Adjacency labeling scheme for power-law graphs

Casper Petersen, Noy Rotbart,
Jakob Grue Simonsen and Christian Wulff-Nilsen

February 12, 2015

Abstract

An adjacency labeling scheme is a method that assigns labels to the vertices of a graph such that adjacency relation can be inferred directly from the assigned label, without using a centralized data structure. We devise adjacency labeling schemes for the family of power-law graphs, a family that has been used to model many types of networks, e.g. the Internet AS-level graph. Furthermore, we prove an almost matching lower bound for this family. We also provide an asymptotically optimal labeling scheme for sparse graphs. Finally, we validate the efficiency of our labeling scheme by an experimental evaluation using both synthetic data and real-world networks of up to hundreds of thousands of vertices.

1 Introduction

A fundamental problem in networks is how to disseminate the structural information of the underlying graph of a network to its vertices in such that the local structure of the network can be inferred using only local information. One way of doing so is via an *adjacency labeling scheme*: an algorithm that assigns a bit string–a *label*–to each vertex so that adjacency between any two vertices can be inferred solely from their respective labels. The main objective is to minimize the maximum label length, i.e., the maximum number of bits used in a label. Moon [37] showed a lower bound of $n/2$ on label size for general graphs, and a corresponding upper bound of $n/2 + 6$ was very recently shown by Alstrup et al. [6]. Upper bounds for adjacency labeling schemes exist for many specific classes of graphs, including trees [7], planar graphs [23], bounded-degree graphs [2], and bipartite graphs [35].

However, for classes of graphs whose statistical properties–in particular their *degree distribution*–more closely resembles that of real-world networks, there has been no research on adjacency labeling schemes. One class of graphs extensively used for modelling real-world networks is *power-law graphs*: roughly, graphs where the number of vertices of degree k is proportional to k^α for some positive α. Power-law graphs[1] have been used, e.g., to model the Internet AS-level graph [40, 4], and many other types of network (see, e.g., [36, 20] for overviews). Both the adequacy of fit of power-law graph models to actual data, as well as the empirical correctness of the conjectured mechanisms giving rise to power-law behaviour have been subject to criticism (see, e.g., [1, 20]). In spite of such criticism, and because their degree distribution affords a reasonable approximation of the degree distribution of many networks, the class of power-law graphs remains a popular tool in network modelling whose statistical behaviour is well-understood: e.g., for graphs with $2 < \alpha < 3$ it is known that with high probability the average distance between any two vertices is $O(\log \log n)$, the diameter is $O(\log n)$ and includes a dense subgraph of $n^{c/\log \log n}$ vertices [17].

Routing labeling schemes for power-law graphs have been investigated by Brady and Cowen [13], and by Chen et. al [16]. Labeling schemes for other properties than adjacency have been investigated for various classes of graphs, e.g., distance [24], and flow [29]. Dynamic labeling schemes were studied by Korman and Peleg [31, 32, 30] and recently by Dahlgaard et. al [21]. Experimental evaluation for some labeling schemes for various properties on general graphs habe been performed by Caminiti et. al [15], Fischer [22] and Rotbart et. al [39].

1.1 Our contribution

In detail, our contribution comprises of:

An $O(\sqrt[c]{n}(\log n)^{1-1/\alpha})$ adjacency labeling scheme for power-law graphs G. The scheme is based on two ideas: (I) a labeling *strategy* that partitions the vertices of G into high ("fat") and low degree ("thin") vertices based on a threshold degree, and (II) a threshold *prediction* that depends only on the coefficient α of a power-law curve fitted to the degree distribution of G. Real-world power-law graphs rarely exceed 10^{10} vertices, implying a label size of maximum 10^5 bits, well within the processing capabilities of current hardware. We claim that our scheme is thus appealing in practice due both to its simplicity and its size. Using the same ideas, we get an asymptotically tight $O(\sqrt{n})$ adjacency labeling scheme for sparse graphs.

A lower bound of $\Omega(\sqrt[c]{n})$ bits on the maximum label size for *any* adjacency labeling scheme for power-law graphs. To that end we define a highly restrictive subclass of power-law graphs and ... The lower bound shows that our upper bound above is asymptotically optimal, bar a $(\log n)^{1-1/\alpha}$ factor.

An experimental investigation of our labeling scheme Using both real-world (23K-325K vertices) and synthetic (300K-1M vertices) data sets, we observe that: (i) Our threshold *prediction* performs close to optimal when using the labeling *strategy* above. (ii) our labeling scheme achieves maximum label size several orders of magnitude smaller than the state-of-the-art labeling schemes for more general graph families.

[1] Power-law graphs are also called scale-free graphs in the literature surveyed.

Our study contributes to the graph-theory related question of induced universal graphs. Given graph family $\mathcal{F}$ the question is to find smallest N such that a graph of N vertices contains all graphs in $\mathcal{F}$ as induced subgraphs. Kannan, Naor and Rudich [28] showed that an $f(n)\log n$ adjacency labeling scheme for $\mathcal{F}$ constructs an induced universal graph for this family of $2^{f(n)}$ vertices. To the best of our knowledge, we are the first to provide an upper and lower bounds on induced universal graphs for power-law graphs.

Finally, our study may contribute to the understanding of the quality of *generative models*[2] for power-law graphs, such as the Barabasi-Albert model [11] and the Aiello-Chung-Lu model [3]. As a first step, we provide an evidence that the Barabasi-Albert model [11] produces only a small fraction of the power-law graphs possible.

2 Preliminaries

Throughout the paper, we consider n-vertex, undirected, finite graphs. For real $c > 0$, a graph is *c-sparse* if it has at most cn edges and *sparse* if it is c-sparse for some c. For $0 < c \leq n-1$, the set of c-sparse graphs with n vertices is denoted by $\mathcal{S}_{c,n}$. If $\mathcal{F}$ is a set of graphs, $\mathcal{F}_n$ denotes the subset of graphs in $\mathcal{F}$ having exactly n vertices. The degree of a vertex v in a graph is denoted by $\Delta(v)$, and for non-negative integers k, the set of vertices in a graph G of degree k is denoted by V_k. The length of a binary string $x \in \{0,1\}^*$ is denoted by $|x|$.

Let $\mathcal{F}$ be a set of graphs. An *adjacency labeling scheme* (from hereon just *labeling scheme*) for $\mathcal{G}$ is a pair consisting of an *encoder* and a *decoder*. The encoder is an algorithm that receives $G \in \mathcal{G}$ as input and outputs a bit string $\mathcal{L}(v) \in \{0,1\}^*$ called the *label* of v. The decoder is an algorithm that receives any two labels $\mathcal{L}(v), \mathcal{L}(u)$ as input and outputs `true` iff u and v are adjacent in G and `false` otherwise. Note that the graph G is not an input to the decoder. The *size* of a labeling scheme is the map size : $\mathbb{N} \longrightarrow \mathbb{N}$ such that size(n) is the maximum length of any label assigned by the encoder to any vertex in any graph $G \in \mathcal{F}_n$. The *degree distribution* of a graph $G = (V, E)$ is the mapping $\text{ddist}_G(k) : \mathbb{N}_0 \longrightarrow \mathbb{Q}$ defined by $\text{ddist}_G(k) := \frac{|V_k|}{n}$.

We treat the family of *power-law* graphs, which is defined in the literature as the class of n vertex graphs G such that $\text{ddist}_G(k)$ is proportional to $k^{-\alpha}$ for some real number $\alpha > 1$ (note that such a graph, in principle, cannot have any isolated vertices). Thus, ideally, and ignoring rounding, we would like $\text{ddist}_G(k) = Ck^{-\alpha}$ for all k. As the degree distribution of a graph must be a probability distribution, we then have $\sum_{k=1}^{\infty} Ck^{-\alpha} = C\sum_{k=1}^{\infty} k^{-\alpha} = 1$, hence $C = 1/\zeta(\alpha)$ where ζ is the Riemann zeta function.

3 Graph families related to power-law graphs

In this section we study two families of graphs $\mathcal{P}_\alpha$ and $\mathcal{P}'_\alpha$ with $\mathcal{P}'_\alpha \subseteq \mathcal{P}_\alpha$. Family $\mathcal{P}_\alpha$ is rich enough to contain the graphs with power-law distribution that we are interested in, and our upper bound on the label size for our labeling scheme holds for any graph in $\mathcal{P}_\alpha$. Family $\mathcal{P}'_\alpha$ is used to show our lower bound. Let $C' \geq (\frac{C}{\alpha-1} + 4)^\alpha + \frac{C}{\alpha-1}$ be a constant.

Definition 1. *Let $\alpha > 1$ be a real number. $\mathcal{P}_\alpha$ is the family of graphs G such that if $n = |V(G)|$ then for all integers k between $\sqrt[\alpha]{n/\log n}$ and $n-1$, $\sum_{i=k}^{n-1} |V_i| \leq C'(\frac{n}{k^{\alpha-1}})$.*

The class of *α-proper* power law graphs contains graphs where the number of vertices of degree k must be $C\frac{n}{k^\alpha}$ rounded either up or down and the number of vertices of degree k is non-increasing with k. Note that the function $k \mapsto C\frac{1}{k^\alpha}$ is strictly decreasing.

Definition 2. *Let $\alpha > 1$ be a real number. We say that an n-vertex graph $G = (V, E)$ is an* α-proper power-law graph *if*

1. *$\lfloor Cn \rfloor - k - 1 \leq |V_1| \leq \lceil Cn \rceil$, where $k = \Theta(\sqrt[\alpha]{n})$ is the smallest integer such that $\lfloor Cn/k^\alpha \rfloor = 1$,*
2. *for every $2 \leq i \leq n$: $|V_i| \in \{\lfloor C\frac{n}{i^\alpha} \rfloor, \lceil C\frac{n}{i^\alpha} \rceil\}$, and*

[2]Procedures that "grow" random graphs whose degree distributions are with high probability "close" to power-law graphs

3. for every $2 \leq i \leq n-1$*:* $|V_i| \geq |V_{i+1}|$.

The family of α*-proper power-law graphs is denoted* $\mathcal{P}'_\alpha$.

We show the following properties of $\mathcal{P}'_\alpha$.

Proposition 1. *The maximum degree in an* n*-vertex graph of* $\mathcal{P}'_\alpha$ *is at most* $\left(\frac{C}{\alpha-1}+2\right)\sqrt[\alpha]{n}+2$.

Proof. Let $n>0$ be an arbitrary integer and let $k' = \lfloor \sqrt[\alpha]{n} \rfloor$. Furthermore, let $S_{k'} = \sum_{i=1}^{k'} |V_i|$, that is $S_{k'}$ is the number of vertices of degree at most k'.

Let $S^-_{k'} = \sum_{i=1}^{k'} \lceil Cni^{-\alpha} \rceil$. Then, $S_{k'} \geq S^-_{k'}$. We now bound $S^-_{k'}$ from below. For every i with $1 \leq i \leq k'$,

$$\begin{aligned}
S^-_{k'} + k' = &\sum_{i=1}^{k'} \left(\left\lfloor Cni^{-\alpha} \right\rfloor + 1 \right) \geq \sum_{i=1}^{k'} Cni^{-\alpha} = Cn \sum_{i=1}^{k'} i^{-\alpha} \geq \\
&n\left(1 - C \sum_{i=k'+1}^{\infty} i^{-\alpha}\right) \geq n\left(1 - C\int_{k'}^{\infty} x^{-\alpha} dx\right) = \\
&n\left(1 - C\left[\frac{1}{\alpha-1}x^{-\alpha+1}\right]_{\infty}^{k'}\right) = n\left(1 - \frac{C}{\alpha-1}\left(\lceil \sqrt[\alpha]{n} \rceil\right)^{-\alpha+1}\right) \geq \\
&n\left(1 - \frac{C}{\alpha-1}\left(\sqrt[\alpha]{n}\right)^{-\alpha+1}\right) = n - \frac{Cn}{\alpha-1} n^{-1+\frac{1}{\alpha}} = \\
&n - \frac{C}{\alpha-1}\sqrt[\alpha]{n},
\end{aligned}$$

giving $S_{k'} \geq S^-_{k'} \geq n - \frac{C}{\alpha-1}\sqrt[\alpha]{n} - \lceil \sqrt[\alpha]{n} \rceil$. There are thus at most $\frac{C}{\alpha-1}\sqrt[\alpha]{n} + \lfloor \sqrt[\alpha]{n} \rfloor$ vertices of degree strictly more than $k' = \lceil \sqrt[\alpha]{n} \rceil$. Since for every $1 \leq i \leq n-1$: $|V_i| \geq |V_{i+1}|$, it follows that the maximum degree of any α-proper power-law graph is at most $\left(\frac{C}{\alpha-1}+2\right)\sqrt[\alpha]{n}+2$. □

Proposition 2. *For* $\alpha > 2$*, all the graphs in* $\mathcal{P}'_\alpha$ *are sparse.*

Proof. By Proposition 1, the maximum degree of an n-vertex α-proper power-law graph is at most $k' \triangleq \left(\frac{C}{\alpha-1}+2\right)\sqrt[\alpha]{n}+2$, whence the total number of edges is at most $\frac{1}{2}\sum_{k=1}^{k'} k|V_k|$. By definition, $|V_k| \leq \lceil \frac{Cn}{k^\alpha} \rceil \leq \frac{Cn}{k^\alpha} + 1$, and thus

$$\begin{aligned}
\frac{1}{2}\sum_{k=1}^{k'} k|V_k| &\leq \frac{1}{2}\sum_{k=1}^{k'} k\left(\frac{Cn}{k^\alpha}+1\right) \leq \frac{k'(k'+1)}{4} + Cn\sum_{k=1}^{\infty} k^{-\alpha+1} \\
&= O(n^{2/\alpha}) + Cn\zeta(\alpha-1) = O(n).
\end{aligned}$$

□

Proposition 3. $\mathcal{P}'_\alpha \subseteq \mathcal{P}_\alpha$.

Proof. Let $d = \lfloor (\frac{C}{\alpha-1}+2)\sqrt[\alpha]{n}+2 \rfloor$. For any α-proper power-law graph with n vertices and for any k, $|V_k| \leq Ck^{-\alpha}n + 1$ and by Proposition 1, $|V_k| = 0$ when $k > d$.

Let k be an arbitrary integer between $\sqrt[\alpha]{n/\log n}$ and $n-1$. We need to show that $\sum_{i=k}^{n-1} |V_i| \leq C'(\frac{n}{k^{\alpha-1}})$.

It suffices to show this for $k \leq d$. We have

$$\begin{aligned}
\sum_{i=k}^{n-1} |V_i| &\leq \sum_{i=k}^{d} (Ci^{-\alpha}n + 1) = d - k + 1 + Cn \sum_{i=k}^{d} i^{-\alpha} \\
&\leq (\frac{C}{\alpha - 1} + 2)\sqrt[\alpha]{n} + 3 - k + Cn \int_{k}^{d} x^{-\alpha} dx \\
&\leq (\frac{C}{\alpha - 1} + 4)\sqrt[\alpha]{n} + Cn \left[\frac{1}{\alpha - 1} x^{-\alpha+1}\right]_{\infty}^{k} \\
&\leq ((\frac{C}{\alpha - 1} + 4)(\frac{\sqrt[\alpha]{n} d^{\alpha-1}}{n}) + \frac{C}{\alpha - 1})nk^{-\alpha+1} \\
&\leq ((\frac{C}{\alpha - 1} + 4)(\frac{C}{\alpha - 1} + 4)^{\alpha-1} + \frac{C}{\alpha - 1})nk^{-\alpha+1} \leq C'nk^{-\alpha+1},
\end{aligned}$$

as desired. □

4 The Labeling Schemes

In this section, we give our upper bounds for c-sparse graphs and for the family $\mathcal{P}_\alpha$. Both labeling schemes partition vertices into *thin* vertices which are of low degree and *fat* vertices of high degree. The *degree threshold* for the scheme is the lowest possible degree of a fat vertex. We start with c-sparse graphs.

Proposition 4. *There is a* $\sqrt{2cn \log n} + 2 \log n + 1$ *labeling scheme for* $\mathcal{S}_{c,n}$.

Proof. Let $G = (V, E)$ be an n-vertex c-sparse graph. Let $f(n)$ be the degree threshold for n-vertex graphs; we choose $f(n)$ below. With k denoting the number of fat vertices of G, assign each of them a unique identifier between 1 and k. Each thin vertex is given a unique identifier between $k + 1$ and n.

For a $v \in V$, the first part of the label $\mathcal{L}(v)$ is a single bit indicating whether v is thin or fat followed by a string of $\log n$ bits representing its identifier. If v is thin, the last part of $\mathcal{L}(v)$ is the concatenation of the identifiers of the neighbors of v. If v is fat, the last part of $\mathcal{L}(v)$ is a *fat bit string* of length k where the ith bit is 1 iff v is incident to the (fat) vertex with identifier i.

Decoding a pair $(\mathcal{L}(u), \mathcal{L}(v))$ is now straightforward: if one of the vertices, say u, is thin, u and v are adjacent iff the identifier of v is part of the label of u. If both u and v are fat then they are adjacent iff the ith bit of the fat bit string of $\mathcal{L}(u)$ is 1 where i is the identifier of v.

Since $|E| \leq cn$, we have $k \leq 2cn/f(n)$. A fat vertex thus has label size $1 + \log n + k \leq 1 + \log n + 2cn/f(n)$ and a thin vertex has label size at most $1 + \log n + f(n) \log n$. To minimize the maximum possible label size, we solve $2cn/x = x \log n$. Solving this gives $x = \sqrt{2cn/\log n}$ and setting $f(n) = \lceil x \rceil$ gives a label size of at most $1 + \log n + (\sqrt{2cn/\log n} + 1) \log n \leq 1 + 2 \log n + \sqrt{2cn \log n}$. □

By Proposition 2, graphs in $\mathcal{P}'_\alpha$ are sparse for $\alpha > 2$. This gives a label size of $O(\sqrt{n \log n})$ with the labeling scheme in Proposition 4. We now show that this label can be significantly improved, by constructing a labeling scheme for $\mathcal{P}_\alpha$ which contains $\mathcal{P}'_\alpha$.

Proposition 5. *There is a* $\sqrt[\alpha]{C'n}(\log n)^{1-1/\alpha} + 2 \log n + 1$ *labeling scheme for* $\mathcal{P}_\alpha$.

Proof. The proof is very similar to that of Proposition 4. We let $f(n)$ denote the degree threshold. If we pick $f(n) \geq \sqrt[\alpha]{n/\log n}$ then by Definition 1 there are at most $C'n/f(n)^{\alpha-1}$ fat vertices. Defining labels in the same way as in Proposition 4 gives a label size for thin vertices of at most $1 + \log n + f(n) \log n$ and a label size for fat vertices of at most $1 + \log n + C'n/f(n)^{\alpha-1}$. We minimize by solving $x \log n = C'n/x^{\alpha-1}$, giving $x = \sqrt[\alpha]{C'n/\log n}$. Setting $f(n) = \lceil x \rceil$ gives a label size of at most $\sqrt[\alpha]{C'n}(\log n)^{1-1/\alpha} + 2 \log n + 1$. □

5 Lower Bounds

We now derive lower bounds for the label size of any labeling schemes for both $\mathcal{S}_{c,n}$ and $\mathcal{P}'_\alpha$. Our proofs rely on Moon's [37] lower bound of $\lfloor n/2 \rfloor$ bits for labeling scheme for general graphs. We first show that the upper bound achieved for sparse graphs is close to the best possible. The following proposition is essentially a more precise version of the lower bound suggested by Spinrad [41].

Proposition 6. *Any labeling scheme for $\mathcal{S}_{c,n}$ requires labels of size at least $\left\lfloor \frac{\sqrt{cn}}{2} \right\rfloor$ bits.*

Proof. Assume for contradiction that there exists a labeling scheme assigning labels of size strictly less than $\lfloor \frac{\sqrt{cn}}{2} \rfloor$. Let G be an n-vertex graph. Let G' be the graph resulting by adding $\lfloor \frac{n^2}{c} \rfloor - n$ isolated vertices to G, and note that now G' is c-sparse. The graph G is an induced subgraph of G'. It now follows that the vertices of G have labels of size strictly less than $\lfloor \frac{\sqrt{c\lfloor n^2/c \rfloor}}{2} \rfloor \leq n/2$ bits. As G was arbitrary, we obtain a contradiction. □

5.1 Lower bound for power-law graphs

In the remainder of this section we are assuming that $\alpha > 2$ and prove the following:

Proposition 7. *For all n, any labeling scheme for n-vertex graphs of $\mathcal{P}_\alpha$ requires label size $\Omega(\sqrt[\alpha]{n})$.*

More precisely, we present a lower bound for $\mathcal{P}'_\alpha$ which is contained in $\mathcal{P}_\alpha$. Let $n \in \mathbb{N}$ be given and let H be an arbitrary graph with k vertices where k is the smallest integer such that $\lfloor Cn/k^\alpha \rfloor = 1$. Note that $k = \Theta(\sqrt[\alpha]{n})$. We show how to construct an α-proper power-law graph $G = (V, E)$ with n vertices which contains H as an induced subgraph. Observe that a labeling of G induces a labeling of H. Since H was chosen arbitrarily and since any labeling scheme for k-vertex graphs requires $\Omega(k)$ label size in the worst case, Proposition 7 follows if we can show the existence of G.

We construct G incrementally where initially $E = \emptyset$. Partition V into subsets $V_1, \ldots, V_n$ as follows. Set V_1 has size $\lfloor Cn \rfloor - k$. For $i = 2, \ldots, k-1$, V_i has size $\lfloor Cn/i^\alpha \rfloor$. Letting $n' = \sum_{i=1}^{k-1} |V_i|$, we set the size of V_i to 1 for $i = k, \ldots, k + n - n' - 1$ and the size of V_i to 0 for $i = k + n - n', \ldots, n$, thereby ensuring that the sizes of all sets sum up to n. Observe that $\sum_{i=1}^{k} \lfloor Cn/i^\alpha \rfloor \leq n$ so that $n' \leq n - k$, implying that $n - n' \geq k$. Hence we have at least k size 1 subsets $V_k, \ldots, V_{k+n-n'-1}$ in each of which the vertex degree allowed by Defintion 2 is at least k.

Let $v_1, \ldots, v_k$ be an ordering of $V(H)$, form a set $V_H \subseteq V$ of k arbitrary vertices from sets $V_k, \ldots, V_{k+n-n'-1}$, and choose an ordering $v'_1, \ldots, v'_k$ of them. For all $i, j \in \{1, \ldots, k\}$, add edge (v'_i, v'_j) to G iff $(v_i, v_j) \in E(H)$. Now, H is an induced subgraph of G and since the maximum degree of H is $k - 1$, no vertex of V_i exceeds the degree bound allowed by Definition 2 for $i = 1, \ldots, n$.

We next add additional edges to G in three phases to ensure that it is an α-proper power law graph while maintaining the property that H is an induced subgraph of G. For $i = 1, \ldots, n$, during the construction of G we say that a vertex $v \in V_i$ is *unprocessed* if its degree in the current graph G is strictly less than i. If the degree of v is exactly i, v is *processed*.

Phase 1: Let $V' = V \setminus (V_1 \cup V_H)$. Phase 1 is as follows: while there exists a pair of unprocessed vertices $(u, v) \in V' \times V_H$, add (u, v) to E.

When Phase 1 terminates, H is clearly still an induced subgraph of G. Furthermore, all vertices of V_H are processed. To see this, note that the sum of degrees of vertices of V_H when they are all processed is $O(k^2) = O(n^{2/\alpha})$ which is $o(n)$ since $\alpha > 2$. Furthermore, prior to Phase 1, each of the $\Theta(n)$ vertices of V' have degree 0 and can thus have their degrees increased by at least 1 before being processed.

Phase 2: Phase 2 is as follows: while there exists a pair of unprocessed vertices $(u, v) \in V' \times V'$, add (u, v) to E. At termination, at most one vertex of V' remains unprocessed. If such a vertex exists we process it by connecting it to $O(\sqrt[\alpha]{n})$ vertices of V_1; since $|V_1| = \Theta(n)$, there are enough vertices of V_1 to accomodate this.

Furthermore, prior to adding these edges, all vertices of V_1 have degree 0 so we do not exceed the bound allowed for vertices of this set.

Phase 3: In Phase 3, we add edges between pairs of unprocessed vertices of V_1 until no such pair exists. If no unprocessed vertices remain we have the desired α-proper power law graph G. Otherwise, let $w \in V_1$ be the unprocessed vertex of degree 0. We add a single edge from w to another vertex w' of V_1, thereby processing w and moving w' from V_1 to V_2. We may assume that $Cn/2^\alpha$ is not an integer and the resulting graph G is an α-proper power law graph since $\lfloor Cn \rfloor - k - 1 \leq |V_1| \leq \lceil Cn \rceil$ and $|V_2| = \lceil Cn/2^\alpha \rceil$. This shows Proposition 7.

6 Scale free graphs from generative models

The Barabási-Albert (BA) model is a very well-known generative model for power-law graphs that, roughly, grows a graph in a sequence of time steps by inserting a single vertex at each step and attaching it to m existing vertices with probability weighted by the degree of each existing vertex [11]. The BA model generates graphs that asymptotically have a power-law degree distribution ($\alpha = 3$) for low-degree nodes [12]. Graphs created by the BA model have low arboricity[3] [25]; we use this to prove the following highly efficient labeling scheme.

Proposition 8. *The family of graphs generated by the BA model has an $O(m \log n)$ adjacency labeling scheme.*

Proof. Let $G = (V, E)$ be an n-vertex graph resulting by the construction by the BA model with some parameter m (starting from some graph $G_0 = (V_0, E_0)$ with $|V_0| \ll n$). While it is not known how to compute the arboricity of a graph efficiently, it is possible in near-linear time to compute a partition of G with at most twice[4] the number of forests in comparison to the optimal [8]. We can thus decompose the graph to $2m$ forests in near linear time and label each forest using Alstrup and Rauhe's [7] $\log n + O(\log^* n)$ labeling scheme for trees, and achieve a $2m(\log n + O(\log^* n))$ labeling scheme for G. □

Note that if the encoder operates at the same time as the creation of the graph, Proposition 8 can be strengthened to yield an an $m \log n$ labeling scheme: simply store the identifiers of the m vertices attached with every vertex insertion. Propositions 7 and 8 strongly suggest that, for each sufficiently large n, the number of power-law graphs with n vertices is vastly larger than the number of graphs with n vertices created by the BA model. In contrast, other generative models such as Waxman [42], N-level Hierarchical [14]. and Chung's [18] (Chapter 3) do not seem to have an obvious smaller label size than the one in Proposition 5.

7 Experimental Study

We now perform an experimental evaluation of our labeling scheme on a number of large networks. The source code for our experiments can be found in: http://www.diku.dk/∼noyro/powerlaw.zip

7.1 Experimental Framework

Performance Indicators. Recall that our labeling scheme consists of two ideas: separation of the nodes according to some threshold, and selecting a threshold depending on the power-law parameter α. The *predicted* threshold of our labeling scheme is $\lceil \sqrt[\alpha]{Cn/(\alpha - 1)} \rceil$. For a concrete graph G, it is conceivable that some other threshold n_0 would yield a labeling scheme with smaller size. Let $\max_t(n_0)$ and $\max_f(n_0)$ be the maximum label sizes of thin, resp. fat vertices in G when the threshold is set at $1 \leq n_0 \leq n - 1$.

[3] The arboricity of a graph is the min. number of spanning forests needed to cover its edges.

[4] More precisely, for any $\epsilon \in (0, 1)$ there exist an $O(|E(G)|/\epsilon)$ algorithm [33] that computes such partition using at most $(1 + \epsilon)$ times more forests than the optimal.

Clearly the maximum label size with the threshold n_0 is $\max\{\max_t(n_0), \max_f(n_0)\}$. Observe further that $\max_t(n_0)$ and $\max_f(n_0)$ are monotonically increasing, resp. decreasing functions of n_0. Hence, the n_0 for which $\max\{\max_t(n_0), \max_f(n_0)\}$ is minimal is where the curves of $\max_t(n_0)$ and $\max_f(n_0)$ intersect. We call this n_0 the *empirical* threshold.

Performance indicator 1: We measure the relative difference in label size between the predicted and empirical threshold. A low relative difference means that the predicted threshold can achieve small label sizes without examining the global properties of the network other than the power-law parameter α.

Performance indicator 2: We compare the label sizes attained by our labeling schemes to other labeling schemes, namely state-of-the art labeling schemes for the classes of bounded-degree, sparse and general graphs using the labeling schemes suggested in [2], Proposition 4 and [6]. We interpret small label sizes for our scheme, especially in comparison with "small" classes like the class of bounded-degree graphs, as a sign that our labeling scheme efficiently utilizes the extra information about the graphs: namely that their degree distribution is reasonably well-approximated by a power-law.

Test Sets. We employ both real-world and synthetic data sets.

The *synthetic* data sets are created by first generating a power-law degree sequence using the method of Clauset et al. [20, App. D], subsequently constructing a corresponding graph for the sequence using the Havel-Hakimi method [27]. We use the range $2 < \alpha < 3$ as suggested in [20] and generate graphs of $300,000$ and $1M.$ vertices denoted $\text{s300}^{\alpha=x}$ and $\text{s1M}^{\alpha=x}$ respectively, for $x \in \{2.2, 2.4, 2.6, 2.8\}$.

The three *real-world* data sets originate from articles that found the data to be well-approximated by a power-law. The WWW dataset [5] shows the distribution of in- and out-degrees of webpages within the nd.edu domain. The ENRON dataset [34] shows the email communication network between Enron employees. The INTERNET dataset [38] provides a snapshot the Internet structure at the level of autonomous systems, reconstructed from BGP tables. For all of these sets, we consider the underlying simple, undirected graphs. For each set, standard maximum likelihood methods were used to compute the parameter α of the best-fitting power-law curve [20]. Additional information on all data sets is found in Table 1.

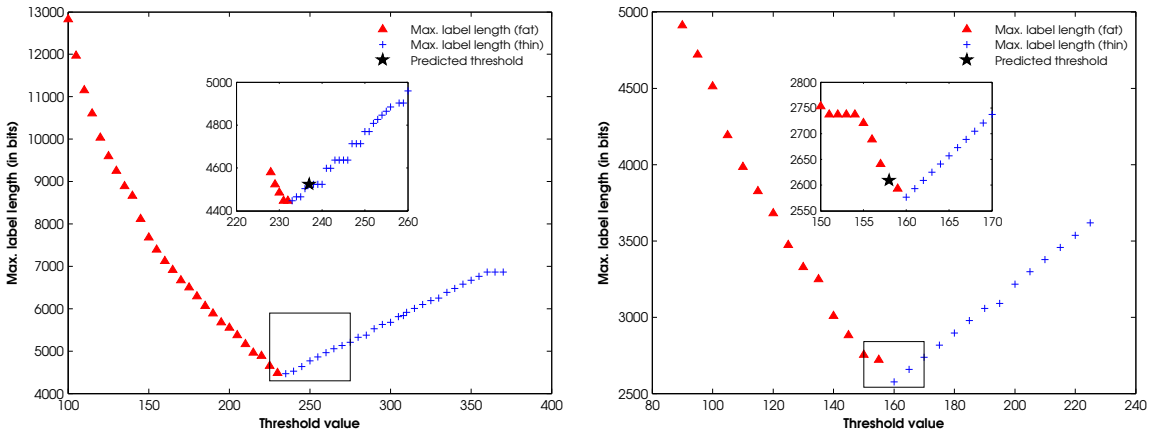
Real-Life					
Dataset	$\|V\|$	$\|E\|$	α	Δ_{max}	Source
WWW	325,729	1,117,563	2.16	10,721	[5]
ENRON	36,692	183,830	1.97	1,383	[34]
INTERNET	22,963	48,436	2.09	2,390	[38]
Synthetic					
$\text{s1M}^{\alpha=2.4}$	1,000,000	1,127,797	2.4	42,683	–
$\text{s1M}^{\alpha=2.6}$	1,000,000	878,472	2.6	12,169	–
$\text{s1M}^{\alpha=2.8}$	1,000,000	751,784	2.8	1,692	–
$\text{s300}^{\alpha=2.2}$	300,000	491,926	2.2	10,906	–
$\text{s300}^{\alpha=2.4}$	300,000	327,631	2.4	3,265	–
$\text{s300}^{\alpha=2.6}$	300,000	261,949	2.6	1,410	–
$\text{s300}^{\alpha=2.8}$	300,000	227,247	2.8	1,842	–

Table 1: Datasets and their properties. All graphs are undirected and simple. Δ_{max} stands for the maximum degree amongst the graph's nodes.

7.2 Findings

Figure 1 shows the distribution of maximum label sizes for one synthetic and one real-world dataset. The maximum label size for the predicted and empirical thresholds as well as upper bounds on the label sizes from different label schemes in the literature can be seen in Table 2 for two synthetic datasets and all three real-world datasets. For Plots for the remaining datasets see Appendix A.

Table 2 shows the maximum label sizes achieved using different labeling schemes on our data sets. "Predicted" shows the experimental maximum label size obtained by running our scheme on the graphs,



(a) $\text{syn300}^{\alpha=2.2}$ (b) ENRON

Figure 1: Maximum label sizes of different threshold values for the $\text{syn300}^{\alpha=2.2}$ and ENRON datasets. The triangles and crosses represent that for the tested threshold the largest label belong to fat, resp. thin node. The star indicate the position of the predicted threshold.

"Empirical" is the label size attained by using the empirical threshold. The remaining columns show non-experimental upper bounds for different label schemes: "Bound" is the upper bound guaranteed in Proposition 5, "C-sparse" is the labeling scheme for sparse graphs defined in Proposition 4" using simple concatenation of labels to represent the fat bit string[5], "BD" is the $\lceil \frac{\Delta}{2} \rceil \lceil \log n \rceil$ bounded degree graph labeling of [2], and AKTZ is the $\lceil n/2 \rceil + 6$ general graph labeling of [6].

Dataset	Predicted	Empirical	Bound	C-sparse	BD [2]	AKTZ [6]
$\text{s1M}^{\alpha=2.4}$	4,841	4,821	19,873	30,060	426,820	500,006
$\text{s1M}^{\alpha=2.6}$	3,361	3,201	12,642	26,540	121,680	500,006
$\text{s1M}^{\alpha=2.8}$	2,101	2,061	8,583	24,560	16,920	500,006
$\text{s300}^{\alpha=2.2}$	4,523	4,447	17,938	18,867	103,607	150,006
$\text{s300}^{\alpha=2.4}$	2,775	2,680	10,996	15,409	31,008	150,006
$\text{s300}^{\alpha=2.6}$	1,958	1,920	7,276	13,775	13,395	150,006
$\text{s300}^{\alpha=2.8}$	1,350	1,312	5,106	12,844	17,499	150,006
WWW	5,245	3,060	20,912	28,443	101,840	162,870
ENRON	2,609	2,577	10,243	9,728	11,056	18,352
INTERNET	1,426	1,156	5,706	4,575	17,925	11,487

Table 2: Label size in bits of labeling schemes. The two leftmost columns are experimental results; the remaining are upper bounds on label sizes computed from the characteristics of the data sets.

Our findings show that our labeling scheme obtains maximum label size at most 3% larger than what would have been obtained by using the empirical threshold for all synthetic datasets. This is expected—the synthetic datasets are graphs generated specifically to have power-law distributed degree distribution. For the real-world data sets, the labeling scheme obtain maximum label size at most 23% larger than by using the empirical threshold; this larger deviation is likely due to degree distributions of the datasets being close to, but not quite, power-law distributions due to natural phenomena or noise. E.g., for the ENRON dataset there is sudden drop in frequency between nodes of degree < 158 and ≥ 158. As expected, both our experimental

[5]Our labeling schemes introduced in this paper all make use of a succinctly represented "fat bit string"; for our experiments, we use simple concatenation of labels instead of a bit string; while this incurs a $(\log n)/\alpha$ factor on the label size, it simplifies the exposition.

results and theoretical upper bounds for our labeling scheme are several orders of magnitudes lower than for labeling schemes aimed at more general classes of graphs. Of these, it is most interesting to compare the upper bound of bounded degree graphs—the most restrictive class of graphs that both contains the class of power-law graphs and has an efficient labeling scheme described in the literature [2]. As seen in Table 2, the upper bound on our labeling schemes for both power-law graphs and sparse graphs have better upper bounds on label sizes, but only marginally so for datasets with low maximum degree and low values of the power-law parameter α, e.g. ENRON ($\alpha = 1.97$).

8 Conclusion and future work

8.1 Future work

It is of theoretical interest to close the gap between upper bound $O(\sqrt[c]{n} \log n^{1-1/\alpha})$ and lower bound $\Omega(\sqrt[c]{n})$ on label size for power-law graphs. Second, our labeling scheme can easily support directed graphs using one more bit per edge in each label; it is interesting to investigate the overhead incurred by distributing the storage of the graph topology to the labels as per our labeling scheme compared to the substantial body of work on storing directed power-law graphs directly in main memory (so-called "web-graph compression") [26, 9, 10, 19]. Finally, we note that the label sizes attained in Sec. 7.1 can be reduced by using the succinctly represented "fat bit string" as well as an additional rule that prevents storing an edge in two labels. Both improvements can improve the label sizes with factors $(\log n)^{1-1/\alpha}$ and 2 respectively. In our data sets, this corresponds to an approximate factor 10, which will make our labeling scheme even more practical.

References

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs. *J. ACM*, 56(4), 2009.

[2] D. Adjiashvili and N. Rotbart. Labeling schemes for bounded degree graphs. In *Automata, Languages, and Programming*, pages 375–386. Springer, 2014.

[3] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

[4] A. Akella, S. Chawla, A. Kannan, and S. Seshan. Scaling properties of the internet graph. In *Proceedings of the Twenty-Second ACM Symposium on Principles of Distributed Computing, PODC 2003*, pages 337–346, 2003.

[5] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[6] S. Alstrup, H. Kaplan, M. Thorup, and U. Zwick. Adjacency labeling schemes and induced-universal graphs. *To appear in the 47th symposium on Theory of computing (STOC)*, 2015.

[7] S. Alstrup and T. Rauhe. Small induced-universal graphs and compact implicit graph representations. In *Proceedings of the 43rd Symposium on Foundations of Computer Science*, FOCS '02, pages 53–62, Washington, DC, USA, 2002. IEEE Computer Society.

[8] S. R. Arikati, A. Maheshwari, and C. D. Zaroliagis. Efficient computation of implicit representations of sparse graphs. *Discrete Applied Mathematics*, 78(1):1–16, 1997.

[9] Y. Asano, T. Ito, H. Imai, M. Toyoda, and M. Kitsuregawa. Compact encoding of the web graph exploiting various power laws. In *Advances in Web-Age Information Management*, pages 37–46. Springer, 2003.

[10] Y. Asano, Y. Miyawaki, and T. Nishizeki. Efficient compression of web graphs. In *Computing and Combinatorics*, pages 1–11. Springer, 2008.

[11] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

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

[13] A. Brady and L. J. Cowen. Compact routing on power law graphs with additive stretch. In *ALENEX*, volume 6, pages 119–128. SIAM, 2006.

[14] K. L. Calvert, M. B. Doar, and E. W. Zegura. Modeling internet topology. *Communications Magazine, IEEE*, 35(6):160–163, 1997.

[15] S. Caminiti, I. Finocchi, and R. Petreschi. Engineering tree labeling schemes: A case study on least common ancestors. In *Algorithms-ESA 2008*, pages 234–245. Springer, 2008.

[16] W. Chen, C. Sommer, S.-H. Teng, and Y. Wang. A compact routing scheme and approximate distance oracle for power-law graphs. *ACM Transactions on Algorithms*, 9(1):4, 2012.

[17] F. Chung and L. Lu. The average distance in a random graph with given expected degrees. *Internet Mathematics*, 1(1):91–113, 2004.

[18] F. R. Chung and L. Lu. *Complex Graphs and Networks*, volume 107. American mathematical society Providence, 2006.

[19] F. Claude and G. Navarro. Fast and compact web graph representations. *ACM Transactions on the Web (TWEB)*, 4(4):16, 2010.

[20] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[21] S. Dahlgaard, M. B. T. Knudsen, and N. Rotbart. Dynamic and multi-functional labeling schemes. In *Algorithms and Computation*, pages 141–153. Springer, 2014.

[22] J. Fischer. Short labels for lowest common ancestors in trees. In *Algorithms-ESA 2009*, pages 752–763. Springer, 2009.

[23] C. Gavoille and A. Labourel. Shorter implicit representation for planar graphs and bounded treewidth graphs. In *Algorithms–ESA 2007*, pages 582–593. Springer, 2007.

[24] C. Gavoille, D. Peleg, S. Pérennesc, and R. Razb. Distance labeling in graphs. *Journal of Algorithms*, 53:85–112, 2004.

[25] G. Goel and J. Gustedt. Bounded arboricity to determine the local structure of sparse graphs. In *Graph-Theoretic Concepts in Computer Science*, pages 159–167. Springer, 2006.

[26] J.-L. Guillaume, M. Latapy, and L. Viennot. Efficient and simple encodings for the web graph. In *Advances in Web-Age Information Management*, pages 328–337. Springer, 2002.

[27] S. L. Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial & Applied Mathematics*, 10(3):496–506, 1962.

[28] S. Kannan, M. Naor, and S. Rudich. Implicit representation of graphs. In *SIAM Journal On Discrete Mathematics*, pages 334–343, 1992.

[29] M. Katz, N. A. Katz, A. Korman, and D. Peleg. Labeling schemes for flow and connectivity. *SIAM Journal on Computing*, 34(1):23–40, 2004.

[30] A. Korman. General compact labeling schemes for dynamic trees. *Distributed Computing*, 20(3):179–193, 2007.

[31] A. Korman and D. Peleg. Compact separator decompositions in dynamic trees and applications to labeling schemes. In *Distributed Computing*, pages 313–327. Springer, 2007.

[32] A. Korman and D. Peleg. Labeling schemes for weighted dynamic trees. *Inf. Comput.*, 205(12):1721–1740, Dec. 2007.

[33] Ł. Kowalik. Approximation scheme for lowest outdegree orientation and graph density measures. In *Algorithms and computation*, pages 557–566. Springer, 2006.

[34] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[35] V. V. Lozin and G. Rudolf. Minimal universal bipartite graphs. *Ars Combinatoria*, 84:345–356, 2007.

[36] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[37] J. Moon. On minimal n-universal graphs. In *Proceedings of the Glasgow Mathematical Association*, volume 7, pages 32–33. Cambridge University Press, 1965.

[38] M. Newman. Network data. `http://www-personal.umich.edu/~mejn/netdata/`, 2013. [Online; accessed 02-Jan-2015].

[39] N. Rotbart, M. V. Salles, and I. Zotos. An evaluation of dynamic labeling schemes for tree networks. In *Experimental Algorithms*, pages 199–210. Springer, 2014.

[40] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power laws and the as-level internet topology. *IEEE/ACM Trans. Netw.*, 11(4):514–524, 2003.

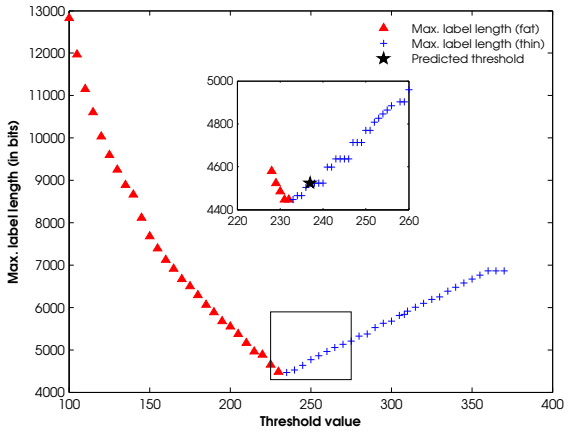
[41] J. P. Spinrad. *Efficient graph representations*. American mathematical society, 2003.

[42] B. M. Waxman. Routing of multipoint connections. *Selected Areas in Communications, IEEE Journal on*, 6(9):1617–1622, 1988.

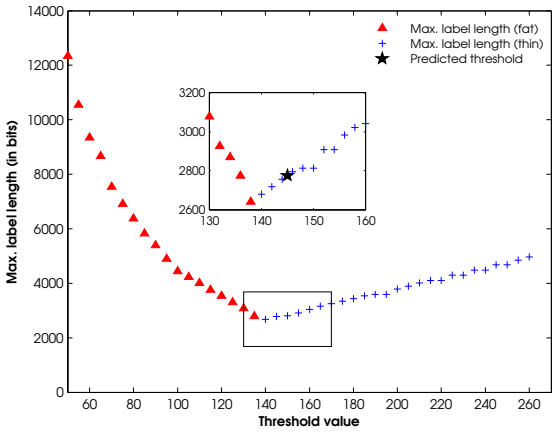
A Experimental results in detail

This appendix is organized as follows: Sections A.1 A.2 include the maximum label sizes for all synthetic and real-world data sets, respectively.

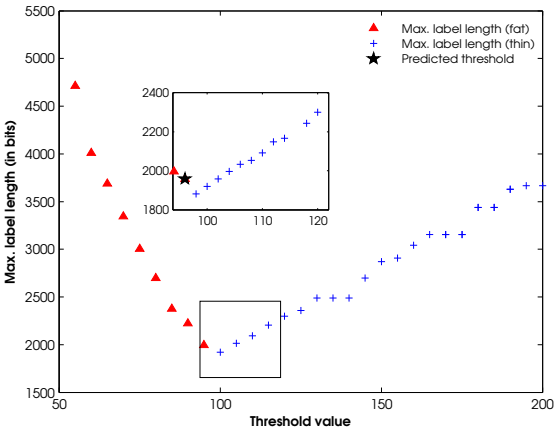
A.1 Maximum label size distribution for synthetic datasets



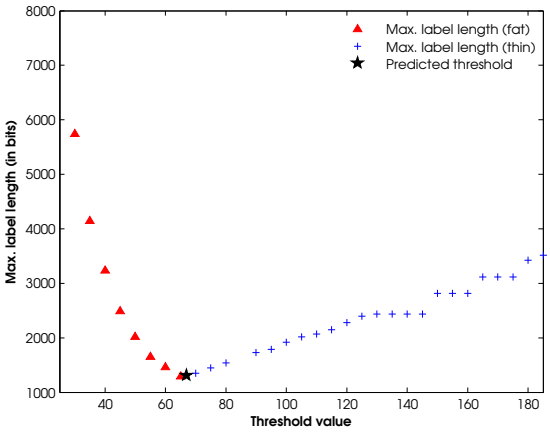
(a) syn300$^{\alpha=2.2}$



(b) syn300$^{\alpha=2.4}$

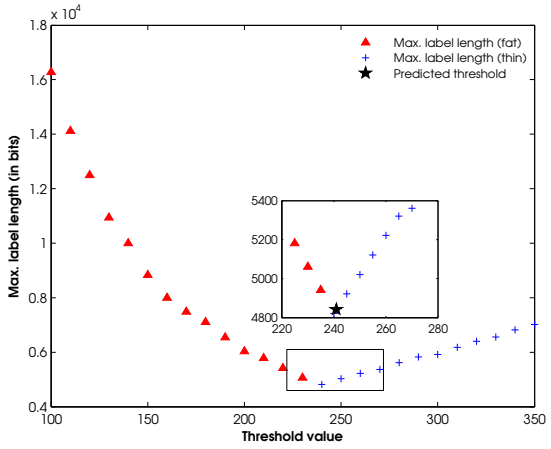


(c) syn300$^{\alpha=2.6}$

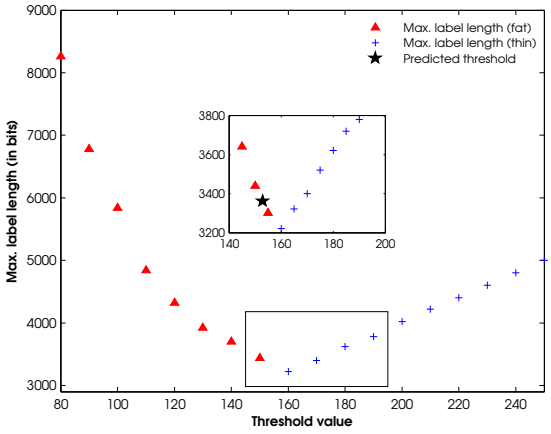


(d) syn300$^{\alpha=2.8}$

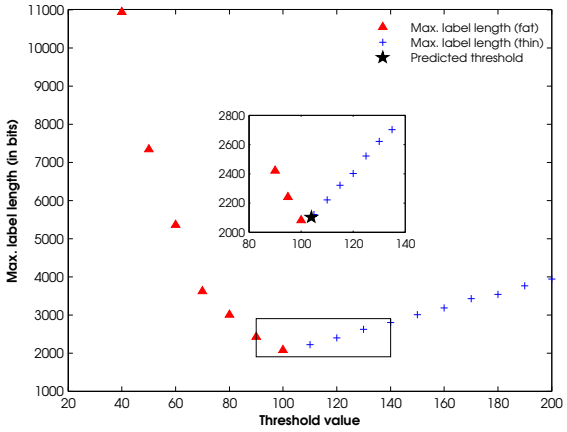
Figure 2: Distribution of maximum label sizes for four different synthetic datasets of $|V| = 300,000$. Each dataset was generated using one of α-values: 2.2, 2.4, 2.6, 2.8 using. Fat vertices are shown as red triangles and thin vertices as blue crosses. The black pentagram is the *predicted* maximum label size. The transition between fat and thin vertices is the *empirical* best maximum label size.



(a) syn1M$^{\alpha=2.4}$



(b) syn1M$^{\alpha=2.6}$

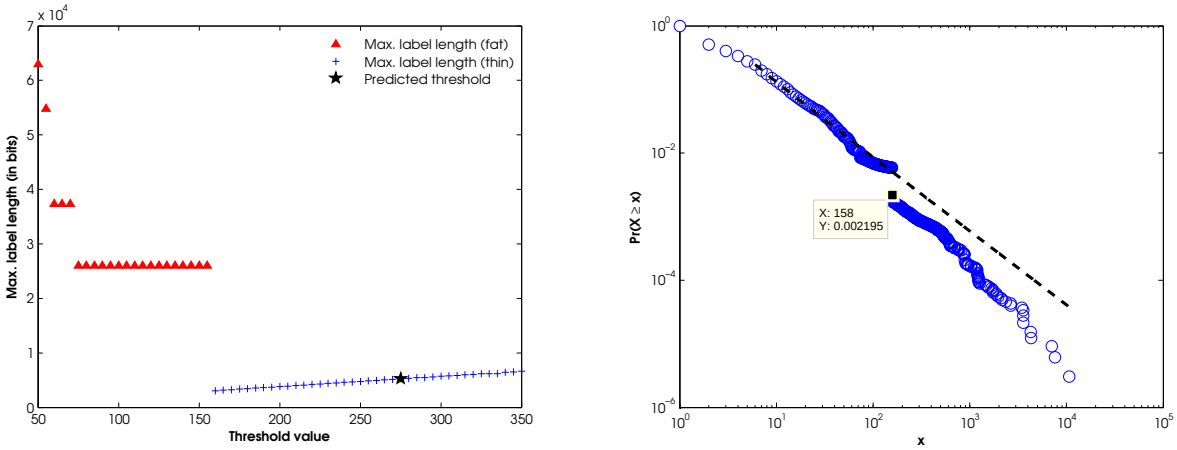


(c) syn1M$^{\alpha=2.8}$

Figure 3: Distribution of maximum label sizes for three different synthetic datasets of $|V| = 1,000,000$. Each dataset was generated using one of α-values: 2.4, 2.6, 2.8 using. Fat vertices are shown as red triangles and thin vertices as blue crosses. The black pentagram is the *predicted* maximum label size. The transition between fat and thin vertices is the *empirical* best maximum label size.

A.2 Maximum label size distribution for real-life datasets

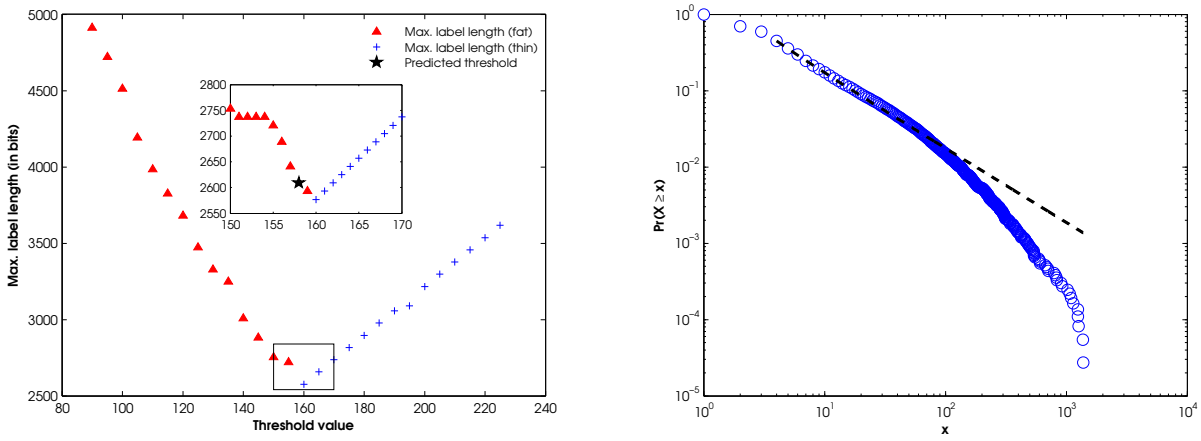
For completeness, we provide an illustration of the best-fitting power law fitted to the probability mass function of the data.



(a) Fat and thin vertices vs. threshold values

(b) Power law fit

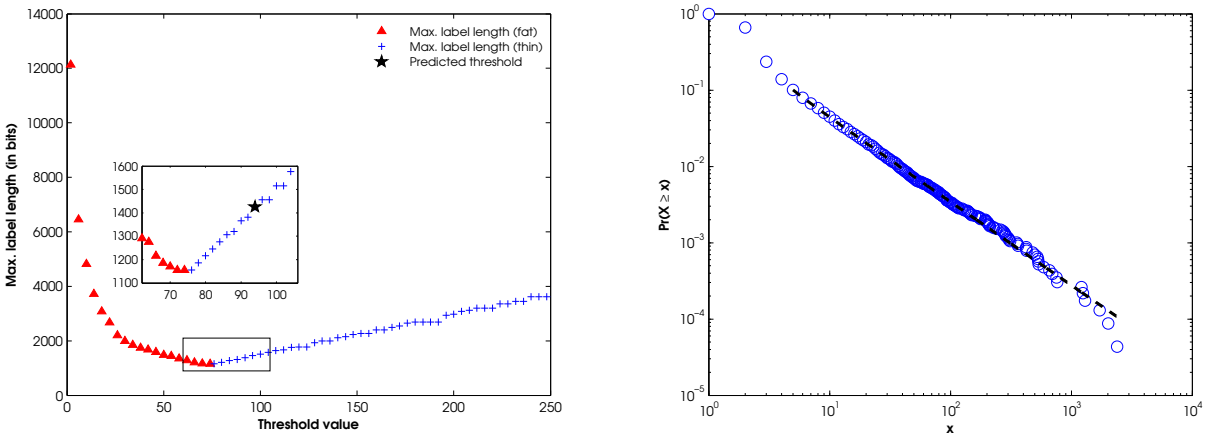
Figure 4: Left: Fat and thin vertices plotted against increasing threshold values for the WWW dataset. The black pentagram is the predicted threshold $(1/\zeta(\alpha)\sqrt[\alpha]{(n)})$ rounded to nearest integer. Right: Best-fitting power law ($\alpha = 2.16$) fitted to the probability mass function of the data. Power law fit plotted on top of the complementary cumulative distribution function (CCDF) using the framework by [20].



(a) Fat and thin vertices vs. threshold values

(b) Power law fit

Figure 5: Left: Fat and thin vertices plotted against increasing threshold values for the ENRON email communication dataset. The black pentagram is the predicted threshold $(1/\zeta(\alpha)\sqrt[\alpha]{(n)})$ rounded to the nearest integer. Right: Right: Best-fitting power law ($\alpha = 1.97$) fitted to the probability mass function of the data. Power law fit plotted on top of the complementary cumulative distribution function (CCDF) using the framework by [20].



(a) Fat and thin vertices vs. threshold values

(b) Power law fit

Figure 6: Left: Fat and thin vertices plotted against increasing threshold values for the INTERNET dataset. The black pentagram is the predicted threshold $(1/\zeta(\alpha)\sqrt[\alpha]{(n)})$ rounded to nearest integer. Right: Right: Best-fitting power law ($\alpha = 2.09$) fitted to the probability mass function of the data. Power law fit plotted on top of the complementary cumulative distribution function (CCDF) using the framework by [20].