Brief Announcement: Labeling Schemes for Power-Law Graphs

Casper Petersen
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
cazz@di.ku.dk

Noy Rotbart
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
noyro@di.ku.dk

Jakob Grue Simonsen
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
jakob@di.ku.dk

Christian Wulff-Nilsen
University of Copenhagen
Universitetsparken 5
2100 Copenhagen, Denmark
koolooz@di.ku.dk

Keywords

Labeling schemes, Power-law graphs

1. INTRODUCTION

A labeling scheme is a method of distributing the information about the structure of a graph among its vertices by assigning short *labels*, such that a selected function on pairs of vertices can be computed using *only* their labels. A labeling scheme consists of an *encoder* that has access to the entire graph and assigns labels to vertices, and a *decoder* that has access to only the labels of a smaller set of vertices (typically a pair) and returns information about this subset (e.g., whether two vertices are adjacent, or the distance between them in the graph). The main objective is to minimize the *maximum label size*: the maximum number of bits used in a label of any vertex. Among the applications of labeling schemes are XML search engines, mapping services, and internet routing.

Adjacency labeling schemes for general graphs, bounded degree graphs, trees and various important graph families were studied (See [2] for a comprehensive table), yielding remarkably close upper and lower bounds[1]. It is thus surprising that the important family of *power-law graphs* has not yet been studied for this question[2].

An n-vertex graph is power-law if the number of its vertices of degree k is proportional to n/k^{α} for some positive α. To solidify this somewhat vague definition, numerous probabilistic and deterministic definitions of power-law graphs are given in the literature. A recent deterministic model, called shifted power-law distribution, has recently proven to capture a vast number of such definitions, both in theory and experimentally in [4].

Power-law graphs (also called scale-free graphs in the literature) have been used to model numerous types of network (see, e.g., [7] for an overview). In our work, we perform the first theoretical and practical study of adjacency labeling schemes for classes of graphs whose statistical properties– in particular their *degree distribution*–more closely resemble that of real-world networks.

We assert that our research can also assist the study of storage of real-world networks, and efficient distribution thereof. Rather than graph compression or dissemination of the underlying graphs of these networks over several machines [8], we propose to disseminate the structural information of the graph to its vertices. This *peer-to-peer* strategy allows inferring the graph's local topology using only local information stored in each vertex without using costly access to large, global data structures. In particular, it can be useful to address privacy concerns and ensure a high survivability rate.

2. OUR CONTRIBUTION

In our full version [11] we contribute the following results for power-law graphs:

A discrete and simple characterisation of power-law graphs. We define and prove useful properties for two simple families of graphs, $\mathcal{P}_h$ and $\mathcal{P}_l$, where $\mathcal{P}_h$ contains and $\mathcal{P}_l$ is contained by the standard definitions of power-law graphs in the literature, including recent ones [4]. We use $\mathcal{P}_h$ and $\mathcal{P}_l$ to study upper and lower bounds respectively.

An $O(\sqrt[\alpha]{n}(\log n)^{1-1/\alpha})$ adjacency labeling scheme. 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. These ideas are illustrated in Figure 1. We also show that applying our labeling scheme for random graphs with a power-law distribution results in a small expected worst-case label size.

[1] e.g. the current label size gap for general graphs stands on 4 bits.

[2] Routing labeling schemes however were studied in [5].

PODC'16 July 25-28, 2016, Chicago, IL, USA

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-3964-3/16/07.

DOI: http://dx.doi.org/10.1145/2933057.2933060

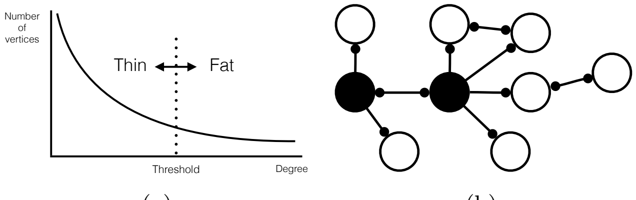
Real-Life Graphs

Data set	Vertices	Edges	Predicted	Empirical	Upper-Bound	c-sparse	Bounded degree
INTERNET	22,963	48,436	1,426	1,156	8,181	4,700	17,925
ENRON	36,692	183,830	2,609	2,577	15,835	9,735	11,056
WWW	325,729	1,117,563	5,245	3,060	29,225	28,445	101,840

Synthetic Graphs

Data set	Vertices	Edges	Predicted	Empirical	Upper-Bound	c-sparse	Bounded degree
s1M$^{\alpha=2.8}$	1,000,000	751,784	2,101	2,061	10,081	24,566	16,920
s300$^{\alpha=2.8}$	300,000	227,247	1,350	1,312	6,244	12,849	17,499

Table 1: Label size in bits for various labeling schemes. The *predicted* and *empirical* label sizes are explained below. The upper bound is the guarantee of performance by our algorithm, and we report the label size for the graphs using the sparse and bounded degree [1] labeling schemes.



(a) (b)

Figure 1: Two illustrations of the main idea: Figure (a) demonstrates the threshold assignment, figure (b) demonstrates the label assignment, in which fat (black) vertices do not store adjacency to thin (white) vertices.

Real-world power-law graphs rarely exceed 10^{10} vertices. Given that those typically have $2 \leq \alpha \leq 3$, our labeling scheme theoretically achieves label size of $10^4 - 10^5$ bits, well within the processing capabilities of current hardware. This along with their simplicity implies that our labeling scheme may be appealing in practice. Using the same ideas, and as a stepping stone, we also get an asymptotically near-tight $O(\sqrt{n \log n})$ labeling scheme for sparse graphs.

A lower bound of $\Omega(\sqrt[\alpha]{n})$ *for any adjacency labeling scheme.* We use our restrictive subclass of power-law graphs and show that it requires label size $\Omega(\sqrt[\alpha]{n})$ for n-vertex graphs. This lower bound shows that our upper bound above is asymptotically optimal, bar a $(\log n)^{1-1/\alpha}$ factor. By the connections between adjacency labeling schemes and universal graphs, we also obtain upper and lower bounds for induced universal graphs for power-law graphs.

An $O(\log n)$ *adjacency labeling scheme in two restricted settings.* We bypass the aforementioned lower bound in two restricted settings: (i) all power law graphs that are created using the popular Barabasi-Albert [3] model, and (ii) using a variant of labeling schemes called "labeling scheme with a query" [9], which allows for the decoder to consult the labels and fetch a third label to determine the query.

An experimental investigation of our labeling scheme. To test our $O(\sqrt[\alpha]{n}(\log n)^{1-1/\alpha})$ labeling scheme, we label both synthetic graphs (300K-1M vertices) and ones that are considered by practitioners as power-law (23K-3M vertices). We first note that our selected *strategy* is optimal when the threshold chosen balances the largest fat and thin vertex. Using this observation, we can estimate the quality of the *predicted* threshold given in our theoretical approach to the *empirical* optimum. We observe the following:

- As seen in Figure 2, our threshold *prediction* performs close to the *empirical* optimum using this *strategy*, in particular to graphs with a degree distribution close to power-law.
- As seen in Table 1, our labeling scheme achieves maximum label size that is several orders of magnitude smaller than the state-of-the-art labeling schemes for more general graph families [1].

A $o(n)$ *distance labeling scheme.* We demonstrate the usefulness of our strategy to arrive at a $o(n)$ distance labeling scheme. Our labeling scheme outperforms competing labeling schemes for small distances, in accordance to Chung and Lu's findings [6] on the small expected diameter of power-law graphs.

3. THE LABELING SCHEME

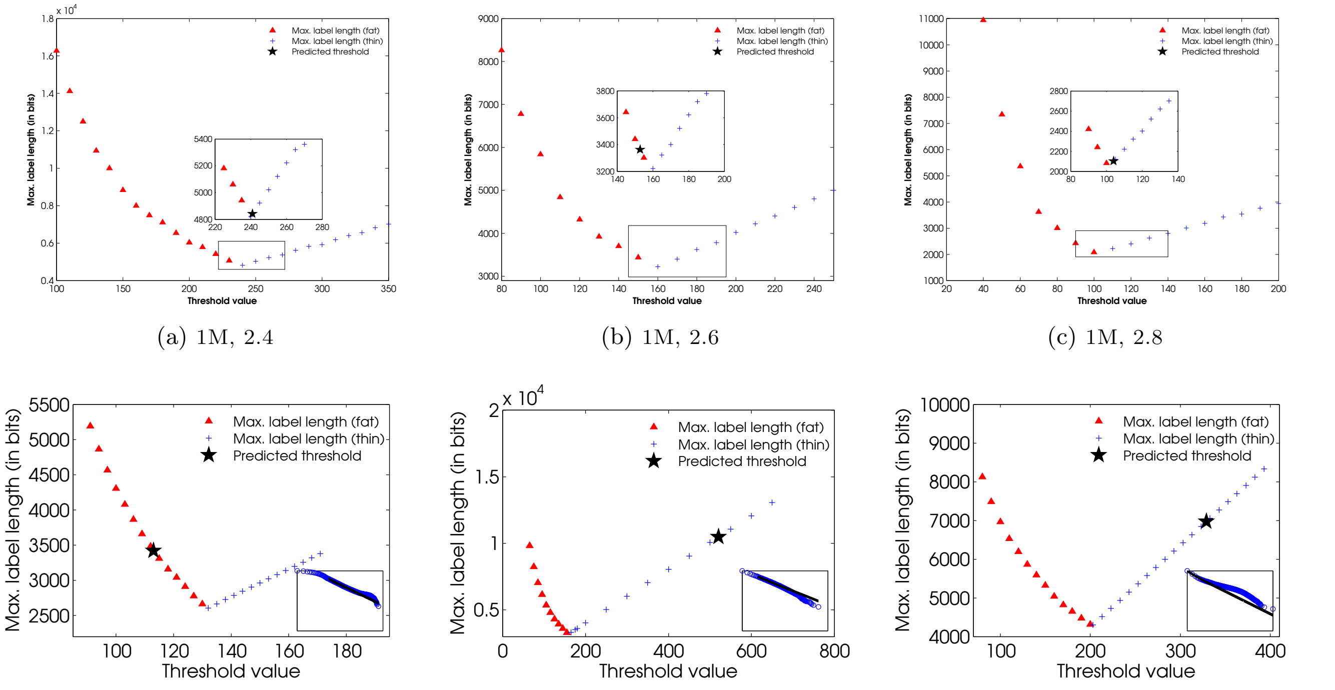
To demonstrate our technique we present our adjacency labeling scheme for the case of c-sparse[3] graphs. This family contain all power-law graphs, and can be explained succinctly without the additional definitions required for the case of power-law graphs as described in the full version [11]. As mentioned, we partition the 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.

Theorem 3.1. *There is a* $\sqrt{2cn \log n} + 2 \log n + 1$ *labeling scheme for c-sparse graphs.*

Proof. Let $G = (V, E)$ be an n-vertex c-sparse graph. Let $\tau(n)$ be the degree threshold for n-vertex graphs; we choose $\tau(n)$ below. Let k denote the number of fat vertices of G, and assign each fat vertex 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*

[3] Graphs of n vertices and at most cn edges.



(a) 1M, 2.4

(b) 1M, 2.6

(c) 1M, 2.8

(d) GOOGLE

(e) YOUTUBE

(f) WIKITALK

Figure 2: A demonstration of the maximum label size given a choice of threshold, for both synthetic graphs of million vertices and $\alpha = 2.4, 2.6, 2.8$ (top row) and for the real-life claimed power-law graphs GOOGLE, YOUTUBE and WIKITALK (taken from [10]). Our predicted threshold, marked in a black star, provides a close estimation for the location of this minimum.

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 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. Both decoding processes can be computed in $O(\log n)$ time using standard assumptions.

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

4. ACKNOWLEDGMENTS

The work is supported in part by the Sapere Aude project Complexity via Logic and Algebra (COLA).

5. REFERENCES

[1] S. Alstrup, S. Dahlgaard, and M. B. T. Knudsen. Optimal induced universal graphs and adjacency labeling for trees. FOCS '15, 2015.

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

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

[4] P. Brach, M. Cygan, J. Lacki, and P. Sankowski. Algorithmic complexity of power law networks. SODA 2016.

[5] W. Chen, C. Sommer, S.-H. Teng, and Y. Wang. A compact routing scheme and approximate distance oracle for power-law graphs. *TALG*, 9, 2012.

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

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

[8] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *OSDI*, 2012.

[9] A. Korman and S. Kutten. Labeling schemes with queries. In *Structural Information and Communication Complexity*, pages 109–123. Springer, 2007.

[10] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data, June 2014.

[11] C. Petersen, N. Rotbart, J. G. Simonsen, and C. Wulff-Nilsen. Near-optimal adjacency labeling scheme for power-law graphs. *To appear in ICALP 16'*.