Sublinear Distance Labeling

Stephen Alstrup*[1], Søren Dahlgaard†[2], Mathias Bæk Tejs Knudsen‡[3], and Ely Porat[4]

1 **University of Copenhagen, Copenhagen, Denmark**
s.alstrup@di.ku.dk
2 **University of Copenhagen, Copenhagen, Denmark**
soerend@di.ku.dk
3 **University of Copenhagen, Copenhagen, Denmark**
knudsen@di.ku.dk
4 **Bar-Ilan University, Ramat Gan, Israel**
porately@cs.biu.ac.il

Abstract

A distance labeling scheme labels the n nodes of a graph with binary strings such that, given the labels of any two nodes, one can determine the distance in the graph between the two nodes by looking only at the labels. A D-preserving distance labeling scheme only returns precise distances between pairs of nodes that are at distance at least D from each other. In this paper we consider distance labeling schemes for the classical case of unweighted and undirected graphs.

We present a $O(\frac{n}{D}\log^2 D)$ bit D-preserving distance labeling scheme, improving the previous bound by Bollobás et al. [SIAM J. Discrete Math. 2005]. We also give an almost matching lower bound of $\Omega(\frac{n}{D})$. With our D-preserving distance labeling scheme as a building block, we additionally achieve the following results:

1. We present the first distance labeling scheme of size $o(n)$ for sparse graphs (and hence bounded degree graphs). This addresses an open problem by Gavoille et al. [J. Algo. 2004], hereby separating the complexity from distance labeling in general graphs which require $\Omega(n)$ bits, Moon [Proc. of Glasgow Math. Association 1965].[1]
2. For approximate r-additive labeling schemes, that return distances within an additive error of r we show a scheme of size $O\left(\frac{n}{r}\cdot\frac{\text{polylog}(r\log n)}{\log n}\right)$ for $r\geq 2$. This improves on the current best bound of $O\left(\frac{n}{r}\right)$ by Alstrup et. al. [SODA 2016] for sub-polynomial r, and is a generalization of a result by Gawrychowski et al. [arXiv preprint 2015] who showed this for $r=2$.

1998 ACM Subject Classification G.2.2 Graph Theory, E.1 Data Structures, G.2.1 Combinatorics

Keywords and phrases Graph labeling schemes, Distance labeling, Graph theory, Sparse graphs

Digital Object Identifier 10.4230/LIPIcs.ESA.2016.5

* Research partly supported by the FNU project AlgoDisc – Discrete Mathematics, Algorithms, and Data Structures.
† Research partly supported by Mikkel Thorup's Advanced Grant from the Danish Council for Independent Research under the Sapere Aude research career programme.
‡ Research partly supported by the FNU project AlgoDisc – Discrete Mathematics, Algorithms, and Data Structures. Research partly supported by Mikkel Thorup's Advanced Grant from the Danish Council for Independent Research under the Sapere Aude research career programme.
[1] This result for sparse graphs was made available online in a preliminary version of this paper [6]. The label size was subsequently slightly improved by an $O(\log\log n)$ factor by Gawrychowski et al. [25].

1 Introduction

The concept of *informative labeling schemes* dates back to Breuer and Folkman [12, 13] and was formally introduced by Kannan et al. [30, 34]. A labeling scheme is a way to represent a graph in a distributed setting by assigning bit strings (called *labels*) to each node of the graph. In a distance labeling scheme we assign labels to a graph G from a family $\mathcal{G}$ such that, given *only* the labels of a pair of nodes, we can compute the distance between them without the need for a centralized data structure. When designing a labeling scheme the main goal is to minimize the *maximum label size* over all nodes of all graphs G in the family $\mathcal{G}$. We call this the size of the labeling scheme. As a secondary goal some papers consider the *ecoding* and *decoding* time of the labeling scheme in various computational models. In this paper we study the classical case of *undirected* and *unweighted* graphs.

Exact distances. The problem of exact distance labeling in general graphs is a classic problem that was studied thoroughly in the 1970/80's. Graham and Pollak [26] and Winkler [39] showed that labels of size $\lceil (n-1) \cdot \log_2 3 \rceil$ suffice in this case. Combining [30] and [33] gives a lower bound of $\lceil n/2 \rceil$ bits (see also [24]). Recently, Alstrup et al. [7] improved the label size to $\frac{\log_2 3}{2} n + O\left(\log^2 n\right)$ bits.

Distance labeling schemes have also been investigated for various families of graphs, providing both upper and lower bounds. For trees, Peleg [35] showed that labels of size $O(\log^2 n)$ suffice with a matching lower bound by Gavoille et. al [24]. Gavoille et. al [24] also showed a $\Omega(n^{1/3})$ lower bound for planar graphs and $\Omega(\sqrt{n})$ bound for bounded degree (and thus sparse) graphs. They also provided an $O(\sqrt{n} \log n)$ labeling scheme for planar graphs, however nothing better than the $O(n)$ scheme for general graphs is known for bounded-degree graphs. It remains a major open problem in the field of labeling schemes whether a scheme of size $O(\sqrt{n})$ or even $o(n)$ exists for bounded-degree graphs as stated in e.g. [24].

Other families of graphs studied include distance-hereditary [22], bounded clique-width [16], some non-positively curved plane [15], as well as interval [23] and permutation graphs [10].

Approximate distances. For some applications, the $\Omega(poly(n))$ requirement on the label size for several graph classes is prohibitive. Therefore a large body of work is dedicated to labeling schemes for approximating distances in various families of graphs [2, 14, 19, 24, 27, 28, 32, 35, 36, 37, 38]. Such labeling schemes often provide efficient implementations of other data structures like distance oracles [37] and dynamic graph algorithms [2].

In [35] a labeling scheme of size $O(\log^2 n \cdot \kappa \cdot n^{1/\kappa})$ was presented for approximating distances up to a factor[2] of $\sqrt{8\kappa}$. In [37] a scheme of poly-logarithmic size was given for planar graphs when distances need only be reported within a factor of $(1+\varepsilon)$. Labeling schemes of additive error have also been investigated. For general graphs Alstrup et. al [7] gave a scheme of size $O(n/r)$ for r-additive distance labeling with $r \geq 2$ and a lower bound of $\Omega(\sqrt{n/r})$ was given by Gavoille et al. [21]. For $r = 1$ a lower bound of $\Omega(n)$ can be established by observing that such a scheme can answer adjacency queries in bipartite graphs.

Distance preserving. An alternative to approximating all distances is to only report exact distances above some certain threshold D. A labeling scheme, which reports exact distances

[2] This does not break the Girth Conjecture, as the labeling scheme may under-estimate the distance as well.

for nodes u, v where $dist(u, v) \geq D$ is called a *D-preserving distance labeling scheme*[3]. Bollobás et al. [11] introduced this notion and gave a labeling scheme of size $O(\frac{n}{D} \log^2 n)$ for both directed and undirected graphs. They also provided an $\Omega(\frac{n}{D} \log D)$ lower bound for directed graphs.

1.1 Related work

A problem closely related to distance labeling is adjacency labeling. For some classes such as general graphs the best-known lower bounds for distance is actually that of adjacency. Adjacency labeling has been studied for various classes of graphs. In [8] the label size for adjacency in general undirected graphs was improved from $n/2 + O(\log n)$ [30, 33] to optimal size $n/2+O(1)$, and in [5] adjacency labeling for trees was improved from $\log_2 n+O(\log^* n)$ [9] to optimal size $\log_2 n + O(1)$.

Distance labeling schemes and related 2-hop labeling are used in SIGMOD and is central for some real-world applications [4, 17, 29]. Approximate distance labeling schemes have found applications in several fields such as reachability and distance oracles [37] and communication networks [35]. An overview of distance labeling schemes can be found in [7].

1.2 Our results

We address open problems of [7, 11, 24] improving the label sizes for *exact distances in sparse graphs*, *r-additive distance in general graphs*, and *D-preserving distance labeling*. We do this by showing a strong relationship between D-preserving distance labeling and several other labeling problems using D-preserving distance labels as a black box. Thus, by improving the result of [11] we are able to obtain the first sublinear labeling schemes for several problems studied at SODA over the past decades. Our results are summarized below.

Sparse graphs. We present the first sublinear distance labeling scheme for sparse graphs giving the following theorem:

▶ **Theorem 1.** *Let $\mathcal{S}_n$ denote the family of undirected and unweighted graphs on n nodes with at most $n^{1+o(1)}$ edges. Then there exists a distance labeling scheme for $\mathcal{S}_n$ with maximum label size $o(n)$.*

As noted, prior to this work the best-known bound for this family was the $O(n)$ scheme of [7] for general graphs. Thus, Theorem 1 separates the family of sparse graphs from the family of general graphs requiring $\Omega(n)$ label size. Our result uses a black-box reduction from sparse graphs to the D-preserving distance scheme of Theorem 3 below. The result of Theorem 1 was made available online in a preliminary version of this paper [6] and was subsequently slightly improved by Gawrychowski et al. [25] by noting, that one of the steps in the construction of our D-preserving distance scheme can be skipped when only considering sparse graphs[4].

Approximate labeling schemes. For r-additive distance labeling Gawrychowski et al. [25] showed that a sublinear labeling scheme for sparse graphs implies a sublinear labeling scheme

[3] In this paper we adopt the convention that the labeling scheme returns an upper-bound if $dist(u, v) < D$.

[4] The scheme presented in this paper has labels of length $O\left(\frac{n \operatorname{polylog} \Delta}{\Delta}\right)$, where $\Delta = \frac{\log n}{1+\log \frac{m+n}{n}}$. In [25] they improve the exponent of the $\operatorname{polylog} \Delta$ term from 2 to 1.

for $r = 2$ in general graphs. We generalise this result to $r \geq 2$ by a reduction to the D-preserving scheme. We note that a reduction to sparse graphs does not suffice in this case, and the scheme of [25] thus only works for $r = 2$. More precisely, we show the following:

▶ **Theorem 2.** *For any $r \geq 2$, there exists an approximate r-additive labeling schemes for the family $\mathcal{G}_n$ of undirected and unweighted graphs on n nodes with maximum label size*

$$O\left(\frac{n}{r} \cdot \frac{\operatorname{polylog}(r \log n)}{\log n}\right) .$$

Theorem 2 improves on the previous best bound of $O\left(\frac{n}{r}\right)$ by [7] whenever $r = 2^{o\left(\sqrt{\log n}\right)}$, e.g. when $r = \operatorname{polylog} n$.

D-preserving labeling schemes. For D-preserving labeling schemes we show that:

▶ **Theorem 3.** *For any integer $D \in [1, n]$, there exists a D-preserving distance labeling scheme for the family $\mathcal{G}_n$ of undirected and unweighted graphs on n nodes with maximum label size*

$$O\left(\frac{n}{D} \max\left\{\log^2 D, 1\right\}\right) .$$

Theorem 3 improves the result of [11] by a factor of $O(\log^2 n / \log^2 D)$ giving the first sublinear size labels for this problem for any $D = \omega(1)$. This sublinearity is the main ingredient in showing the results of Theorems 1 and 2. Our scheme uses sampling similar to that of [11]. By sampling fewer nodes we show that not "too many" nodes end up being problematic and handle these separately by using a tree structure similar to [7][5].

Finally, we give an almost matching lower bound showing:

▶ **Theorem 4.** *A D-preserving distance labeling scheme for the family $\mathcal{G}_n$ of undirected and unweighted graphs on n nodes require label size $\Omega(\frac{n}{D})$, when D is an integer in $[1, n-1]$.*

This bound is a slight modification of the $\Omega(\frac{n}{D} \log D)$ lower bound for directed graphs given in [11].

2 Preliminaries

Throughout the paper we adopt the convention that $\lg x = \max(\log_2 x, 1)$ and $\log x = \ln x$. When $x \leq 0$ we define $\lg x = 1$. In this paper we assume the word-RAM model, with word size $w = \Theta(\log n)$. If s is a bitstring we denote its length by $|s|$ and will also use s to denote the integer value of s when this is clear from context. We use $s \circ s'$ to denote concatenation of bit strings. Finally, we use the Elias γ code [18] to encode a bitstring s of unknown length using $2|s|$ bits such that we may concatenate several such bitstrings and decode them again.

Labeling schemes. A *distance labeling scheme* for a family of graphs $\mathcal{G}$ consists of an encoder e and a decoder d. Given a graph $G \in \mathcal{G}$ the encoder computes a *label assignment* $e_G : V(G) \to \{0,1\}^*$, which assigns a *label* to each node of G. The decoder is a function such that given any graph $G \in \mathcal{G}$ and any pair of nodes $u, v \in V(G)$ we have $d(e_G(u), e_G(v)) =$

[5] We note that after making this result available online in a preliminary version [6], the bound of Theorem 3 was slightly improved by Gawrychowski et al. [25] to $O(\frac{n}{D} \log D)$.

$dist_G(u,v)$. Note that the decoder is oblivious to the actual graph G and is only given the two labels $e_G(u)$ and $e_G(v)$.

The *size* of a labeling scheme is defined as the maximum label size $|e_G(u)|$ over all graphs $G \in \mathcal{G}$ and all nodes $u \in V(G)$. If for all graphs $G \in \mathcal{G}$ the mapping e_G is injective we say that the labeling scheme assigns *unique labels* (note that two different graphs $G, G' \in \mathcal{G}$ may share a label).

If the encoder and graph is clear from the context, we will sometimes denote the label of a node u by $\ell(u) = e_G(u)$.

Various computability requirements are sometimes imposed on labeling schemes [1, 30, 31].

3 D-preserving distance labeling schemes

In this section we will prove Theorem 3. Observe first that for $D = 1$ Theorem 3 is exactly the classic problem of distance labeling and we may use the result of [7]. We will therefore assume that $D \geq 2$ for the remainder of this paper. Let us first formalize the definition of a D-preserving distance labeling scheme.

▶ **Definition 5.** Let D be a positive integer let $\mathcal{G}$ be a family of graphs. For each graph $G \in \mathcal{G}$ let $e_G : V(G) \to \{0,1\}^*$ be a mapping of nodes to labels. Let $d : \{0,1\}^* \times \{0,1\}^* \to \mathbb{Z}$ be a decoder. If e and d satisfy the following two properties, we say that the pair (e,d) is a *D-distance preserving labeling scheme* for the graph family $\mathcal{G}$.

1. $d(e_G(u), e_G(v)) \geq dist_G(u,v)$ for all $u, v \in G$ for any $G \in \mathcal{G}$.
2. $d(e_G(u), e_G(v)) = dist_G(u,v)$ for all $u, v \in G$ with $dist_G(u,v) \geq D$ for any $G \in \mathcal{G}$.

The idea of the labeling scheme presented in this section is to first make a labeling scheme for distances in the range $[D, 2D]$ and use this scheme for increasingly bigger distances until all distances of at least D are covered. Loosely speaking, the scheme is obtained by sampling a set of nodes R, such that *most* shortest paths of length at least D contain a node from R. Then all nodes are partitioned into *sick* and *healthy* nodes adding the sick nodes to the set R. All nodes then store their distance to each node of R and healthy nodes will store the distance to all nodes, for which the shortest path is not *covered* by some node in R.

3.1 A sample-based approach

As a warm-up, we first present the $O\left(\frac{n}{D}\log^2 n\right)$ scheme of Bollobás et al. in [11] with a slight modification.

Given a graph $G = (V,E) \in \mathcal{G}$ we pick a random multiset $R \subseteq V$ consisting of $\left\lceil c \cdot \frac{n}{D} \log n \right\rceil$ nodes for a constant c to be decided. Each element of R is picked uniformly and independently at random from V (i.e. the same node might be picked several times)[6]. We order R arbitrarily as $(w_1, \ldots, w_{|R|})$ and assign the label of a node $u \in V$ as

$$\ell(u) = dist_G(u, w_1) \circ dist_G(u, w_2) \circ \ldots \circ dist_G(u, w_{|R|})$$

▶ **Lemma 6.** *Let u and v be two nodes of some graph $G \in \mathcal{G}$. Set*

$$d = \min_{w \in R} dist_G(u,w) + dist_G(v,w) \ . \tag{1}$$

Then $d \geq dist_G(u,v)$ and $d = dist_G(u,v)$ if R contains a node from a shortest path between u and v.

[6] In [11] they instead picked R by including each node of G with probability $\frac{c \log n}{D}$.

Proof. Let $z \in R$ be the node corresponding to the minimum value of (1). We then have $d = dist_G(u, z) + dist_G(z, v)$. By the triangle inequality this implies $d \geq dist(u, v)$.

Now let p be some shortest path between u and v in G and assume that $z \in p$. Then $dist_G(u, v) = dist_G(u, z) + dist_G(z, v)$, implying that $d \leq dist_G(u, v)$, and thus $d = dist(u, v)$. ◀

By Lemma 6 it only remains to show that the set R is likely to contain a node on a shortest path between any pair of nodes $u, v \in V$ with $dist_G(u, v) \geq D$.

▶ **Lemma 7.** *Let R be defined as above. Then the probability that there exists a pair of nodes $u, v \in V$ such that $dist_G(u, v) \geq D$ and no node on the shortest path between u and v is sampled is at most n^{2-c}.*

Proof. Consider a pair of nodes $u, v \in V$ with $dist_G(u, v) \geq D$. Let p be a shortest path between u and v, then $|p| \geq D$. Each element of R has probability at least D/n of belonging to p (independently), so the probability that no element of R belonging to p is at most

$$\left(1 - \frac{D}{n}\right)^{|R|} \leq \exp\left(-\frac{D}{n} \cdot |R|\right) \leq \exp(-c \log n) = n^{-c} \ . \tag{2}$$

Since there are at most n^2 such pairs, by a union bound the probability that there exists a pair u, v with $dist_G(u, v) \geq D$, such that no element on a shortest path between u and v is sampled in R is thus at most $n^2 \cdot n^{-c} = n^{2-c}$ ◀

By setting $c > 2$ we can ensure that the expected number of times we have to re-sample the set R until the condition of Lemma 7 is satisfied is $O(1)$. The labels can be assigned using $O(|R| \log n) = O(\frac{n}{D} \log^2 n)$ bits as each distance can be stored using $O(\log n)$ bits.

3.2 A scheme for medium distances

We now present a scheme, which preserves distances in the range $[D, 2D]$ using $O\big(\frac{n}{D} \log^2 D\big)$ bits. More formally, we present a labeling scheme such that given a family of unweighted undirected graphs $\mathcal{G}$ the encoder and the decoder satisfies the following constraints for any $G \in \mathcal{G}$:

1. $d(e_G(u), e_G(v)) \geq dist_G(u, v)$ for any $u, v \in G$.
2. $d(e_G(u), e_G(v)) = dist_G(u, v)$ for any $u, v \in G$ with $dist_G(u, v) \in [D, 2D]$.

Let such a labeling scheme be called a $[D, 2D]$-*preserving distance labeling scheme.*

The labeling scheme is based on a sampling procedure similar to that presented in Section 3.1, but improves the label size by introducing the notion of *sick* and *healthy* nodes.

Let $G = (V, E) \in \mathcal{G}$. We sample a multiset R of size $2 \cdot \frac{n}{D} \log D$. Similar to Section 3.1, each element of R is picked uniformly at random from V.

▶ **Definition 8.** Let R be as defined above and fix some node u. We say that a node v is *uncovered* for u if $dist_G(u, v) \geq D$ and no node in R is contained in a shortest path between u and v. A node u with more than $\frac{n}{D}$ uncovered nodes is called *sick* and all other nodes are called *healthy.*

Let S denote the set of sick nodes and let $\mathrm{uc}(u)$ denote the set of uncovered nodes for u. The main outline of the scheme is as follows:

1. Each node u stores the distance from itself to each node of $R \cup S$ using a tree structure to be described.

2. If u is healthy, u stores the distance from itself to every $v \in \text{uc}(u)$ for which $dist_G(u,v) \in [D, 2D]$.

We start by showing that the set of sick nodes has size $O(n/D)$ with probability at least $1/2$. This is captured by the following lemma.

▶ **Lemma 9.** *Let R be defined as above and let S be the set of sick nodes. Then*

$$\Pr\left[|S| \geq 2\frac{n}{D}\right] \leq 1/2 \ .$$

Proof. Fix some node $u \in V$ and let $v \in V$ be a node such that $dist_G(u,v) \geq D$. Using the same argument as in (2) of Lemma 7 we see that the probability that v is uncovered for u is at most D^{-2}. Therefore $\mathbf{E}[|\text{uc}(u)|] \leq \frac{n}{D^2}$. By Markov's inequality we have

$$\Pr[u \text{ is sick}] = \Pr\left[|\text{uc}(u)| \geq D \cdot \frac{n}{D^2}\right] \leq \frac{1}{D} \ ,$$

and thus $\mathbf{E}[|S|] \leq n/D$. We again use Markov's inequality to conclude that

$$\Pr\left[|S| \geq 2\frac{n}{D}\right] \leq 1/2 \, . \qquad \blacktriangleleft$$

The goal is now to store the distances to the nodes of $R \cup S$ as well as $\text{uc}(u)$ using few bits. First consider the distances to $R \cup S$. We will store these distances using a tree structure similar to that of [7]. To do this we will use the following algorithm:

1. Let $r \in V$ be an arbitrary node.
2. Let T' be the BFS-tree of r in G rooted in r.
3. For $i \in \{0, \ldots, D-1\}$, let $A_i = \{u \in V \mid dist_G(r,u) \equiv i \mod D\}$.
4. Let $j = \arg\min_{i \in \{0,\ldots,D-1\}} |A_i|$.
5. Let T be a graph with $V(T) = A_j \cup \{r\} \cup R \cup S$ and $E(T) = \emptyset$.
6. For each $u \in V(T) \setminus \{r\}$ let v be the nearest ancestor of u in $T' \setminus \{u\}$ such that $v \in V(T)$. Add the edge (v,u) to T with weight $dist_G(v,u) = dist_{T'}(v,u)$.

This process is illustrated in Figure 1.

▶ **Lemma 10.** *Let T be the tree created by the algorithm above. Then T contains $O(\frac{n}{D} \log D)$ nodes with probability at least $1/2$ and each edge of T has weight at most D.*

Proof. Let T', r and A_j be as defined in the algorithm above.

The size of T is at most $|S| + |R| + |A_j| + 1$. By our choice of A_j and R this is bounded by

$$|S| + 2 \cdot \frac{n}{D} \log D + \left\lfloor \frac{n}{D} \right\rfloor + 1 \ .$$

Using Lemma 9 we see that this is $O\big(\frac{n}{D} \log D\big)$ with probability at least $1/2$.

Consider now any edge $(u, p(u)) \in E(T)$ and let $d = dist_G(u,r)$. If $d \leq D$ it follows from the definition of T that $dist_G(u,p(u)) \leq D$, as r is an ancestor of all nodes, and thus also u, in T'. If $d > D$ consider the unique path from u to r in T' and denote the nodes on this path as $(u, v_1, v_2, \ldots, v_k, r)$. It follows that $dist_G(v_1, r) = d - 1$, $dist_G(v_2, r) = d - 2$, etc. Since $d > D$ we have $k \geq D$ and thus one of $v_1, \ldots, v_D$ is contained in the set A_j and has distance at most D to u. It now follows that $dist_G(u,p(u)) \leq D$ and thus $dist_T(u,p(u)) \leq D$. $\blacktriangleleft$

Using Lemma 10 we are able to store the distance from any node u to all nodes of T by storing the differences between the distance from u to adjacent nodes in T. This is captured in the following lemma:

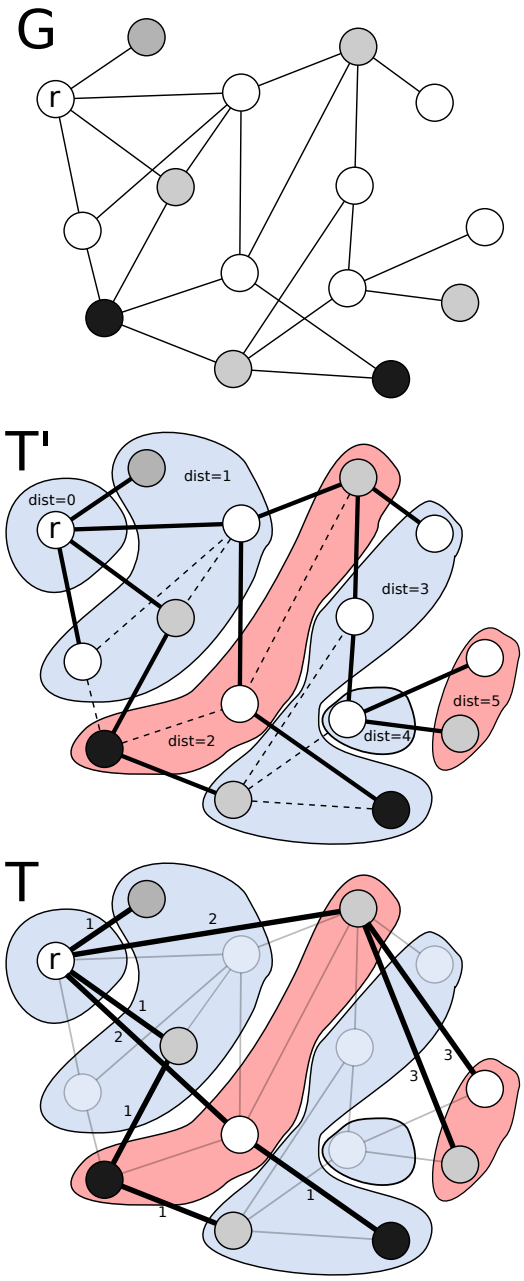

Figure 1 The process of creating T as described above. Gray nodes are the sampled nodes, R, and black nodes are the sick nodes, S. We assume $D = 3$ and pick A_2 as the smallest set (marked in red). *Note that the black nodes are only for illustration and might not actually be sick by our definition.*

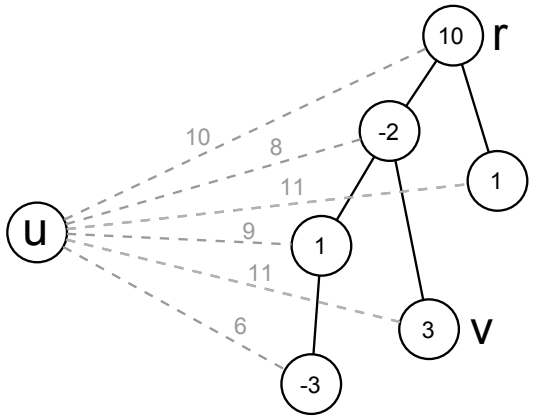

Figure 2 Storing the tree $T = (V', E')$ using few bits. For each node $u \in V'$, we store $dist(u, v) - dist(u, p(v))$. Shortest path distances from u in G are denoted in gray. The distance from u to v is calculated as $10 + (-2) + 1 + (-3) - (-3) - 1 + 3 = 11$.

▶ **Lemma 11.** *Let u be some node in $G = (V, E)$ and let $T = (V', E')$ be the tree resulting from the algorithm above rooted in r. Then we can store the distance from u to every node in T using $O\left(\frac{n}{D} \log^2 D\right)$ bits.*

Proof. Consider the following encoding: We fix some canonical DFS ordering of T and describe it using $2|T|$ bits. This will be the same for all nodes $u \in V$. Next, we store $dist_G(u, r)$ using $\lceil \lg n \rceil$ bits. For each node $v \in V' \setminus \{r\}$ taken in the DFS ordering of T we store $dist_G(u, v) - dist_G(u, p(v))$. Using this description, we can calculate $dist_G(u, v)$ for any $v \in V'$ by summing up the differences on the path from v to r and adding the distance from u to r.

We now argue that $dist_G(u, v) - dist_G(u, p(v))$ can be stored using $\lceil \lg(2D + 1) \rceil$ bits for any node $v \in V' \setminus \{r\}$. Set $t = dist_G(u, p(v))$. By Lemma 10 and the triangle inequality it holds that

$$dist_G(u, v) \leq dist_G(u, p(v)) + dist_G(p(v), v) \leq t + D \ .$$

Similarly,

$$dist_G(u, v) \geq dist_G(u, p(v)) - dist_G(v, p(v)) \geq t - D \ .$$

Thus, it follows that

$$dist_G(u, v) - dist_G(u, p(v)) \in \{-D, \ldots, 0, \ldots, D\} \ ,$$

which can be stored using $\lceil \lg(2D + 1) \rceil$ bits. We can thus store all the information using

$$O(2|T| + \log n + |T| \log D) = O\left(\frac{n}{D} \log^2 D\right)$$

bits. ◀

The values $dist_G(u, v) - dist_G(u, p(v))$ are illustrated in Figure 2.

We may now assign the label $\ell(u)$ of a node u to be $id(u)$ concatenated with the bitstring resulting for Lemma 11 and if u is healthy this is concatenated with the id of the nodes in $uc(u)$ whose distance from u is in the interval $[D, 2D]$ along with these distances. The decoder works by simply checking if one nodes stores the others distance or by taking the minimum of going via any node in T.

Label size. In order to bound the size of the label we only need to bound the size of storing id's and distances to the nodes of $uc(u)$ whose distance is in $[D, 2D]$. Since we only store this for healthy nodes this set has size at most n/D and can be described using at most $O(\frac{n}{D} \log D)$ bits. Since each distance can be stored using $O(\log D)$ bits we conclude that the total label size is bounded by $O\big(\frac{n}{D} \log^2 D\big)$.

▶ **Theorem 12.** *There exists a $[D, 2D]$-preserving distance labeling scheme for the family $\mathcal{G}_n$ of undirected and unweighted graphs on n nodes with maximum label size*

$$O\Big(\frac{n}{D} \log^2 D\Big) \ .$$

Proof. This is a direct corollary of the discussion above. ◀

3.3 Bootstrapping the scheme

In order to show Theorem 3 we will concatenate several instances of the label from Theorem 12. First define $\ell_D(u)$ to be the $[D, 2D]$-preserving distance label for the node u assigned by the scheme of Theorem 12. Now assign the following label to each node u:

$$\ell(u) = \ell_D(u) \ \circ \ \ell_{2D}(u) \ \circ \ \ell_{4D}(u) \ \circ \ \ldots \ \circ \ \ell_{2^k D}(u) \ , \tag{3}$$

where $k = \lfloor \lg(n/D) \rfloor$. Let d_D be the distance returned by running the decoder of Theorem 12 on the corresponding component of the label $\ell(u)$. Then we let the decoder of the full labeling scheme return

$$\hat{d} = \min(d_D, d_{2D}, \ldots, d_{2^k D}) \ , \tag{4}$$

with k defined as above. We are now ready to prove Theorem 3.

Proof of Theorem 3. Consider any pair of nodes u, v in some graph $G \in \mathcal{G}_n$ and let $d = dist_G(u, v)$. Also, let $\hat{d}$ be the value returned by the decoder for $\ell(u)$ and $\ell(v)$. If $d \leq D$ we have $\hat{d} \geq d$. Now assume that $d \in [2^i \cdot D, 2^{i+1} \cdot D]$ for some non-negative integer i. Then, by Theorem 12 and (4) we have $\hat{d} = d$.

The size of the label assigned by (3) is bounded by

$$\begin{aligned}
\sum_{i=0}^{\lfloor \lg_2(n/D) \rfloor} O\Big(\frac{n}{2^i \cdot D} \log^2(2^i \cdot D)\Big) &\leq \sum_{i=0}^{\infty} O\Big(\frac{n}{2^i \cdot D} \log^2(2^i \cdot D)\Big) \\
&\leq O\left(\frac{n}{D} \log^2(D) \sum_{i=1}^{\infty} \frac{i^2 + 1}{2^i}\right) \\
&= O\Big(\frac{n}{D} \log^2(D)\Big) \ .
\end{aligned}$$

◀

3.4 Lower bound

Proof of Theorem 4. Let $k = \left\lfloor \frac{n}{D+1} \right\rfloor$ and let L and R be sets of k nodes which make up the left and right side of a bipartite graph respectively. Furthermore, let each node of R be the first node on a path of D nodes.

Consider now the family of all such bipartite graphs (L, R) with the attached paths. There are exactly 2^{k^2} such graphs.

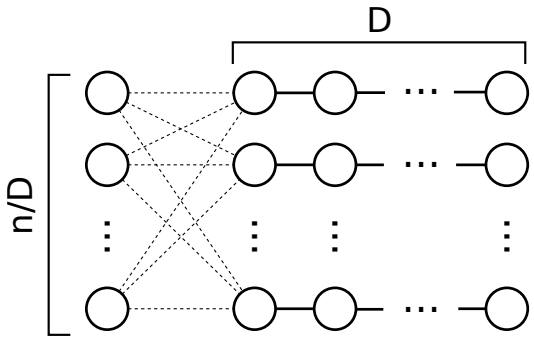

Figure 3 Illustration of the graph family used in the proof of Section 3.4.

Now observe, that a node $u \in L$ is adjacent to a node $v \in R$ if and only if $dist(u, w) = D$, where w is the last node on the path starting in v. By querying all such pairs (u, w) we obtain k^2 bits of information using only $2k$ labels, thus at least one label of size

$$\frac{k^2}{2k} = \frac{\left\lfloor \frac{n}{D+1} \right\rfloor}{2} \geq \frac{n}{8D}$$

is needed. Since the graph has $\leq n$ nodes this implies the result. ◀

This is illustrated in Figure 3.

4 Sparse and bounded degree graphs

We are now ready to prove Theorem 1. In fact we will show the following more general lemma:

▶ **Lemma 13.** *Let $\mathcal{H}_{n,m}$ denote the family of undirected and unweighted graphs on n nodes with at most m edges. Then there exists a distance labeling scheme for $\mathcal{H}_{n,m}$ with maximum label size*

$$O\Big(\frac{n}{D} \cdot \log^2 D\Big), \text{ where } D = \frac{\log n}{1 + \log \frac{m+n}{n}}$$

Since $\frac{\log n}{1+\log \frac{m+n}{n}} = \omega(1)$ when $m = n^{1+o(1)}$ it will suffice to prove Lemma 13. In order to do so we first show the following lemma for bounded-degree graphs:

▶ **Lemma 14.** *Let $\mathcal{B}_n(\Delta)$ be the family of graphs on n nodes with maximum degree Δ. There exists a distance labeling scheme for $\mathcal{B}_n(\Delta)$ with maximum label size*

$$O\Big(\frac{n}{D} \log^2 D\Big), \text{ where } D = \frac{\log n}{1 + \log \Delta}$$

Proof. Suppose we are labeling some graph $G \in \mathcal{B}_n(\Delta)$ and let $u \in G$. Let $D = \left\lceil \frac{\log n}{1+2\log \Delta} \right\rceil$ and let $\ell_D(u)$ be the D-distance preserving label assigned by using Theorem 3 with parameter D. Using this label we can deduce the distance to all nodes of distance at least D to u.

Since $G \in \mathcal{B}_n(\Delta)$ there are at most $\Delta^D = O(\sqrt{n})$ nodes closer than distance D to u. Thus, we may describe the IDs and distances of these nodes using at most $O(\sqrt{n} \log n)$ bits. This gives the desired total label size of

$$|\ell(u)| = O\Big(\sqrt{n} \log n + \frac{n}{D} \log^2 D\Big) = O\Big(\frac{n}{D} \log^2 D\Big) . \quad ◀$$

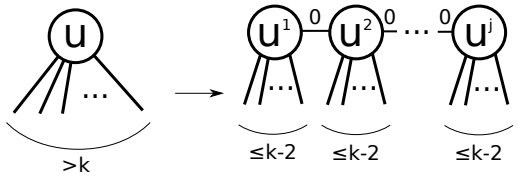

Figure 4 Illustration of the transformation from sparse graph to bounded degree graph.

Using this result we may now prove Lemma 13 by reducing to the bounded degree case in Lemma 14. This has been done before e.g. in distance oracles [20, 3].

Proof of Lemma 13. Let $G \in \mathcal{H}_{n,m}$ be some graph and let $k = \max\left\{\left\lceil \frac{m}{n} \right\rceil, 3\right\}$. Let $u \in G$ be some node with more than k incident edges. If no such node exists, we may apply Lemma 14 directly and we are done. Otherwise we split u into $\lceil \deg(u)/(k-2) \rceil$ nodes and connect these nodes with a path of 0-weight edges. Denote these nodes $u^1, \ldots, u^{\lceil \deg(u)/(k-2) \rceil}$. For each edge (u, v) in G we assign the end-point at u to a node u^i with $\deg(u^i) < k$. This process is illustrated in Figure 4.

Let the graph resulting from performing this process for every node $u \in G$ be denoted by G'. We then have $\Delta(G') \leq k$. Furthermore it holds that for every pair of nodes $u, v \in G$ we have $dist_G(u, v) = dist_{G'}(u^1, v^1)$. Consider now using the labeling scheme of Lemma 14 on G' and setting $\ell(u) = \ell(u^1)$ for each node $u \in G$. By observing that the labeling scheme of Theorem 3 preserves distances for nodes who have at least D *edges* on the shortest path we see that this is actually a distance labeling scheme for G. The number of nodes in G' is bounded by

$$\sum_{u \in G} \left\lceil \frac{\deg(u)}{k-2} \right\rceil \leq \sum_{u \in G} \left(\frac{\deg(u)}{k-2} + 1 \right) = \frac{2m}{k-2} + n = O(n) \ ,$$

which means that Lemma 14 gives the desired label size. ◀

5 Additive error

We will now show how we can use our D-preserving labeling scheme of Theorem 3 to generalize the 2-additive distance labeling scheme of Gawrychowski et al. [25]. We will assume that $r \leq n^{1/10}$ for simplicity.

Let $t = r \log^{10} n$ and let $D = \frac{r \log n}{4 \log t}$. We describe the scheme in three parts:

1. Let G^r be a copy of G, where an edge is added between any pair of nodes whose distance is at most $r/2$ in G. Let $V^r_{\geq t}$ be the set of nodes in G^r with degree at least t and let S be a minimum dominating set of $V^r_{\geq t}$ in G_r. Then $|S| = O(\frac{n \log t}{t})$.

   For all nodes $u \in G$ we store $dist(u, v)$ and $id(v)$ for all $v \in S$.
2. Consider now the subgraph of G induced by $V \setminus V^r_{\geq t}$. For a node $u \notin V^r_{\geq t}$, let $B_u(D)$ be the ball of radius D around u in this induced subgraph Then $|B_u(D)| \leq t^{2D/r} = O(\sqrt{n})$. This follows from the definition of $V^r_{\geq t}$: There are at most t nodes within distance $r/2$ from u and thus at most t^2 nodes within distance r from u, etc.

   For all $u \notin V^r_{\geq t}$ we store $dist(u, v)$ and $id(v)$ for all $v \in B_u(D)$.
3. Finally we store a D-preserving distance label for all $u \in G$.

The total label size is then

$$O\left(n \cdot \frac{\log n \log t}{t} + \sqrt{n} \log n + \frac{n \log t}{r \log n} \cdot (\log(r \log n))^2\right) = O\left(\frac{n}{r \log n} \cdot \text{polylog}(\log n \cdot r)\right) ,$$

as stated in Theorem 2.

Decoding. To see that the distance between two nodes u and v can be calculated within an additive error r we split into several cases:

- If $dist(u, v) \geq D$ we can report the exact distance between u and v using the D-preserving distance scheme.
- If $dist(u, v) \leq D$ and $deg_{G_r}(v) \geq t$ we can find a node $z \in S$ such that $dist(z, v) \leq r/2$ and thus

$$dist(u, z) + dist(z, v) \leq dist(u, v) + dist(v, z) + dist(z, v) \leq dist(u, v) + r ,$$

  and symmetrically if $deg_{G_r}(u) \geq t$.
- Finally, if $dist(u, v) \leq D$ and $deg_{G_r}(u) < t$ and $deg_{G_r}(v) < t$, then we $v \in B_u(D)$ and we can thus report the exact distance between u and v.

Acknowledgements. We would like to thank Noy Rotbart for helpful discussions and observations.

References

1. S. Abiteboul, H. Kaplan, and T. Milo. Compact labeling schemes for ancestor queries. In *Proc. of the 12th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 547–556, 2001.
2. I. Abraham, S. Chechik, and C. Gavoille. Fully dynamic approximate distance oracles for planar graphs via forbidden-set distance labels. In *Proc. 44th Annual ACM Symp. on Theory of Computing (STOC)*, pages 1199–1218, 2012.
3. R. Agarwal, P. B. Godfrey, and S. Har-Peled. Approximate distance queries and compact routing in sparse graphs. In *INFOCOM 2011. 30th IEEE International Conference on Computer Communications*, pages 1754–1762, 2011.
4. T. Akiba, Y. Iwata, and Y. Yoshida. Fast exact shortest-path distance queries on large networks by pruned landmark labeling. In *ACM International Conference on Management of Data (SIGMOD)*, pages 349–360, 2013. doi:10.1145/2463676.2465315.
5. S. Alstrup, S. Dahlgaard, and M. B. T. Knudsen. Optimal induced universal graphs and labeling schemes for trees. In *Proc. 56th Annual Symp. on Foundations of Computer Science (FOCS)*, 2015.
6. S. Alstrup, S. Dahlgaard, M. B. T. Knudsen, and E. Porat. Sublinear distance labeling for sparse graphs. *CoRR*, abs/1507.02618, 2015. URL: http://arxiv.org/abs/1507.02618.
7. S. Alstrup, C. Gavoille, E. B. Halvorsen, and H. Petersen. Simpler, faster and shorter labels for distances in graphs. In *Proc. 27th Annual ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 338–350, 2016.
8. S. Alstrup, H. Kaplan, M. Thorup, and U. Zwick. Adjacency labeling schemes and induced-universal graphs. In *Proc. of the 47th Annual ACM Symp. on Theory of Computing (STOC)*, 2015.
9. S. Alstrup and T. Rauhe. Small induced-universal graphs and compact implicit graph representations. In *Proc. 43rd Annual Symp. on Foundations of Computer Science (FOCS)*, pages 53–62, 2002.

10 F. Bazzaro and C. Gavoille. Localized and compact data-structure for comparability graphs. *Discrete Mathematics*, 309(11):3465–3484, 2009. doi:10.1016/j.disc.2007.12.091.

11 B. Bollobás, D. Coppersmith, and M. Elkin. Sparse distance preservers and additive spanners. *SIAM J. Discrete Math.*, 19(4):1029–1055, 2005. See also SODA'03. doi:10.1137/S0895480103431046.

12 M. A. Breuer. Coding the vertexes of a graph. *IEEE Trans. on Information Theory*, IT–12:148–153, 1966.

13 M. A. Breuer and J. Folkman. An unexpected result on coding vertices of a graph. *J. of Mathemathical analysis and applications*, 20:583–600, 1967.

14 V. D. Chepoi, F. F. Dragan, B. Estellon, M. Habib, and Y. Vaxès. Diameters, centers, and approximating trees of delta-hyperbolic geodesic spaces and graphs. In 24^{th} *Annual ACM Symp. on Computational Geometry (SoCG)*, pages 59–68, 2008. doi:10.1145/1377676.1377687.

15 V. D. Chepoi, F. F. Dragan, and Y. Vaxès. Distance and routing labeling schemes for non-positively curved plane graphs. *J. of Algorithms*, 61(2):60–88, 2006. doi:10.1016/j.jalgor.2004.07.011.

16 B. Courcelle and R. Vanicat. Query efficient implementation of graphs of bounded clique-width. *Discrete Applied Mathematics*, 131:129–150, 2003. doi:10.1016/S0166-218X(02)00421-3.

17 D. Delling, A. V. Goldberg, R. Savchenko, and R. F. Werneck. Hub labels: Theory and practice. In 13^{th} *International Symp. on Experimental Algorithms (SEA)*, pages 259–270, 2014. doi:10.1007/978-3-319-07959-2_22.

18 P. Elias. Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory*, 21(2):194–203, 1975.

19 M. Elkin, A. Filtser, and O. Neiman. Prioritized metric structures and embedding. In *Proc. of the 47th Annual ACM Symp. on Theory of Computing (STOC)*, pages 489–498, 2015.

20 M. Elkin and S. Pettie. A linear-size logarithmic stretch path-reporting distance oracle for general graphs. In *Proc. of the 26th Annual Symp. on Discrete Algorithms (SODA)*, pages 805–821, 2015.

21 C. Gavoille, M. Katz, N. A. Katz, C. Paul, and D. Peleg. Approximate distance labeling schemes. In *Proc. of the 9th annual European Symp. on Algorithms (ESA)*, pages 476–488, 2001.

22 C. Gavoille and C. Paul. Distance labeling scheme and split decomposition. *Discrete Mathematics*, 273(1-3):115–130, 2003.

23 C. Gavoille and C. Paul. Optimal distance labeling for interval graphs and related graphs families. *SIAM J. Discrete Math.*, 22(3):1239–1258, 2008. doi:10.1137/050635006.

24 C. Gavoille, D. Peleg, S. Pérennes, and R. Raz. Distance labeling in graphs. *J. of Algorithms*, 53(1):85–112, 2004. See also SODA'01. doi:10.1016/j.jalgor.2004.05.002.

25 P. Gawrychowski, A. Kosowski, and P. Uznanski. Even simpler distance labeling for (sparse) graphs. *CoRR*, abs/1507.06240, 2015. URL: http://arxiv.org/abs/1507.06240.

26 R. L. Graham and H. O. Pollak. On embedding graphs in squashed cubes. In *Lecture Notes in Mathematics*, volume 303. Springer-Verlag, 1972.

27 A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *44th Annual Symp. on Foundations of Computer Science (FOCS)*, pages 534–543, 2003. doi:10.1109/SFCS.2003.1238226.

28 A. Gupta, A. Kumar, and R. Rastogi. Traveling with a pez dispenser (or, routing issues in mpls). *SIAM J. on Computing*, 34(2):453–474, 2005. See also FOCS'01.

29 R. Jin, N. Ruan, Y. Xiang, and V. Lee. A highway-centric labeling approach for answering distance queries on large sparse graphs. In *ACM International Conference on Management of Data (SIGMOD)*, pages 445–456, May 2012. doi:10.1145/2213836.2213887.

30 S. Kannan, M. Naor, and S. Rudich. Implicit representation of graphs. *SIAM J. Disc. Math.*, pages 596–603, 1992. See also STOC'88.

31 M. Katz, N. A. Katz, A. Korman, and D. Peleg. Labeling schemes for flow and connectivity. *SIAM J. Comput.*, 34(1):23–40, 2004. See also SODA'02. doi:10.1137/S0097539703433912.

32 R. Krauthgamer and J. R. Lee. Algorithms on negatively curved spaces. In *47th Annual Symp. on Foundations of Computer Science (FOCS)*, pages 119–132, 2006. doi:10.1109/FOCS.2006.9.

33 J. W. Moon. On minimal n-universal graphs. *Proc. of the Glasgow Mathematical Association*, 7(1):32–33, 1965.

34 J. H. Müller. *Local structure in graph classes*. PhD thesis, Georgia Institute of Technology, 1988.

35 D. Peleg. Proximity-preserving labeling schemes. *J. Graph Theory*, 33(3):167–176, 2000.

36 K. Talwar. Bypassing the embedding: algorithms for low dimensional metrics. In *Proc. of the 36th Annual ACM Symp. on Theory of Computing (STOC)*, pages 281–290, 2004. doi:10.1145/1007352.1007399.

37 M. Thorup. Compact oracles for reachability and approximate distances in planar digraphs. *J. ACM*, 51(6):993–1024, 2004. See also FOCS'01. doi:10.1145/1039488.1039493.

38 M. Thorup and U. Zwick. Approximate distance oracles. *J. of the ACM*, 52(1):1–24, 2005. See also STOC'01.

39 P. M. Winkler. Proof of the squashed cube conjecture. *Combinatorica*, 3(1):135–139, 1983. doi:10.1007/BF02579350.