boldi and Vigna [5]. First, once web pages are sorted lexicographically by URL, the set of out-links of a page exhibits locality; this can plausibly be attributed to the fact that nearby pages are likely to come from the same website's template. Second, the distribution of the lengths of edges follows a power law with exponent > 1 (the length of an edge is the distance of its endpoints in the ordering); this turns out to be crucial for high compressibility. This prompts the natural question: can we model the compressibility of the web graph, in particular mirroring the properties of locality and edge length distribution, while maintaining other well-known properties such as power law degree distribution?

Main results. Our first set of results in this paper is to show that the best known models for the web graph cannot account for compressibility, in the sense that they require $\Omega(\log n)$ bits storage per edge on average. This holds even when these graphs are represented just in terms of their topology (i.e., with all labels stripped away). Specifically, we show that the preferential attachment (PA) model [3], [7], the ACL model [2],

the copying model [21], the Kronecker product model [24], and Kleinberg's model for navigability[1] on social networks [19], all have large entropy in the above sense.

We then show our main result: a new model for the web graph that has constant entropy per edge, while preserving crucial properties of previous models such as the power law distribution of in-degrees, a large number of communities (i.e., bipartite cliques), small diameter, and a high clustering coefficient. In this model, nodes lie on the line and when a new node arrives it selects an existing node uniformly at random, placing itself on the line to the immediate left of the chosen node. An edge from the new to the chosen node is added, and moreover all outgoing edges of the chosen node but one are copied (these edges are chosen at random); thus, the edges have some locality. We then show a crucial property of our model: the power law distribution of edge lengths. Intuitively, this long-get-longer effect is caused since a long edge is likely to receive the new node (which selects its position uniformly at random) under its protective wing, and the longer it gets, the more likely it is to attract new nodes. Using this, we show that the graphs generated by our model are compressible to $O(1)$ bits per edge; we also provide a linear-time algorithm to compress an unlabeled graph generated by our model.

Technical contributions and guided tour. In Section III we prove that several well-known web graph models are not compressible, i.e., they need $\Omega(\log n)$ bits per edge. In fact, we prove incompressibility even after the labels of nodes and orientations of edges are removed.

Section IV presents our new model and Sections V and VI present the basic properties of our model. Although our new model might at first sight closely resemble a prior *copying* model of [21], it differs in fundamental respects. First, our new model successfully admits the global property of compressibility which the copying model *provably* does not. Second, while the analysis of the distribution of the in-degrees is rather standard, the crucial property that edge lengths are distributed according to a power law requires an entirely novel analysis; in particular, the proof requires a very delicate understanding of the structural properties of the graphs generated by our model in order to establish the concentration of measure. Section VII addresses the compressibility of our model, where we also provide an efficient algorithm to compress graphs generated by our model.

It is difficult to distinguish experimentally between graphs that require only $O(1)$ bits per edge and those requiring, say, $\epsilon \lg n$ bits. The point however is that the compressibility of our model relies upon other important structural properties of real web graphs that previous models, in view of our lower bounds, provably cannot have.

Related prior work. The observation of power law degree distributions in behavioral (and other) graphs has a long history [3], [22]; indeed, such distributions predate the modern interest in social networks through observations in linguistics [36] and sociology [30]; see the survey by Mitzenmacher [28]. Simon [30], Mandelbrot [27], Zipf [36] and others have provided a number of explanations for these distributions, attributing them to the dependencies between the interacting humans who collectively generate these statistics. These explanations have found new expression in the form of rich-get-richer and herd-mentality theories [3], [34]. Early rigorous analyses of such models include [2], [7], [13], [21]. Whereas Kumar et al. [21] and Borgs et al. [8] focused on modeling the web graph, the models of Aiello, Chung, and Lu (ACL) [2], Kleinberg [19], Lattanzi and Sivakumar [23], and Leskovec et al. [24] addressed social graphs in which people are nodes and the edges between them denote friendship. The ACL model is in fact known not to be a good representation of the web graph [22], but is a plausible model for human social networks. Kleinberg's model of social networks focuses on their *navigability*: it is possible for a node to find a short route to a target using only local, myopic choices at each step of the route. The papers by Boldi, Santini, and Vigna [4]–[6] suggests that the web graph is highly compressible (see also [1], [10], [12], [31]).

II. Preliminaries

The graph models we study will either have a fixed number of nodes or will be evolving models in which nodes arrive in a discrete-time stochastic process; for many of them, the number of edges will be linear in the number of nodes. We analyze the space needed to store a graph randomly generated by the models under study; this can be viewed in terms of the entropy of the graph generation process. Note that a naive representation of a graph would require $\Theta(\log n)$ bits per edge; entropically, one can hope for no better for an Erdös–Rényi graph. We are particularly interested in the case when the amortized storage per edge can be reduced to a constant. As in the work of Boldi and Vigna [5], [6], we view the nodes as being arranged in a linear order. To prove compressibility we then study the distribution of edge *lengths* — the distance in this linear order between the end-points of an edge.

Background. Given a function $f : A_1 \times \cdots \times A_n \to \mathbb{R}$, we say that f satisfies the *c-Lipschitz property* if, for any

[1] Since navigability is a crucial property of real-life social networks (cf. [16], [26], [33]), it is tempting to conjecture that social networks are incompressible; see, for instance, [12].

sequence $(a_1, \ldots, a_n) \in A_1 \times \cdots \times A_n$, and for any i and $a_i' \in A_i$, we have

$$|f(a_1, \ldots, a_{i-1}, a_i, a_{i+1}, \ldots, a_n) - f(a_1, \ldots, a_{i-1}, a_i', a_{i+1}, \ldots, a_n)| \leq c.$$

In order to establish that certain events occur w.h.p., we will make use of the following concentration result known as the *method of bounded differences* (cf. [15]).

Theorem 1 (Method of bounded differences). *Let $X_1, \ldots, X_n$ be independent r.v.'s. Let f be a function on $X_1, \ldots, X_n$ satisfying the c-Lipschitz property. Then,*

$$\Pr\left[|f(X_1, \ldots, X_n) - \mathrm{E}\left[f(X_1, \ldots, X_n)\right]| > t\right] \leq 2\exp\left(-\frac{t^2}{c^2 n}\right).$$

The *Gamma function* is defined as $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. We use the following properties of the Gamma function: (i) $\Gamma(x+1) = x\Gamma(x)$, (ii) $\Gamma(x)\Gamma(x+\frac{1}{2}) = \Gamma(2x)2^{1-2x}\sqrt{\pi}$, (iii) for constants $a, b \in \mathbb{R}$, $\lim_{n\to\infty} \frac{\Gamma(n+a)}{\Gamma(n+b)} n^{b-a} = 1$; the following lemma, whose proof we omit for lack of space, will also be used in our analysis.

Lemma 2. *Let $a, b \in \mathbb{R}^+$ be such that $b \neq a+1$. For each $t \in \mathbb{Z}^+$, it holds that*

$$\sum_{i=1}^{t} \frac{\Gamma(i+a)}{\Gamma(i+b)} = \frac{1}{b-a-1}\left(\frac{\Gamma(a+1)}{\Gamma(b)} - \frac{\Gamma(t+a+1)}{\Gamma(t+b)}\right).$$

Throughout the paper, we will use $\lg x$ and $\ln x$ for denoting, respectively, the binary and the natural logarithm of x.

III. Incompressibility of the existing models

In this section we prove the inherent incompressibility of commonly-studied random graph models for social networks and the web. We show that on average $\Omega(\log n)$ bits per edge are necessary to store graphs generated by several well-known models for web/social networks, including the PA and the copying models. In our lower bounds, we show that the random graph produced by the models we consider are incompressible, *even after removing the labels of their nodes and orientations of their edges*. Given a labeled/directed graph and its unlabeled/undirected counterpart, the latter is more compressible than the former; in fact, the gap can be arbitrarily large. Thus the task of proving incompressibility of unlabeled/undirected versions of graphs generated by various models is made more challenging. Note that it is crucial to analyze the compressibility of unlabeled graphs — the experiments on web graph [5], [6] show how its edges alone can be compressed using ≈ 2 bits per edge.

A. Proving incompressibility

Let $\mathcal{G}_n$ denote the set of all directed labeled graphs on n nodes. Let $P_n^\theta : \mathcal{G}_n \to [0,1]$ denote the probability distribution on $\mathcal{G}_n$ induced by the random graph model θ. In this paper we consider the PA model ($\theta = \mathsf{pref}$), the ACL model ($\theta = \mathsf{acl}$), the copying model ($\theta = \mathsf{copy}$), the Kronecker multiplication model ($\theta = \mathsf{kron}$), and Kleinberg's model ($\theta = \mathsf{kl}$).

For a given θ, let $\mathrm{H}(P_n^\theta)$ denote the Shannon entropy of the distribution P_n^θ, that is, the average number of bits needed to represent a directed labeled random graph generated by θ. Our goal is to obtain lower bounds on the representation. This is accomplished by the following min-entropy argument.

Lemma 3 (Min-entropy argument). *Let $\mathcal{G}_n^* \subseteq \mathcal{G}_n$, $P^+ \leq \sum_{G\in\mathcal{G}_n^*} P_n^\theta(G)$, and $P^* \geq \max_{G\in\mathcal{G}_n^*} P_n^\theta(G)$. Then, $\mathrm{H}(P_n^\theta) \geq P^+ \cdot \lg(1/P^*)$.*

Proof:

$$\begin{aligned}\mathrm{H}(P_n^\theta) &= \sum_{G\in\mathcal{G}_n} P_n^\theta(G) \lg \frac{1}{P_n^\theta(G)} \\ &\geq \sum_{G\in\mathcal{G}_n^*} P_n^\theta(G) \lg \frac{1}{P_n^\theta(G)} \\ &\geq \sum_{G\in\mathcal{G}_n^*} P_n^\theta(G) \lg \frac{1}{P^*} \geq P^+ \cdot \lg \frac{1}{P^*}.\end{aligned}$$

■

Thus, to obtain lower bounds on $\mathrm{H}(P_n^\theta)$, we will upper bound $\max_{G\in\mathcal{G}_n^*} P_n^\theta(G)$ by P^* and lower bound $\sum_{G\in\mathcal{G}_n^*} P_n^\theta(G)$ by P^+, for a suitably chosen $\mathcal{G}_n^* \subseteq \mathcal{G}_n$. For good lower bounds on $\mathrm{H}(P_n^\theta)$, $\mathcal{G}_n^*$ has to be chosen judiciously. For instance, choosing a large $\mathcal{G}_n^*$ (say, $\mathcal{G}_n$) might only yield a P^* that is moderately small, while at the same time, it is important to choose a $\mathcal{G}_n^*$ such that P^+ is large.

Let $\mathcal{H}_n$ denote the set of all undirected unlabeled graphs on n nodes. Let $\varphi : \mathcal{G}_n \to \mathcal{H}_n$ be the many-to-one map that discards node and edge labels and edge orientations. For a given model θ, let $Q_n^\theta : \mathcal{H}_n \to [0,1]$ be the probability distribution such that $Q_n^\theta(H) = \sum_{\varphi(G)=H} P_n^\theta(G)$. Clearly, $\mathrm{H}(Q_n^\theta) \leq \mathrm{H}(P_n^\theta)$ and therefore, lower bounds on $\mathrm{H}(Q_n^\theta)$ are stronger and harder to obtain.

B. Incompressibility of the PA model

Consider the PA model ($\mathsf{pref}[\mathsf{k}]$) defined in [7]. This model is parametrized by an integer $k \geq 1$. At time 1, the (undirected) graph consists of a single node x_1 with 1 self-loop. At time $t > 1$,

(i) a new node x_t, labeled t, is added to the graph;

(ii) a random node y is chosen from the graph with probability proportional to its current degree (in this phase, the degree of x_t is taken to be 1);

(iii) the edge $x_t \to y$, labeled $t \mod k$, is added to the graph;[2] and

(iv) if t is a multiple of k, nodes $t-k+1,\ldots,t$ are merged together, preserving self-loops and multi-edges.

For $k=1$, note that the graphs generated by the above model are forests. Since there are $2^{O(n)}$ unlabeled forests on n nodes (e.g., [29]), we have $\mathrm{H}(Q_n^{\mathsf{pref[k]}}) = O(n)$, i.e., the graph without labels and edge orientations is compressible to $O(1)$ bits per edge. The more interesting case is when $k \geq 2$ for which we show an incompressibility bound.

We underscore the importance of a good choice of $\mathcal{G}_n^*$ in applying Lemma 3. Consider the graph G having the first node of degree $k(n+1)$ and the other $n-1$ nodes of degree k. Clearly, $P_n^{\mathsf{pref[k]}}(G) = \prod_{i=k+1}^{nk} \frac{k-1+i}{2i-1} \geq 2^{-nk}$. Thus, choosing a set $\mathcal{G}_n^*$ containing G, would force us to have $P^* \geq 2^{-nk}$ so that the entropy bound given by Lemma 3 would only be $H(P_n^{\mathsf{pref[k]}}) \geq nk = \Theta(n)$. (A similar issue would be encountered in the unlabeled case as well.) A careful choice of $\mathcal{G}_n^*$, however, yields a better lower bound.

Theorem 4. $\mathrm{H}(Q_n^{\mathsf{pref[k]}}) = \Omega(n \log n)$*, for* $k \geq 2$.

Proof: Let G be a graph generated by $\mathsf{pref[k]}$. Let $\deg_t(x_i)$, for $i \leq t$, be the degree of the i-th inserted node at time t in G. By [14, Lemma 6], with probability $1-O(n^{-3})$, for each $1 \leq t \leq n$, each node $x_i, 1 \leq i \leq t$, will have degree $\deg_t(x_i) < (\sqrt{t/i}) \ln^3 n$ in G.

In particular, let $t^* = \lceil \sqrt[3]{n} \rceil$. Let ξ be the event: "$\exists t \geq t^*, \sum_{i=1}^{t^*} \deg_t(x_i) \geq n^{3/4}$." At time n, the sum of the degrees of nodes $x_1,\ldots,x_{t^*}$ can be upper bounded by

$$\sum_{i=1}^{t^*} \deg_n(x_i) \leq \sum_{i=1}^{t^*} \sqrt{\frac{n}{i}} \ln^3 n$$
$$= \sqrt{n} \ln^3 n \sum_{i=1}^{t^*} i^{-1/2} < O(n^{3/4}),$$

w.h.p. Indeed, $\Pr[\xi] \leq O(n^{-3})$.

Now define $t^+ = \lceil \epsilon n \rceil$, for some small enough $\epsilon > 0$; let n be large enough such that $t^* < t^+$. We call a node added after time t^+ *good* if it is not connected to any of the first t^* nodes. To bound the number of good nodes from below, we condition on ξ, and we upper bound the number of bad nodes. Using a union bound, the probability that node x_t for $t \geq t^*$ is bad can be upper bounded by $kn^{3/4}/(\epsilon n) \leq O(n^{-1/4})$.

[2]In the original PA model, edges are both undirected and unlabeled: we direct and label them for simplicity of exposition. The entropy lower bound will hold for the undirected and unlabeled version of these graphs.

Let ξ' be the event: "at least $(1-2\epsilon)n$ nodes are good"; by stochastic dominance, the event ξ' happens w.h.p. In our application of Lemma 3, we will choose $\mathcal{G}_n^* \subseteq \mathcal{G}_n$ to be the set of graphs satisfying $\xi \cap \xi'$. Thus, $P^+ = \Pr[\xi \cap \xi'] = 1 - o(1)$. Moreover,

$$\max_{G \in \mathcal{G}_n^*} P_n^{\mathsf{pref[k]}}(G) \leq \left(\frac{\sqrt{\frac{n}{\sqrt[3]{n}}} \ln^3 n}{kn} \right)^{(1-2\epsilon)kn}$$
$$\leq \left(O(n^{-2/3+\epsilon}) \right)^{2(1-2\epsilon)n} \leq n^{-\frac{4}{3}n + \frac{14}{3}\epsilon n} = \rho.$$

Notice how, by applying Lemma 3 at this point, we already have that $\mathrm{H}(P_n^{\mathsf{pref[k]}}) \geq \Omega(n \log n)$.

Now, we proceed to lower bound $\mathrm{H}(Q_n^{\mathsf{pref[k]}})$ through an upper bound on $|\varphi^{-1}(H)|$ for $H \in \mathcal{H}_n^{\mathsf{pref[k]}}$, by a careful counting argument. Given a H, it is possible to determine for each of its edges, which of the two endpoints of the edge was responsible for adding the edge to the graph. This task is trivial for edges incident to any node of degree k, as that node will have necessarily added all k edges to the graph. So, we can remove all degree k nodes from the graph and repeat this process until the graph becomes empty.

Thus, H could have been produced from at most $n! \cdot (k!)^n$ labeled graphs, since there are at most $n!$ ways of labeling the nodes, and $k!$ ways of labeling each of the "outgoing" edges of each node. That is, $\left|\varphi^{-1}(H)\right| \leq n! \cdot (k!)^n \leq n^n k^{kn}$. Then, choosing $\mathcal{H}_n^* \subseteq \mathcal{H}_n$ to be the set of unlabeled graphs obtained by removing labels from $\mathcal{G}_n^*$, $\mathcal{H}_n^* = \{\varphi(G) \mid G \in \mathcal{G}_n^*\}$, we obtain $P^+ = 1 - o(1)$, and

$$\max_{H \in \mathcal{H}_n^*} Q_n^{\mathsf{pref[k]}}(H) \leq \rho n^n k^{kn} = n^{-\Omega(n)} k^{kn} = P^*.$$

Finally, an application of Lemma 3 gives $\mathrm{H}(Q_n^{\mathsf{pref[k]}}) \geq P^+ \cdot \lg \frac{1}{P^*} \geq \Omega(n \log n)$, completing the proof. ∎

C. Incompressibility of other graph models

We now state the incompressibility results for other well-known graph models. Due to lack of space, the definitions of the models along with the proofs of the following results are omitted in this version.

Theorem 5. $\mathrm{H}(Q_n'^{\mathsf{acl}[\alpha]}) = \Omega(n \log n)$*, for*[3] $\alpha > 1/2$.

Theorem 6. $\mathrm{H}(Q_n^{\mathsf{copy}[\alpha,\mathsf{k}]}) = \Omega(n \log n)$*, for* $k > 2/\alpha$.

Theorem 7. *Let* $\ell \geq 2$ *and* $1/\ell < \alpha < 1$*. Then, w.h.p.,* $\mathrm{H}(Q_n^{\mathsf{krm}[\mathsf{M},\mathsf{s}]}) = \Omega(m \log n)$*, where* $n = \ell^s$*,* $M = \alpha J_\ell$*, and* m *is the number of edges.*

Theorem 8. $\mathrm{H}(Q_n^{\mathsf{kl}}) = \Omega(n \log n)$.

[3]Here we do not use the probability distribution Q on the graphs of n nodes — in the $\mathsf{acl}[\alpha]$ model the number of nodes is a r.v. . $Q_n'^{\mathsf{acl}[\alpha]}$ denotes the probability distribution on the graphs that can be generated by the $\mathsf{acl}[\alpha]$ model in n steps.

IV. The new web graph model

In this section we present our new web graph model. Let $k \geq 2$ be a fixed positive integer. Our new model creates a directed simple graph (i.e., no self-loops or multi-edges) by the following process.

The process starts at time t_0 with a simple directed *seed graph* G_{t_0} whose nodes are arranged on a (discrete) line, or list. The graph G_{t_0} has t_0 nodes, each of out-degree k. Here, G_{t_0} could be, for instance, a complete directed graph with $t_0 = k + 1$ nodes.

At time $t > t_0$, an existing node y is chosen uniformly at random (u.a.r.) as a prototype:

(i) a new node x is placed to the immediate left of y (so that y, and all the nodes on its right, are shifted one position right in the ordering),

(ii) a directed edge $x \to y$ is added to the graph, and

(iii) $k - 1$ edges are "copied" from y, i.e., $k - 1$ successors (i.e., out-neighbors) of y, say $z_1, \ldots, z_{k-1}$, are chosen u.a.r. without replacement and the directed edges $x \to z_1, \ldots, x \to z_{k-1}$ are added to the graph.

An intuitive explanation of this process is as follows. Consider the list of webpages ordered lexicographically by their URLs (for this ordering, a url a.b.com/d/e is to be interpreted as com/b/a/d/e.) A website owner might decide to add a new webpage to her site; to do this, she could take one of the existing webpages from her site as a prototype, modify it as needed, add an edge to the prototype for reference, and publish the new page on her site. Thus the new webpage and the prototype will be close in the URL ordering. See Figure 1 for an illustration of the model.

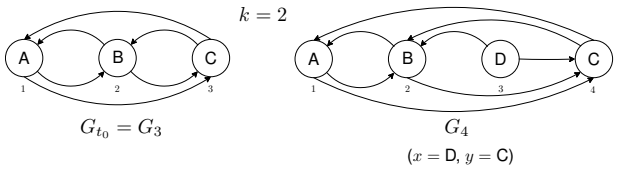

Fig. 1. The new node $x = D$ chooses $y = C$ as its prototype. The edge $C \to B$ is copied and the new edge $D \to C$ is added for reference. Notice that all the edges incident to C in $G_{t_0} = G_3$ increase their length by 1 in $G_{t_0+1} = G_4$.

In our model, we can show the following:

(i) The fraction of nodes of in-degree i is asymptotic to $\Theta(i^{-2-\frac{1}{k-1}})$; this power law is often referred to as "the rich get richer."

(ii) The fraction of edges of length[4] ℓ in the given embedding is asymptotic to $\Theta(\ell^{-1-\frac{1}{k}})$; analogously, we refer to this as "the long get longer."

[4]The *length* of an edge $x \to y$ is the absolute difference between the positions of node x and y in the given embedding.

Boldi and Vigna [5] study the distribution of *gaps* in the web graph, defined as follows. Sort the webpages lexicographically by URLs and this gives an embedding of nodes on the line. Now, if a webpage $x = z_0$ has edges to $z_1, \ldots, z_j$ in this order, the gaps are given by $|z_{i-1} - z_i|$, $1 \leq i \leq j$. They observe how the gap distribution in real web graph snapshots follows a power law with exponent ≈ 1.3. Our model can capture a similar distribution for the edge *lengths*, by an appropriate choice of k. In fact, both the average edge length and the average gap in our model are small; intuitively, though not immediately, this leads to the compressibility result of Section VII. It turns out that a power law distribution of either the lengths or the gaps (with exponent > 1) is sufficient to show compressibility; for sake of simplicity, we focus on the former in Section VI.

V. The rich get richer

In this section we characterize the in-degree distribution of our graph model. We show that the expected in-degree distribution follows a power law. We then show the distribution is tightly concentrated.

Let

$$f(i) = \frac{k \cdot 2^{1+\frac{2}{k-1}}\,\Gamma\left(\frac{3}{2} + \frac{1}{k-1}\right)}{(k-1)\sqrt{\pi}} \cdot \frac{\Gamma\left(i + 1 + \frac{1}{k-1}\right)}{\Gamma\left(i + 3 + \frac{2}{k-1}\right)}.$$

It is easy to show that $\lim_{i\to\infty} f(i) \Big/ \left(\frac{k\cdot 2^{1+\frac{2}{k-1}}\,\Gamma\left(\frac{3}{2}+\frac{1}{k-1}\right)}{(k-1)\sqrt{\pi}} \cdot i^{-2-\frac{1}{k-1}}\right) = 1$, i.e., $f(i) = \Theta(i^{-2-\frac{1}{k-1}})$. Let X_i^t denote the number of nodes of in-degree i at time t. We first show that $\mathrm{E}\left[X_i^t\right]$ can be bounded by $f(i) \cdot t \pm c$, for some constant c.

Theorem 9. *There is a constant $c = c(G_{t_0})$ such that*

$$f(i) \cdot t - c \leq \mathrm{E}\left[X_i^t\right] \leq f(i) \cdot t + c, \qquad (1)$$

for all $t \geq t_0$ and $i \in [t]$.

Proof: For now, assume $t > t_0$. Let x be the new node, and let y be the node we will copy edges from; recall that y is chosen u.a.r. First, we focus on the case $i = 0$. We have

$$\mathrm{E}\left[X_0^t \mid X_0^{t-1}\right] = X_0^{t-1} - \Pr\left[y \text{ had in-degree } 0\right] + 1,$$

as at each time step a new node (i.e., x) of in-degree 0 is added, and the only node that could change its in-degree to 1 is y. The probability of the latter event is exactly $X_0^{t-1}/(t-1)$. By the linearity of expectation, we get

$$\mathrm{E}\left[X_0^t\right] = \left(1 - \frac{1}{t-1}\right)\mathrm{E}\left[X_0^{t-1}\right] + 1. \qquad (2)$$

Next, consider $i \geq 1$. According to our model, nodes $z_1, \ldots, z_{k-1}$, will be chosen without replacement from

$\Gamma(y)$, the successors of y. The successors of the new node x will then be $\Gamma(x) = \{y, z_1, \ldots, z_{k-1}\}$. Since $z_1, \ldots, z_{k-1}$ are all distinct, the graph remains simple and $|\Gamma(x)| = k$.

For each $j = 1, \ldots, k-1$, the node z_j is chosen with probability proportional to its in-degree; this follows since node z_j was the endpoint of an edge chosen u.a.r. The probability that a particular node of in-degree $i \geq 1$ gets chosen as a successor is $\frac{1}{t-1} + \frac{i(k-1)}{k(t-1)}$ (recall that all the k successors of x will be distinct). Thus, for $i \geq 1$,

$$\mathrm{E}\left[X_i^t\right] = \left(1 - \frac{1}{t-1} - \frac{i}{t-1}\frac{k-1}{k}\right)\mathrm{E}\left[X_i^{t-1}\right] + \left(\frac{1}{t-1} + \frac{i-1}{t-1}\frac{k-1}{k}\right)\mathrm{E}\left[X_{i-1}^{t-1}\right]. \quad (3)$$

For the base cases, note that $X_t^t = 0$ for each $t \geq t_0$. Also, the variables $X_i^{t_0}$ are completely determined by G_{t_0}. For each fixed k, we have $f(t) = \Theta(t^{-2-\frac{1}{k-1}})$. Thus, there is a constant c_0 such that for any $c \geq c_0$, and for all $t \geq t_0$, $\mathrm{E}[X_t^t]$ follows (1). The base cases $\mathrm{E}\left[X_i^{t_0}\right]$, $i = 1, 2, \ldots$, can also be covered with a sufficiently large c (that has to be greater than some function of the initial graph G_{t_0}).

For the inductive case, we have $f(0) = \frac{1}{2}$ (by applying $\Gamma(x)\Gamma(x+\frac{1}{2}) = \Gamma(2x)\,2^{1-2x}\sqrt{\pi}$, and $\Gamma(2x+1) = 2x\,\Gamma(2x)$, with $x = 1 + \frac{1}{k-1}$). Using this, (2), and simple calculations, we can show that if X_0^{t-1} satisfies (1), then X_0^t also satisfies (1). For $i \geq 1$, we have $f(i-1) = f(i) \cdot (ik - i + 2k + 2)/(ik - i + 1)$. An easy induction on (3) completes the proof. ∎

Thus, in expectation, the in-degrees follow a power law with exponent $-2 - 1/(k-1)$. We prove a $O(1)$-Lipschitz property for the r.v.'s X_i^t, if $k = O(1)$. The concentration immediately follows from Theorem 1.

Lemma 10. *Each r.v. X_i^t satisfies the $(2k)$-Lipschitz property.*

Proof: Our model can be interpreted as the following stochastic process: at step t, two independent dice, with $t-1$ and k faces respectively, are thrown. Let Q_t and R_t be the respective outcomes of these two trials. The new node x will position itself to the immediate left of the node y that was added at time Q_t. Suppose that the (ordered) list of successors of y is $(z_1, \ldots, z_k)$. The ordered list of successors of x will be composed of y followed by the nodes $z_1, \ldots, z_k$ with the exception of node z_{R_t}. Thus, the number of nodes X_i^τ of in-degree i at time τ can be interpreted as a function of the trials $(Q_1, R_1), \ldots, (Q_\tau, R_\tau)$.

We want to show that changing the outcome of any single trial $(Q_{t'}, R_{t'})$, changes the r.v. X_i^τ (for fixed i) by an amount not greater than $2k$. Suppose we change $(q_{t'}, r_{t'})$ to $(q'_{t'}, r'_{t'})$, going from graph G to G'. Let x be the node added at time t' with the choice $(q_{t'}, r_{t'})$, and x' be the node added with the choice $(q'_{t'}, r'_{t'})$.

Let S, S' be the successors of x in G and x' in G', respectively. We complete complete the proof by showing inductively that at any time step t, and for any nodes y, y' added at the same time respectively in G, G', the (ordered) list of successors of y and y' are *close*, i.e., in each of their positions, they either have the same successor, or they have two different elements of $S \cup S'$.

If $t \leq t'$, then the proof is immediate. For $t > t'$, it is easy to see that the only edges we need to consider are the copied edges. By induction, we know that at time $t-1$, the lists of successors of the node we are copying from, in G and G', were close. Since the two lists are sorted, either the i-th copied edges in G and G' will either be the same or will both point to nodes in $S \cup S'$. Thus the lists of the time t node are close and the proof is complete. ∎

VI. The long get longer

In this section we analyze the edge length distribution in our graph model. We show it follows a power law with exponent larger than 1. Later, we will use this to establish the compressibility of graphs generated by our model. Let

$$g(\ell) = \frac{\Gamma\left(\ell + 1 - \frac{1}{k}\right)}{\Gamma\left(2 - \frac{1}{k}\right)\Gamma\left(\ell + 2\right)}.$$

It holds that $\lim_{\ell\to\infty} g(\ell) \Big/ \left(\ell^{-1-\frac{1}{k}} / \Gamma\left(2 - \frac{1}{k}\right)\right) = 1$, i.e., $g(\ell) = \Theta(\ell^{-1-\frac{1}{k}})$. Recall that the *length* of an edge from a node in position i to a node in position j is equal to $|i - j|$; we define its *circular directed length*, denoted *cd-length*, to be $j - i$ if $j > i$, and $t - (i - j)$ otherwise. Let Y_ℓ^t be the number of edges of length ℓ at time t. We aim to show that $Y_\ell^t \approx g(\ell) \cdot t$. It turns out to be useful to consider a related r.v. Z_ℓ^t, which denotes the number of edges of cd-length ℓ at time t. We will first show that, w.h.p., $Z_\ell^t \approx g(\ell) \cdot t$. We will then argue that Y_ℓ^t is very close to Z_ℓ^t.

The following shows that $\mathrm{E}[Z_\ell^t]$ is bounded by $g(\ell) \cdot t \pm O(1)$.

Theorem 11. *There exists some constant $c = c(G_{t_0})$ such that*

$$g(\ell) \cdot t - c \leq \mathrm{E}\left[Z_\ell^t\right] \leq g(\ell) \cdot t + c,$$

for all $t \geq t_0$ and $\ell \in [t]$.

Proof: As in the proof of Theorem 9, we start by obtaining a recurrence on the r.v.'s Z_i^t. Let x be the node added at time t, and let y, y' be the nodes to the immediate right and left of x respectively (where y' equals the last node in the ordering if x is placed before the first node y).

Consider Z_1^t. For $t > t_0$,

$$\mathrm{E}\left[Z_1^t \mid Z_1^{t-1}\right] = Z_1^{t-1} - \Pr\left[x \text{ enlarges an edge of cd-length } 1\right] + 1,$$

as an edge $x \to y$ of length 1 is necessarily added to the graph, and adding x can enlarge at most one edge of cd-length 1 (that is, the edge $y' \to y$ if it exists). The probability of the latter event is equal to $Z_1^{t-1}/(t-1)$. By the linearity of expectation,

$$\mathrm{E}\left[Z_1^t\right] = \left(1 - \frac{1}{t-1}\right)\mathrm{E}\left[Z_1^{t-1}\right] + 1.$$

Now consider Z_ℓ^t, for $\ell \geq 2$ and $t > t_0$. We have,

$$\mathrm{E}\left[Z_\ell^t \mid Z_\ell^{t-1}, Z_{\ell-1}^{t-1}\right] = Z_\ell^{t-1} - N_1 + N_2 + N_3,$$

where $N_1 = \mathrm{E}[$ # edges of cd-length ℓ that x enlarged $\mid Z_\ell^{t-1}, Z_{\ell-1}^{t-1}]$, $N_2 = \mathrm{E}[$ # edges of cd-length $(\ell-1)$ that x enlarged $\mid Z_\ell^{t-1}, Z_{\ell-1}^{t-1}]$, and $N_3 = \mathrm{E}[$ # edges of cd-length $(\ell-1)$ that x copied from $y \mid Z_\ell^{t-1}, Z_{\ell-1}^{t-1}]$. Recall that x is placed to the left of a node y chosen u.a.r. Thus, given a fixed edge of length ℓ, the probability this edge is enlarged by x is $\ell/(t-1)$. Thus,

$$N_1 = \frac{\ell}{t-1} Z_\ell^{t-1},$$

$$N_2 = \frac{\ell-1}{t-1} Z_{\ell-1}^{t-1},$$

$$N_3 = \sum_{j=1}^{k-1} \Pr[j\text{th copied edge had cd-length } (\ell-1) \mid Z_\ell^{t-1}, Z_{\ell-1}^{t-1}].$$

Note that, for each $j = 1, \ldots, k-1$, the jth copied edge is chosen uniformly at random over all the edges (even if the $k-1$ copied edges are not independent). Thus,

$$N_3 = \frac{(k-1) Z_{\ell-1}^{t-1}}{k(t-1)}.$$

By the linearity of expectation, we get for $\ell \geq 2$,

$$\mathrm{E}\left[Z_\ell^t\right] = \left(1 - \frac{\ell}{t-1}\right)\mathrm{E}\left[Z_\ell^{t-1}\right] + \left(\frac{\ell-1}{t-1} + \frac{1}{t-1}\frac{k-1}{k}\right)\mathrm{E}\left[Z_{\ell-1}^{t-1}\right].$$

The base cases can be handled as in Theorem 9. The inductive step for $\ell = 1$ can be easily shown. For $\ell \geq 2$, it suffices to note that $g(\ell-1) = k\frac{\ell+1}{\ell k - 1} g(\ell)$. ∎

Thus, the expectation of the edge lengths follows a power law with exponent $-1 - 1/k$.

To establish the concentration result, we need to analyze quite closely the combinatorial structure of the graphs generated by our model. Recall that the nodes in our graphs are placed contiguously on a discrete line (or list). At a generic time step, we use x_i to refer to the ith node in the ordering from left to right. Given an ordering $\pi = (x_1, x_2, \ldots, x_t)$ of the nodes, and an integer $0 \leq k < t$, a *k-rotation*, $\rho_k(x_i)$ maps the generic node x_i, $1 \leq i \leq t$, to position $1 + ((i+k) \mod t)$.

We say that two nodes x, x' are *consecutive* if there exists a k such that $|\rho_k(x) - \rho_k(x')| = 1$, i.e., they are consecutive if in the ordering either they are adjacent or one is the first and the other the last. Further, we say that an edge $x'' \to x'''$ *passes over* an node x if there exists k such that $\rho_k(x'') < \rho_k(x) < \rho_k(x''')$. Finally, two edges $x \to x'$ and $x'' \to x'''$ are said to *cross* if there exists a k such that after a k-rotation exactly one of x and x' is within the positions $\rho_k(x'')$ and $\rho_k(x''')$. We prove the following characterization that will be used later in the analysis.

Lemma 12. *At any time, given any two consecutive nodes x, x', and any positive integer ℓ, the number of edges of cd-length ℓ that pass over x or x' (or both) is at most $C = (k+2)t_0 + 1$.*

Proof: Let us define G_t^- as the graph G_t minus the edges incident to the nodes that were originally in G_{t_0}. Note that, for each cd-length ℓ, the number of the edges of cd-length ℓ that we remove is upper-bounded by $2t_0$ as each node can be incident to at most two edges of cd-length ℓ, one going in, and one going out of the node. Unless otherwise noted, we will consider G_t^- for the rest of the proof.

Fix the time t, and take any rotation ρ; let $x_1, \ldots, x_t$ be the nodes in the list in the left-right order given by the rotation (i.e., node x_i is in position i according to ρ). For a set of edges of the same cd-length to pass over at least one of two consecutive nodes x, x' it is necessary for every pair of them to cross. We will bound, for a generic edge e, the number of edges that cross e and have the same length as e. Let $t(x_a)$ be the time when x_a was added to the graph. First, by definition we have that if $x_a \to x_b$, then $t(x_a) > t(x_b)$.

Second, we claim that if there exists a rotation ρ' such that x_a, x_b, x_c are three nodes with $\rho'(x_a) < \rho'(x_b) < \rho'(x_c)$ and $t(x_c) > t(x_b)$, then the edge $x_a \to x_c$ cannot exist. To see this, for $x_a \to x_c$ to exist it must be that $t(x_a) > t(x_c)$. We want to show inductively that all the nodes that will point to x_c will be both to the left of x_c and to the right of x_b, in the ordering implied by ρ'. Note that x_c was not in G_{t_0} since its insertion time is larger than that of x_b. Thus, each node placed to the immediate left of x_c will point to it, and will trivially satisfy the induction hypothesis. Furthermore, each node that copies an edge to x_c must be placed to the immediate left of a node pointing to x_c. Thus, the second claim is proved.

Third, we claim that if x_a, x_b, x_c, x_d are four nodes such that the edges $x_a \to x_c$ and $x_b \to x_d$ exist, and

cross each other, then there exists an edge $x_c \to x_d$. To see this, first note that none of these four nodes could have been part of G_{t_0}, for otherwise at least one of the two edges could not have been part of G_t^-. Fix a rotation ρ'' s.t. $\rho''(x_a) < \rho''(x_b) < \rho''(x_c)$; by the second claim, it must be that $t(x_b) > t(x_c)$. Thus, the edge $x_b \to x_d$ has necessarily been copied from some node, say x_{b_1}. Note that $\rho''(x_{b_1}) \leq \rho(x_c)$. Indeed by assumption $\rho''(x_c) > \rho''(x_b)$ and it is impossible that $\rho''(x_c) < \rho''(x_{b_1})$, for otherwise x_b could not have copied from x_{b_1} as $t(x_b) > t(x_c)$. Now, we know that the edge $x_{b_1} \to x_d$ exists (as before, x_{b_1} is not part of G_{t_0}). If $x_{b_1} = x_c$, then we are done. Otherwise, there must exist an x_{b_2} pointing to x_d from which x_{b_1} has copied the edge. Note that $\rho''(x_{b_1}) < \rho''(x_{b_2}) \leq \rho''(x_c)$. By iterating this reasoning, the claim follows.

Take any set S of edges having the same length, and such that any pair of them *cross*. Given an arbitrary ρ''', let x be the node with the smallest $\rho'''(x)$ such that, for some x', the edge $x \to x'$ is in S (the nodes x and x' are unique). For any other edge $y \to y'$ in S, by the third claim, there must exist the edge $x' \to y'$. As x' has out-degree k, it follows that $|S| \leq k+1$.

Finally, since the seed graph G_{t_0} had kt_0 edges and we removed at most $2t_0$ edges of cd-length ℓ (for an arbitrary $\ell \geq 1$) in the cut $[G_{t_0}, G_t \setminus G_{t_0}]$, we have refrained from counting at most kt_0+2t_0 edges of length ℓ passing over one of the nodes x, x'. The proof follows. ∎

Now we prove the $O(1)$-Lipschitz property of the r.v.'s Z_ℓ^t, if $t_0, k = O(1)$. The concentration of the Z_ℓ^t will follow immediately from Theorem 1.

Lemma 13. *Each r.v. Z_ℓ^t satisfies the $((k+2)t_0+k+1)$-Lipschitz property.*

Proof: We use the stochastic interpretation as in the proof of Lemma 10. For each τ, let Z_ℓ^τ be the r.v. representing the number of edges of cd-length ℓ at time τ. We consider Y_ℓ^τ as a function of the trials (Q_1, R_1), …, (Q_τ, R_τ). We show that changing the outcome of any single trial $(Q_{t'}, R_{t'})$, changes the r.v. Z_ℓ^τ, for fixed ℓ, by an amount not greater than $C+k = (k+2)t_0+k+1$.

Suppose we change $(q_{t'}, r_{t'})$ to $(q'_{t'}, r'_{t'})$, going from graph G to G'. Let x be the node added at time t' with the choice $(q_{t'}, r_{t'})$, and x' be its equivalent with the choice $(q'_{t'}, r'_{t'})$. We show that choosing two different positions for x and x' can change the number of edges of cd-length ℓ by at most $C+k$ at any time step. Note that before time step t', the cd-lengths are all equal.

By Lemma 12, at time $t > t'$, for all ℓ, the number of edges of cd-length ℓ that pass over x (resp., x') is upper bounded by C. For an edge e, let S_e be the set of edges that have been copied from e, directly or indirectly, including e itself, i.e., $e \in S_e$ and if an edge e' is copied from some edge in S_e, then $e' \in S_e$. It is easy to note that no two edges in S_e have the same cd-length, since they all start from different nodes, but end up at the same node.

For any node z, if $e_1, \ldots, e_k$ are the successors of z, we define $S_z = S_{e_1} \cup \cdots \cup S_{e_k}$. The last observation implies that, for any fixed ℓ, no more than k edges of cd-length ℓ are in S_v (or $S_{v'}$) at any single time step. Now, consider the following edge bijection from G to G': the ith edge of the jth inserted node in G is mapped to the ith edge of the jth inserted node in G'. It is easy to see that if an edge e in G (resp., G') does not pass over x (resp., x') and is not in S_x (resp., $S_{x'}$), then e gets mapped to an edge of the same cd-length in G' (resp., G). Thus, the difference in the number of edges of the cd-length ℓ in G and G' is at most $C+k$. ∎

We now show that the number D_t of edges whose length and cd-length are different (at time t) is very small. Since the maximum absolute difference between Y_ℓ^t and Z_ℓ^t is bounded by D_t, this will show that these r.v.'s are close to each other. First note that if an edge $x_i \to x_j$ has different length and cd-length, then $j < i$; call such an edge *left-directed* and let R_t be the set of left-directed edges. Since $D_t \leq R_t$, it suffices to bound the latter.

Lemma 14. *With probability $1 - O\left(\frac{1}{t}\right)$, $R_t \leq O\left(t^{1-\frac{1}{k}+\epsilon}\right)$, for each constant $\epsilon > 0$.*

Proof: Observe that each edge $x_i \to x_j$ counted by R_t is such that $j < i$. Thus, R_{t_0} is equal to the number of left-directed edges in G_{t_0} with its given embedding.

Further, R_t's increase over R_{t-1} equals the number of left-directed edges copied at step t (the proximity edge is always not left-directed).

Thus, $E[R_t|R_{t-1}] = \left(1 + (k-1)\cdot\frac{1}{k(t-1)}\right)\cdot R_{t-1}$ and $E[R_t] = \left(1 + (k-1)\cdot\frac{1}{k(t-1)}\right)\cdot E[R_{t-1}]$, for each $t > t_0$. Therefore,

$$\begin{aligned} E[R_t] &= R_{t_0}\cdot\prod_{i=t_0+1}^{t}\left(1+\frac{k-1}{k}\cdot\frac{1}{i}\right) \\ &= R_{t_0}\cdot\prod_{i=t_0+1}^{t}\frac{i+\frac{k-1}{k}}{i} \\ &= R_{t_0}\cdot\frac{\Gamma\left(t+\frac{k-1}{k}+1\right)\cdot\Gamma\left(t_0+1\right)}{\Gamma\left(t_0+\frac{k-1}{k}+1\right)\cdot\Gamma\left(t+1\right)}. \end{aligned}$$

Thus, $E[R_t] = \Theta\left(t^{1-\frac{1}{k}}\right)$. We note how a $O(1)$-Lipschitz condition holds (at most $k-1$ new left-directed edges can be added at each step). Thus, Theorem 1 can be applied with an error term of $O\left(\sqrt{t\log t}\right) \leq O\left(t^{\frac{1}{2}+\epsilon}\right) \leq O\left(t^{1-\frac{1}{k}+\epsilon}\right)$. The result follows. ∎

Applying Theorem 1, Theorem 11, Lemma 13, and Lemma 14, we obtain the following.

Corollary 15. *With probability* $\geq 1 - O\left(\frac{1}{t^2}\right)$,

i. $|Z_\ell^t - E[Z_\ell^t]| \leq O\left(\sqrt{t \log t}\right)$ *and*
ii. $|Y_\ell^t - E[Z_\ell^t]| \leq O\left(t^{1-1/k+\epsilon}\right)$.

Note that the concentration error term, $O(\sqrt{t \log t})$, is upper bounded by R_t, for each $k \geq 2$. Also, the corollary is vacuous if $\ell > t^{1/(k+2)}$.

VII. Compressibility of our model

We now analyze the number of bits needed to compress the graphs generated by our model. Recall that the web graph has a natural embedding on the line via the URL ordering that experimentally gives very good compression [5], [6]. Our model generates a web-like random graphs and an embedding "à-la-URL" on the line. We work with the following *BV-like compression scheme*: a node at position p on the line stores its list of successors at positions $p_1, \ldots, p_k$ as a list $(p_1 - p, \ldots, p_k - p)$ of compressed integers. An integer $i \neq 0$ will be compressed using $O(\log(|i| + 1))$ bits, using Elias γ-code, for instance [35]. We show that our graphs can be compressed using $O(1)$ bits per edge using above scheme.

Theorem 16. *The above BV-like scheme compresses the graphs generated by our model using* $O(n)$ *bits, with probability at least* $1 - O\left(\frac{1}{n}\right)$.

Proof: Let $\epsilon > 0$ be a small constant. At time n, consider the number of edges of length at most $L = \lceil n^\epsilon \rceil$. Note that by Corollary 15, for each $1 \leq \ell \leq L$, it holds that $|Y_\ell^n - E[Z_\ell^n]| \leq O\left(n^{1-1/k+\epsilon}\right)$, with probability $1 - O\left(n^{-1}\right)$. For the rest of the proof, we implicitly condition on these events.

Lower bounding $E[Z_\ell^n]$ as in Theorem 11, we obtain the following lower bound on the number of edges of length $\leq L$, using standard algebraic manipulation and Lemma[5] 2

$$\begin{aligned} S &\geq \sum_{\ell=1}^{L} \left(\frac{\Gamma\left(\ell + 1 - \frac{1}{k}\right)}{\Gamma\left(2 - \frac{1}{k}\right)\Gamma(\ell+2)} \cdot n - c - O\left(n^{1-1/k+\epsilon}\right) \right) \\ &\geq nk \left(1 - \frac{\Gamma\left(L + 2 - \frac{1}{k}\right)}{\Gamma(L+2)\Gamma\left(2 - \frac{1}{k}\right)}\right) - O\left(Ln^{1-1/k+\epsilon}\right) \\ &\geq nk - O\left(nkL^{-1/k}\right) - O\left(Ln^{1-1/k+\epsilon}\right) \\ &\geq nk - O\left(n^{1-\epsilon_1}\right), \end{aligned}$$

where ϵ_1 is a small constant.

[5]Which we apply to the sum, to conclude that $\frac{1}{\Gamma\left(2-\frac{1}{k}\right)} \sum_{\ell=1}^{L} \frac{\Gamma\left(\ell+1-\frac{1}{k}\right)}{\Gamma(\ell+2)} = k \cdot \left(1 - \frac{\Gamma\left(L+2-\frac{1}{k}\right)}{\Gamma(L+2)\Gamma\left(2-\frac{1}{k}\right)}\right)$.

At time n, the total number of edges of the graph is nk. Thus the number of edges of length more than L is at most $O\left(n^{1-\epsilon_1}\right)$. (Notice how, for this argument to work, it is crucial to have a very strong bound on the behavior of the Y_ℓ^n random variables; this is why we used the Gamma function in their expressions.) The maximum edge length is $O(n)$ and so each edge can be compressed in $O(\log n)$ bits. The overall contribution, in terms of bits, of the edges longer than L will then be $o(n)$.

Now, we calculate the bit contribution B of the edges of length at most L.

$$\begin{aligned} B &\leq \sum_{\ell=1}^{L} (O(\log(\ell+1)) \left(\frac{\Gamma\left(\ell + 1 - \frac{1}{k}\right)}{\Gamma\left(2 + \frac{1}{k}\right)\Gamma(\ell+2)} n + c \right. \\ &\qquad \left. + O\left(n^{1-1/k+\epsilon}\right)\right)) \\ &\leq n \cdot O\left(\sum_{\ell=1}^{L} \ell^{-1-1/k} \log(\ell+1)\right) \\ &\qquad + O\left(Ln^{1-1/k+\epsilon} \log L\right) \\ &\leq O(n), \end{aligned}$$

where the penultimate inequality follows since the fraction involving the Gamma function can be upper bounded by $O(\ell^{-1-1/k})$, and the last inequality from $O\left(\ell^{-1-2\epsilon} \cdot \log \ell\right) \leq O\left(\ell^{-1-\epsilon}\right)$ and from the convergence of the Riemann series. The proof is complete. ■

Thus, given an ordering of nodes, we can compress the graph to use $O(1)$ bits per edge using a very simple linear-time algorithm. A natural question is if it is still possible to compress this graph *without* knowing the ordering. We show that this is still possible.

Theorem 17. *The graphs generated by our model can be compressed using* $O(n)$ *bits in linear time, even if ordering of the nodes is not available.*

Proof: Given a node v in G, just by looking at two-neighborhood, we can either (i) find an out-neighbor w of v having exactly $k-1$ out-neighbors in common with v, or (ii) we can conclude that v was part of the "seed" graph G_{t_0} (having constant order). This step takes time $O(k^2) = O(1)$.

Indeed, if v was not part of G_{t_0}, during its insertion, v added a proximity edge to its "real prototype" w, and copied $k-1$ of w's out-links. If more than one out-neighbor of v has $k-1$ out-neighbors in common with v, we choose one arbitrarily and we call it the "possible prototype" of v.

For compressing, we create an unlabeled rooted forest out of the nodes in G. A node v will look for a possible prototype w. If such a w is found, then v will choose w as its parent. Otherwise v will be a root in the forest.

To describe G, it will suffice to (i) describe the unlabeled rooted forest, (ii) describe the subgraph induced by the roots of the trees in the forest, and (iii) for each non-root node v in the forest, use $\lceil \lg k \rceil$ bits to describe which of its parent's out-neighbors was not copied by v in G. The forest can be described with $O(n)$ bits, for instance, by writing down the *down / up* steps made when visiting each tree in the forest. This requires $O(n)$ bits. The graph induced by the roots of the trees (i.e., a subgraph of G_{t_0}) can be stored in a non-compressed way using $O(t_0^2) = O(1)$ bits. The third part of the encoding will require at most $O(n \log k) = O(n)$ bits. Note that it is trivial to compute each of the three encodings in linear time. ■

REFERENCES

[1] M. Adler and M. Mitzenmacher. Towards compressing web graphs. In *Data Compression Conference*, pages 203–212, 2001.

[2] W. Aiello, F. Chung, and L. Lu. Random evolution in massive graphs. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 510–519, 2001.

[3] R. Albert and A.-L. Barabasi. Emergence of scaling in random networks. *Science*, pages 509–512, 1999.

[4] P. Boldi, M. Santini, and S. Vigna. Permuting web graphs. In *Proc. of WAW 2009*, 2009.

[5] P. Boldi and S. Vigna. The webgraph framework I: Compression techniques. In *Proc. 13th International World Wide Web Conference*, pages 595–601, 2004.

[6] P. Boldi and S. Vigna. Codes for the world-wide web. *Internet Mathematics*, 2(4):405–427, 2005.

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

[8] C. Borgs, J. T. Chayes, C. Daskalakis, and S. Roch. First to market is not everything: An analysis of preferential attachment with fitness. In *Proc. 39th Annual ACM Symposium on Theory of Computing*, pages 135–144, 2007.

[9] A. Z. Broder, S. R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. In *Proc. 9th International World Wide Web Conference*, pages 309–320, 2000.

[10] G. Buehrer and K. Chellapilla. A scalable pattern mining approach to web graph compression with communities. In *Proc. 1st International Conference on Web Search and Data Mining*, pages 95–106, 2008.

[11] J. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E*, 60:1412, 1999.

[12] F. Chierichetti, R. Kumar, M. Mitzenmacher, A. Panconesi, P. Raghavan, and S. Lattanzi. On compressing social networks, 2009. Manuscript.

[13] C. Cooper and A. M. Frieze. A general model of web graphs. *Random Structures and Algorithms*, 22(3):311–335, 2003.

[14] C. Cooper and A. M. Frieze. The cover time of the preferential attachment graph. *Journal of Combinatorial Theory, Ser. B*, 97(2):269–290, 2007.

[15] D.Dubhashi and A.Panconesi. Concentration of measure for the analysis of randomized algorithms, 1998. Draft available at: `http://www.dsi.uniroma1.it/~ale/papers.html`.

[16] M. Dodds and Watts. An experimental study of search in global social networks. *Science*, 301(5634):827–829, 2003.

[17] A. Fabrikant, E. Koutsoupias, and C. H. Papadimitriou. Heuristically optimized trade-offs: A new paradigm for power laws in the internet. In *Proc. 29th International Colloquium on Automata, Languages and Programming*, pages 110–122, 2002.

[18] C. Karande, K. Chellapilla, and R. Andersen. Speeding up algorithms on compressed web graphs. In *Proc. 2nd International Conference on Web Search and Data Mining*, pages 272–281, 2009.

[19] J. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[20] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proc. 37th Annual ACM Symposium on Theory of Computing*, pages 163–170, 2000.

[21] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal. Stochastic models for the web graph. In *Proc. 41st IEEE Symposium on Foundations of Computer Science*, pages 57–65, 2000.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cybercommunities. In *Proc. 8th International World Wide Web Conference*, pages 403–416, 1999.

[23] S. Lattanzi and D. Sivakumar. Affiliation networks. In *Proc. 41st ACM Symposium on Theory of Computing*, 2009.

[24] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 133–145, 2005.

[25] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters, and possible explanations. In *Proc. 11th Conference on Knowledge Discovery and Data Mining*, pages 177–187, 2005.

[26] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proc. National Academy of Sciences*, 102(33):11623–11628, 2005.

[27] B. Mandelbrot. An informational theory of the statistical structure of languages. In W. Jackson, editor, *Communication Theory*, pages 486–502. Butterworth, 1953.

[28] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2), 2003.

[29] L. Mutafchiev. The largest tree in certain models of random forests. *Random Structures and Algorithms*, 13(3-4):211–228, 1998.

[30] H. Simon. On a class of skew distribution functions. *Biometrika*, 42:425–440, 1955.

[31] T. Suel and J. Yuan. Compressing the graph structure of the web. In *Data Compression Conference*, pages 213–222, 2001.

[32] J. Szymanski. On the complexity of algorithms on recursive trees. *Theoretical Computer Science*, 74(3):355–361, 1990.

[33] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969.

[34] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–410, 1998.

[35] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kauffman Publishers, 2 edition, 1999.

[36] G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.