

# Hardness of exact distance queries in sparse graphs through hub labeling

Adrian Kosowski<sup>1</sup>, Przemysław Uznański<sup>2</sup>, and Laurent Viennot<sup>1</sup>

<sup>1</sup>Inria, Paris University, France\*

<sup>2</sup>Institute of Computer Science, University of Wrocław, Poland<sup>†</sup>

## Abstract

A *distance labeling scheme* is an assignment of bit-labels to the vertices of an undirected, unweighted graph such that the distance between any pair of vertices can be decoded solely from their labels. An important class of distance labeling schemes is that of *hub labelings*, where a node  $v \in G$  stores its distance to the so-called hubs  $S_v \subseteq V$ , chosen so that for any  $u, v \in V$  there is  $w \in S_u \cap S_v$  belonging to some shortest  $uv$  path. Notice that for most existing graph classes, the best distance labelling constructions existing use at some point a hub labeling scheme at least as a key building block.

Our interest lies in hub labelings of sparse graphs, i.e., those with  $|E(G)| = \mathcal{O}(n)$ , for which we show a lowerbound of  $\frac{n}{2^{\mathcal{O}(\sqrt{\log n})}}$  for the average size of the hubsets. Additionally, we show a hub-labeling construction for sparse graphs of average size  $\mathcal{O}(\frac{n}{\text{RS}(n)^c})$  for some  $0 < c < 1$ , where  $\text{RS}(n)$  is the so-called Ruzsa-Szemerédi function, linked to structure of induced matchings in dense graphs. This implies that further improving the lower bound on hub labeling size to  $\frac{n}{2^{(\log n)^{o(1)}}$  would require a breakthrough in the study of lower bounds on  $\text{RS}(n)$ , which have resisted substantial improvement in the last 70 years.

For general distance labeling of sparse graphs, we show a lowerbound of  $\frac{1}{2^{\Theta(\sqrt{\log n})}} \text{SUMINDEX}(n)$ , where  $\text{SUMINDEX}(n)$  is the communication complexity of the Sum-Index problem over  $Z_n$ . Our results suggest that the best achievable hub-label size and distance-label size in sparse graphs may be  $\Theta(\frac{n}{2^{(\log n)^c}})$  for some  $0 < c < 1$ .

## 1 Introduction

Hub labeling schemes are a popular way of distributed encoding shortest path structure for easy retrieval. Given a graph  $G = (V, E)$  (across this paper we assume, unless stated otherwise, that  $G$  is undirected and unweighted, with  $|V| = n$  and  $|E| = m$ ), we store with each vertex  $v \in V(G)$  the so-called hubset  $S_v \subseteq V(G)$  together with shortest path distances between  $v$  and its hubs. The distance query  $uv$  is resolved by returning

$$\min_{w \in S(u) \cap S(v)} \text{dist}(u, w) + \text{dist}(w, v).$$

This framework was introduced in [CHKZ03] under the name of 2-hop covers, further explored by [AG11] (as landmark labelings) and [ADGW12]. The computed distance between all pairs of nodes  $u$  and  $v$  is exact

\*Supported by Irif CNRS laboratory and ANR projects DESCARTES (ANR-16-CE40-0023), DISTANCIA (ANR-17-CE40-0015) and MULTIMOD (ANR-17-CE22-0016).

<sup>†</sup>Supported by Polish National Science Centre grant 2016/22/E/ST6/00499.

if set  $S(u) \cap S(v)$  contains at least one node on some shortest  $u - v$  path. This property of the family of sets  $(S(u) : u \in V)$  is known as *shortest path cover*.

Hub labeling schemes are a special case of the more general *distance labeling* schemes, in which the task is the assignment of a binary string  $\text{label}(u)$  to each node  $u \in V$ , so that the graph distance between  $u$  and  $v$  is uniquely determined by the pair of labels:  $\text{label}(u)$  and  $\text{label}(v)$ , using a decoding function specified as part of the distance labeling. A central question in the distributed computing literature consists in identifying, for a given graph class, the optimal size of a distance labeling, expressed in terms on a bound on the average or maximum size of the labels used, taken over all nodes of a graph, for graphs belonging to the considered graph class.

In general graphs, the distance labeling problem was first investigated by Graham and Pollak [GP72], who provided the first labeling scheme with labels of size  $\mathcal{O}(n)$ . A subsequent line of research contributed to reducing the *decoding time*, with results of [GPPR04, WP11] and finally [AGHP16a] presented a scheme with labels of size  $\frac{\log_2 3}{2}n + o(n)$  and  $\mathcal{O}(1)$  decoding time. This result is asymptotically tight: by a simple counting argument, the label size of a node has to be at least  $\frac{1}{2}n - \mathcal{O}(1)$ , regardless of decoding time [GPPR04].

By contrast, when considering classes of *sparse graphs* (where  $m = \mathcal{O}(n)$ ), no similar asymptotically tight bounds on distance label size are known. The first sublinear-space distance labeling schemes were proposed only recently: [ADKP16] and later [GKU16] presented a hub labeling scheme with hubs of size  $\mathcal{O}(\frac{n}{\log n} \log \log n)$  and  $\mathcal{O}(\frac{n}{\log n})$ , respectively. Interpreted in terms of distance labeling schemes, these works provide  $\mathcal{O}(\frac{n}{\log n} (\log \log n)^2)$  and  $\mathcal{O}(\frac{n}{\log n} \log \log n)$  bits per label respectively, through some careful encoding of distances to vertices of hubsets. This leaves a substantial gap with respect to the best known lower bound of  $\Omega(\sqrt{n})$  for the bit size of distance labels and of hub labeling schemes per node, given in [GPPR04].

The objective of this paper is to help understand the source of hardness of distance labeling in sparse graphs. We show a basic obstacle on the path to the construction of better distance labelings: a lower bound of  $\frac{n}{2^{\mathcal{O}(\sqrt{\log n})}}$  on the hub set size in hub labeling. As we discuss further on in the related work section, this is an issue as all currently known approaches to distance labeling rely inherently (possibly implicitly) on the construction of some form of hub labels. More rigorously, we succeed in tying-in the distance label size to two long-standing open questions: one from combinatorics on the value of the so-called Ruzsa-Szemerédi function  $\text{RS}(n)$  on graphs, one from communication complexity on the bit complexity  $\text{SUMINDEX}(n)$  of a basic 3-party communication problem known as Sum-Index.

We remark that our results can perhaps help to shed further light on the question of designing (centralized) distance oracles for sparse graphs. A natural objective, so far unachieved in general, would be to obtain for  $n$ -node sparse graphs a spectrum of data structures, using space  $S$  and resolving exact distance queries in time  $T$ , with a time-space tradeoff of  $ST = \tilde{O}(n^2)$ . This tradeoff is trivially achieved at its endpoints ( $S = \tilde{O}(n)$ ,  $T = \tilde{O}(n^2)$ ), but the existence of such oracles appears open, e.g., for  $S = \tilde{O}(n^{3/2})$  and  $T = \tilde{O}(n^{1/2})$  (cf. [CP10, SVY09]). Our result precludes the existence of such a centralized oracle relying on an application of hub labeling.

## 1.1 Related work

**Distance labelings vs. Hub labelings.** It is noteworthy that for most existing graph classes, the best known distance labelling constructions are based on hub labeling schemes, either explicitly or as a key building block. Such constructions usually involve some form of compression and/or encoding of all distances (from a vertex to its hubs), to avoid  $\log n$  overhead when going from hubsets to binary labels (see e.g. [GKU16]).

For example, for the case of arbitrary graphs, a distance labeling may be constructed as follows: an additive approximation scheme for hub-labeling is constructed, that is for each pair  $uv$ , there is  $w \in S(u) \cap S(v)$  such that either  $w$  or some neighbor  $x \in N(w)$  is on shortest  $uv$  path. This guarantees that the absolute error of estimation is either 0, 1 or 2. Constructing such (small) approximate hub-set and complementing it with explicit correction tables (which require  $\frac{\log_2 3}{2}n$  bits per vertex) suffices. Many more ingredients and insights

are required to achieve constant decoding time as described in [AGHP16a], which goes beyond the scope of this overview.

**Distance labeling of sparse graphs and sub-classes.** Constructions for distance labeling in sparse graphs were considered in the previously mentioned papers [ADKP16] and [GKU16]. Both of them use hubsets as the underlying technique. The first of these constructions relies on the observation that, by selecting randomly a hubset  $S$  of size  $\mathcal{O}(\frac{n}{D} \log D)$  and assigning  $S \subseteq S(u)$  for any  $u$ , one already covers almost all pairs of vertices which are at distance  $D$  to each other, except at most a  $1/D$  fraction of these pairs. Selecting  $D = \Theta(\log n)$  and storing vertices closer than  $D$  as hubs explicitly leads to desired hubset sizes (however, a careful approach is required when dealing with large degree vertices which may appear in a sparse graph, guaranteed only to have constant average degree). Second work shaves  $\log \log n$  term by derandomizing the construction.

For the class of trees, the constructions of [Pel00] and [AGHP16b] are based on selection of central vertices as hubs, and proceeding recursively on obtained subtrees. In terms of bit size, those constructions give  $\Theta(\log^2 n)$  bits per label (which is asymptotically optimal due to lowerbound of [GPPR04]), which corresponds to  $\log n$  hubs per vertex. The construction of [FGNW17] through very careful assignment of pieces of information through the construction achieves  $\frac{1}{4} \log^2 n + o(\log^2 n)$  bits per label, which is optimal up to lower order terms due to the lowerbound of [AGHP16b].

For planar graphs, the main technical ingredient is an existence of small size separators. Specifically, in any planar graph there is a  $S \subseteq V$  such that  $|S| = \mathcal{O}(\sqrt{n})$  and  $S$  separates  $V$  into  $V_1$  and  $V_2$  that are balanced (up to a factor 2) in sizes. Taking advantage of this, [GPPR04] described  $\mathcal{O}(\sqrt{n})$  hub labeling and  $\mathcal{O}(\sqrt{n} \log n)$  bits distance labeling schemes, by applying the separation recursively. They also present  $\Omega(\sqrt[3]{n})$  lower bound (for both hub and distance labeling schemes). The distance labeling size was later improved in [GU16] to  $\mathcal{O}(\sqrt{n})$ , through better encoding of distances from vertex to its hub (and thus avoiding  $\log n$  overhead).

**Lower bounds.** The aforementioned lowerbounds on size of distance labels and/or hubsets for sparse graphs, and especially for planar graphs, follow a particular line of thought, which can be called a *counting argument*. One can construct a family of graphs  $\mathcal{F}$  on  $n$  vertices within the considered graph class, with a preselected subset of vertices  $V' \subseteq V$ , such that knowing  $|V'|^2$  pairwise distances in graph  $G$  is enough to identify the graph  $G \in \mathcal{F}$ . One can reason then that the total length of the  $|V'|$  labels in some graph in  $\mathcal{F}$  needs to be at least  $\log_2 |\mathcal{F}|$ , thus giving a lowerbound of  $\frac{1}{|V'|} \log_2 |\mathcal{F}|$  bits per label (the same reasoning applies to hub-labeling schemes as well). This reasoning has one fundamental limitation, in that it is unable to separate a *distributed data structure* from a *centralized* one. For example, it is believed that unweighted planar graphs do in fact admit  $\Omega(\sqrt{n})$  lowerbound for size of distance labels (as is the case for weighted planar graphs). Current lower bound construction of [GPPR04] uses counting technique construction with  $|V'| = \Theta(n^{1/3})$  and  $\log_2 |\mathcal{F}| = \Theta(n^{2/3})$ . This cannot be improved with counting technique: [AGMW18] shows that planar graphs do in fact admit *distance oracles* which take in total  $\mathcal{O}(\min |V'|^2, \sqrt{n|V'|})$  bits to encode all  $|V'|^2$  pairwise distances, which is  $\mathcal{O}(\min |V'|, \sqrt{n/|V'|})$  bits per vertex and is maximized when  $|V'| = n^{1/3}$  (one needs to argue that counting technique is not able to separate distance oracles from distance labelings, which is the case). This motivates the search for alternative techniques when proving lower bounds.

**Hub labeling in practice.** We remark that the hub-based method of distance computation is efficient for many real-world networks (even in centralised scenarios, as a distance oracle) for at least two reasons. First of all, for transportation-type networks it is possible to show bounds on the sizes of sets  $S$ , which follow from the network structure. Notably, Abraham et al. [ADF<sup>+</sup>16] introduce the notion of highway dimension  $h$  of a network, which is presumed to be a small constant e.g. for road networks, and show that an appropriate cover

of all shortest paths in the graph can be achieved using sets  $S$  of size  $\tilde{O}(h)$ . Moreover, the order in which elements of sets  $S(u)$  and  $S(v)$  is browsed when performing the minimum operation is relevant, and in some schemes, the operation can be interrupted once it is certain that the minimum has been found, before probing all elements of the set. This is the principle of numerous heuristics for the exact shortest-path problem, such as contraction hierarchies and algorithms with arc flags [KMS06, BD10].

## 1.2 Overview of our results and proof techniques

In this work we first show a lower bound on the size of hub labelings in sparse graphs. Note that in the regimes of both distance labeling and hub labeling, the state-of-the-art results for sparse graphs were leaving a huge gap: while upper bounds are of the form  $\mathcal{O}(n/\log n)$  at best (ignoring poly-loglog terms), the lower bounds are  $\Omega(\sqrt{n})$ . We close this gap for hub labeling with the following result.

**Theorem 1.1.** *Graphs of max-degree 3 require average hub size to be at least  $\frac{n}{2^{\Theta(\sqrt{\log n})}}$ , where  $n$  is the number of vertices.*

(See Section 2 for a detailed proof.)

This result is complemented by an upper bound on hub set size, given by Theorem 1.4.

Our technical contributions are obtained by using the following observations as a starting point. First, let us fix  $D = n^{o(1)}$  to be a threshold value, such that we consider distances up to  $D$  as small, and large otherwise. Simply using random hubsets of size roughly  $n/D$  we can cover all  $uv$  pairs with  $uv$ -distance being at least  $D$ . We therefore see immediately that only small distances are crucial (regardless of whether we are constructing a hard instance for a lowerbound, or small hubsets).

Next, we observe that in the regime of small distances, we can afford to fix our attention to *monotone* hubsets, where we require that for any  $u$ , if  $x \in S(u)$ , then all vertices on some chosen shortest  $ux$  path belong to  $S(u)$  as well. It trivially follows that minimal *monotone* hubset covering all small distances is within a factor of  $D$  in size from minimal arbitrary hubset. Monotone hubsets have a following advantage in analysis: for  $uv$  connected by an *unique* shortest path, for any  $x$  belonging to such path, either  $u$  or  $v$  has to “pay” for hub  $x$ , that is  $x \in S(u)$  or  $x \in S(v)$ . Thus, looking for hard instances, it is advantageous to consider the following: fix  $u \in V$ , and consider set  $S'(u) = \{h : \text{dist}(u, h) = \text{dist}(h, v) = D/2\}$  where  $v \in V$  iterates over vertices such that  $\text{dist}(u, v) = D$ . If we ensure that  $S'(u)$  need to be almost-linear in size (on average), since  $\sum |S'(u)|$  is within a  $D$  factor from total size of hubsets, we have our desired hard instance for hubset covering.

**Induced matchings in dense graphs.** It turns out that the connection between structure of shortest paths and induced matchings is at the heart of the hardness of the hub labeling problem in sparse graphs. This connection as such has already been noted in the literature (cf. [Bod17]); here, we provide a brief exposition from a perspective most relevant to our study. Let us first recall the terminology.

**Definition 1.2.** *We say that  $M \subseteq E(G)$  is an induced matching of  $G$  if (i) it is a matching and (ii) there is a set of vertices  $V' \subseteq V(G)$  such that  $M$  is a subgraph of  $G$  induced by  $V'$ .*

The connection between induced matchings and the shortest path structure follows from considering the bipartite graph  $G' = (V, V, E')$ , where  $(u, v) \in E'$  iff  $u$  and  $v$  are at distance at most  $D$  in  $G$ , for some fixed threshold value  $D$ . Now consider a quadruple of vertices  $u, v, u', v'$  such that  $(u, v) \in G'$  and  $(u', v') \in G'$ ,  $\text{dist}(u, v) = \text{dist}(u', v') \leq D$ . Take now some hub candidate  $h \in V$  and some integers  $a, b$  such that  $a + b = \text{dist}(u, v)$ . If  $h$  is on the unique shortest  $uv$  path and  $h \in S(u) \cap S(v)$  and  $\text{dist}(u, h) = a$ ,  $\text{dist}(h, v) = b$ , we can mark it by selecting  $uv$  as an edge in some subgraph  $M_h \subseteq E'$ . We do the same for  $u'v'$  and every other pair of vertices, adding edges to  $M_h$  if  $h$  is a hub on shortest path located at distance  $a$

from one endpoint and  $b$  from other. It is now an easy observation that e.g.  $\text{dist}(u, v') \leq a + b$ , and if the distance is indeed  $a + b$ , pair  $u, v'$  is already covered by hub  $h$ . Since  $h \in S(u)$  and  $h \in S(v')$  was already amortized by paying for edges  $uv$  and  $u'v'$  in  $M_h$ , there is no need for adding edge  $uv'$ , and we can ensure that  $M_h$  is indeed an induced matching (repeating the reasoning for  $u'v$  pair).

We are thus interested in structure of (linear in  $n$ ) induced matchings in (potentially dense) graphs. Specifically, when looking for hubset constructions, it is advantageous for us to have an upper bound on number of induced matchings graph can be edge partitioned into. Similarly, when looking for a lowerbound construction, we are interested in an explicit construction of dense graphs edge partitioned into large induced matchings, and moreover in having such graphs realisable as a structure of unique shortest paths. The first of these questions has been studied extensively in the combinatorics literature.

The Ruzsa-Szemerédi function  $\text{RS}(n)$  is a graph-theoretic function, defined as follows.

**Definition 1.3.** *An undirected, unweighted graph  $G$  on  $n$  vertices whose edges can be partitioned into at most  $n$  induced matchings is called a Ruzsa-Szemerédi graph.<sup>1</sup> Then  $\text{RS}(n)$  is the largest value such that every Ruzsa-Szemerédi graph has at most  $n^2/\text{RS}(n)$  edges.*

The study of values of  $\text{RS}(n)$  was initiated by [RS78] who showed that  $\text{RS}(n) = \omega(1)$ . Current bounds state that

$$2^{\Omega(\log^* n)} \leq \text{RS}(n) \leq 2^{\mathcal{O}(\sqrt{\log n})}$$

due to results of [Fox11] and [Beh46] (see [Elk10] for current best upperbound up to lower-order terms).  $\text{RS}(n)$  finds applications in algorithms and complexity theory, e.g. communication over a shared channel [BLM93], PCP Theorem [HW03], property testing [AS04] and streaming algorithms [GKK12, Kap13, Kon15]. We note that another work to exploit the connections between  $\text{RS}(n)$  and the structure of shortest paths in graphs was [Bod17], although this was done for the different (distantly related) problem of constructing *distance preservers*. Whereas the lite-motif of noting connections between sets of shortest paths and induced matchings is inherent to both [Bod17] and our work, the combinatorial arguments used in the respective constructions are significantly different.

Using a number of combinatorial insights and properties of the hub labeling problem, we show the following result.

**Theorem 1.4.** *Any graph  $G = (V, E)$  on  $n$  vertices and  $\mathcal{O}(n)$  edges admits a hub labeling  $\{S_v\}$  of average size  $\frac{1}{n} \sum_v |S(v)| = \mathcal{O}(\frac{n}{\text{RS}(n)^{1/c}})$  for some constant  $c \leq 7$ .*

(See Section 4 for a detailed proof.)

This result can be read twofold. First, it hints at the possibility that there are hub labeling schemes for sparse graphs with average hub size  $\mathcal{O}(\frac{n}{2^{\log^c n}})$  for some  $0 < c < 1$ , if one is to believe that the true value of  $\text{RS}(n)$  is on the side of the currently best upperbound. Secondly, it gives a (conditional) lowerbound on our lowerbound technique: one cannot hope to significantly improve the bound from Theorem 1.1 without strengthening existing bounds on value of  $\text{RS}(n)$ , which is a hard open problem.

The connection between induced matchings and the structure of shortest paths leads us to the missing ingredient necessary for proving Theorem 1.1. Namely, in hard instances we are looking for, in addition to already discussed structure, we need the structure of shortest paths to realize some form of Ruzsa-Szemerédi structure, that is admit edge partion into preferably  $\mathcal{O}(n/\text{RS}(n))$  induced matchings of  $n$  edges. Luckily, related ideas for such constructions were explored by [AMS12]. The construction presented there involves graph with vertex set  $[C]^d$  for some constants  $C, d$ , and interpreting vertices as  $d$ -dimensional vectors, the rule for connecting with edge  $\vec{x}$  and  $\vec{y}$  iff  $|\|\vec{x} - \vec{y}\|_2^2 - \mu| \leq n$  for some chosen constant  $\mu$ . We tweak the

<sup>1</sup>In the literature there exists term of  $(r, t)$ -Ruzsa-Szemerédi graph, where edges are partitioned into  $t$  induced matchings of size  $r$ . We follow the convention of [Bod17] and drop the indices.



construction by using  $[C]^d \times [\ell]$  as our vertex set, interpreted as  $\ell$   $d$ -dimensional layers, and connecting with an edge only between neighbouring layers (edges go always between  $i$ -th and  $(i + 1)$ -th layer for some  $i$ ). We fix  $\ell$  to be a small value being in the order of *diameter* of the graph in the original construction - thus there is an equivalence between paths in the original graph of length  $\ell$  and path in the new graph from first to last layer. Finally, we simplify the predicate deciding the existence of an edge (edges connect  $\vec{x}$  and  $\vec{y}$  if they differ in at most one coordinate), and ensure the uniqueness of a shortest paths by introducing weights over the edges of the graph.

**Distance labelings and the Sum-Index problem.** Our graph family used in the proof of Theorem 1.1 has a very regular structure. Namely, one can distinguish three (large) layers of vertices  $A, B, C \subset V$ , such that each layer is labeled with vectors from a high-dimensional integer grid, and the following property holds: for  $u_{\vec{x}} \in A$  and  $w_{\vec{z}} \in C$ , the unique shortest path intersects  $B$  in a single vertex that is precisely  $v_{(\vec{x}+\vec{z})/2} \in B$ . It is tempting to construct a family of (sparse) graphs, where shortest path length from vertices in  $A$  to vertices in  $C$  is sensitive to existence (or removal) of vertices in  $B$ . We can imagine this scenario as a game, where two players, one residing on  $A$  side, and one residing on  $C$  side of the graph, are evaluating in coordination a  $A \times C \rightarrow \{0, 1\}$  function. To our advantage, such a problem has been widely considered in *communication complexity* in the 1990's.

**Definition 1.5** (Sum-Index Problem). *Let  $S = S_0 S_1 \dots S_{n-1} \in \{0, 1\}^n$ . Alice holds  $S$  and  $a \in [0, n - 1]$ . Bob holds  $S$  and  $b \in [0, n - 1]$ . They both simultaneously send a messages  $M_a$  and  $M_b$  to a referee, whose goal is to compute  $f(M_a, M_b) = S_{(a+b) \bmod n}$ .*

This problem has been first stated explicitly in [Pud94], as a single-bit-output “extract” of a  $\{0, 1\}^n \times [n] \rightarrow \{0, 1\}^n$  shift function:  $\text{Shift}_k(x) = y$ , where  $y_i = x_{(i+k) \bmod n}$ . Analysis of Shift is closely connected to a program of proving highly non-trivial lowerbounds in circuit complexity, c.f. [HG91] for a precise reduction. Informally, the goal is to prove super-linear lowerbounds on the size of circuits for some function  $\{0, 1\}^n \rightarrow \{0, 1\}^n$ , and Shift was considered a strong candidate. However, the result of [Pud94] and later [Amb96] have shown that such direction is unfeasible (at least with Shift), by showing existence of sub-linear communication complexity protocols for evaluation of Sum-Index. Those results were called “unexpected” upperbounds, and the best construction up-to-date is due to [Amb96] with  $\text{SUMINDEX}(n) = \mathcal{O}\left(\frac{n \log^{0.25} n}{2^{\sqrt{\log n}}}\right)$  bits complexity, where  $\text{SUMINDEX}(n)$  denotes the exact bit complexity of the problem. It is also known that  $\text{SUMINDEX}(n) = \Omega(\sqrt{n})$ , cf. [BGKL03, BKL95, PRS97, NW93], and the precise complexity is still a major open problem. We prove the following reduction.

**Theorem 1.6.** *Distance labeling in graphs on  $n$  vertices and max-degree 3 requires at least  $\frac{1}{2^{\Theta(\sqrt{\log n})}} \text{SUMINDEX}(n)$  bits per vertex.*

(See Section 3 for a detailed proof.)

We note that our linkage between distance labeling size and communication complexity lowerbounds appears to be the first advancement of the techniques in the area going beyond the graph-counting technique of [GPPR04], and actually using distribution of information as a source of hardness in labeling schemes (although links between communication complexity and related to ours topic of compact routing have been investigated in [Twi06]). One might hope that, with the advancement of communication complexity bounds, this approach will eventually result in non-trivial *unconditional* lowerbounds for distance labelings.

## 2 Lower bound on Hub Labeling

This section is devoted to the proof of the following theorem.

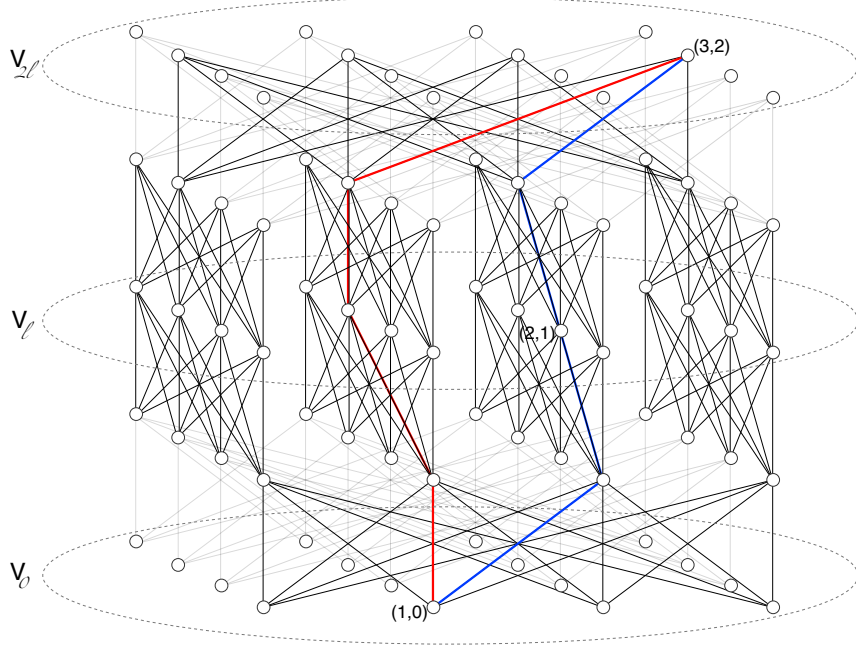


Figure 1: Graph  $H_{b,\ell}$  with  $b = 2$  and  $\ell = 2$  ( $s = 4$ ). Some edges are drawn in light gray for better readability. The blue path is the only shortest path from  $v_{0,(1,0)}$  to  $v_{4,(3,2)}$ . It passes through  $v_{2,(2,1)}$  (which is a point of symmetry for the path) and has length  $4A + 4$ . The red path has length  $4A + 8$ .

**Theorem 2.1.** For any positive integers  $\ell \in \mathbb{N}$  (the number of levels) and  $b \in \mathbb{N}$  (the side length parameter) there exists a graph  $G_{b,\ell}$  such that:

- (i)  $|V(G_{b,\ell})| = 2^{b\ell} \cdot 2^{\Theta(b+\log \ell)}$  (number of nodes).
- (ii)  $\Delta(G_{b,\ell}) = 3$  (maximum degree).
- (iii) any hub labeling  $\{S_v\}$  of  $G_{b,\ell}$  satisfies:  $\frac{1}{|V(G_{b,\ell})|} \sum_{v \in V(G_{b,\ell})} |S_v| \geq 2^{b\ell} \cdot 2^{-\Theta(b+\ell)}$  (average size of hub sets).

*Proof.* In order to construct  $G_{b,\ell}$ , we more conveniently describe a weighted graph with non-uniform length edges,  $H_{b,\ell} = (V, E, w)$  with integer edge weights appropriately chosen from the range  $w(e) \in [1, (3\ell + 1) \cdot 2^{2b}]^2$  for all  $e \in E$ .

In following we set  $s = 2^b$  (the side length). We define vertex set of  $H_{b,\ell}$  as  $V = \bigcup_{r=0}^{2\ell} V_i$ , where each level  $V_i$ ,  $i \in [0, 2\ell]$ , satisfies  $|V_i| = s^\ell$  and is identified with a set of  $\ell$ -dimensional vectors,  $V_i = \{v_{i,\vec{j}} : \vec{j} \in [0, s-1]^\ell\}$ .

The edges of  $H_{b,\ell}$  are given so that we put an edge between  $v_{i,\vec{j}}$  and  $v_{i+1,\vec{j}'}$  when  $\vec{j}$  and  $\vec{j}'$  differ at most on one coordinate  $c$ : that is  $j_k = j'_k$  for all  $k \neq c$ , where  $c = i + 1$  for  $i < \ell$  and  $c = 2\ell - i$  for  $i \geq \ell$ . The weight of such an edge is then given as  $w(\{v_{i,\vec{j}}, v_{i+1,\vec{j}'}\}) = A + (j_c - j'_c)^2$  where  $A = 3\ell s^2$ . Note that each node  $v_{i,\vec{j}}$  has exactly  $s = 2^b$  neighbors in  $V_{i+1}$  (if  $i < 2\ell$ ) and  $s$  neighbors in  $V_{i-1}$  (if  $i > 0$ ). See Figure 1 for an illustration.

We convert  $H_{b,\ell}$  into  $G_{b,\ell}$  as follows.

- We associate each vertex  $v$  of  $H_{b,\ell}$  with two disjoint perfectly balanced binary trees  $T_v^{\text{in}}$  and  $T_v^{\text{out}}$  in  $G_{b,\ell}$ , both roots being linked to  $v$ . Each tree has  $s$  leaves and depth  $\log_2 s = b$ . If  $v \in V_i$ , then leaves

<sup>2</sup>Throughout the text, we use the integer range notation  $[a, b] \equiv \{a, a + 1, \dots, b\}$  and  $[a, b) \equiv [a, b - 1]$ , for  $a, b \in \mathbb{N}$ .

of  $T_v^{\text{in}}$  are labeled as  $v_u^{\text{in}}$ , with  $u \in V_{i-1}$ ,  $\{u, v\} \in E(H_{b,\ell})$  ( $T_v^{\text{in}}$  is omitted when  $i = 0$ ), and leaves of  $T_v^{\text{out}}$  are labeled as  $v_u^{\text{out}}$ , with  $u \in V_{i+1}$ ,  $\{v, u\} \in E(H_{b,\ell})$  ( $T_v^{\text{out}}$  is omitted when  $i = 2\ell$ ).

- We associate each edge  $e = \{u, v\}$  of  $H_{b,\ell}$ ,  $u \in V_i, v \in V_{i+1}$  of length  $w(e)$  to a path of length  $w(e) - 2b + -2$  with  $w(e) - 2b - 3$  auxiliary vertices in  $G_{b,\ell}$ , starting at  $u_v^{\text{out}}$  and ending at  $v_u^{\text{in}}$ . Note that the path length satisfies  $w(e) - 2b + -2 \geq 3\ell s^2 - 2b - 2 > 0$  for  $\ell \geq 1$  and  $b \geq 1$ , and that using  $T_u^{\text{out}}$  and  $T_v^{\text{in}}$ , we obtain a path of length  $w(\{u, v\})$  in  $G_{b,\ell}$  for each edge  $\{u, v\} \in E(H_{b,\ell})$ .

We have  $|V(G_{b,\ell})| \leq |V(H_{b,\ell})| \cdot 4s + \sum_{e \in E(H_{b,\ell})} w(e) \leq 4s \cdot s^\ell \cdot (2\ell + 1) + (3\ell + 1)s^2 \cdot s^\ell \cdot 2\ell \cdot s$ , so the claims (i) and (ii) follow directly from the construction. To show claim (iii), we first observe that the following property holds.

**Lemma 2.2.** *Let  $\vec{x}, \vec{z} \in [0, s - 1]^\ell$  be such that for all  $k \in [1, \ell]$ ,  $z_k - x_k$  is even. Then there exists a unique shortest path in  $H_{b,\ell}$  (and consequently also in  $G_{b,\ell}$ ) between vertices  $v_{0,\vec{x}}$  and  $v_{2\ell,\vec{z}}$ . Moreover, this shortest path passes through vertex  $v_{\ell,(\vec{x}+\vec{z})/2}$ .*

We defer the proof of the Lemma 2.2.

Now, let  $\{S_v\}$  be any fixed hub labeling of  $G_{b,\ell}$ . Fix arbitrarily shortest path trees  $T_v$  rooted at each vertex  $v$  of  $G_{b,\ell}$ , and let  $S_v^*$  be the vertex set of the minimal subtree of  $T_v$  rooted at  $v$  which contains all vertices of  $S_v$ . We obviously have:

$$|S_v^*| \leq \text{diam}(G_{b,\ell}) \cdot |S_v| \leq (3\ell + 1)s^2 \cdot 4\ell \cdot |S_v|. \quad (1)$$

Thus, the sets  $S_v$  and  $S_v^*$  are of equivalent size up to lower-order terms, and from now on, we focus on showing a lower bound on  $\sum_{v \in V} S_v^*$ .

Consider all triplets  $(\vec{x}, \vec{y}, \vec{z})$  such that  $\vec{x}, \vec{y}, \vec{z} \in [0, s - 1]^\ell$  and that  $\vec{y} = (\vec{x} + \vec{z})/2$ . There is  $s^\ell \cdot (s/2)^\ell$  such triplets. Denote  $x = v_{0,\vec{x}}, y = v_{\ell,\vec{y}}$  and  $z = v_{2\ell,\vec{z}}$ , then by Lemma 2.2  $y$  lies on the unique shortest  $xz$  path, thus  $y \in S_x^*$  or  $y \in S_z^*$ . Since  $\vec{z} = 2\vec{x} - \vec{y}$ , we have that value of  $\vec{z}$  is uniquely determined by fixing  $\vec{x}$  and  $\vec{y}$ , and similarly by  $\vec{x} = 2\vec{z} - \vec{y}$ , value of  $\vec{x}$  is uniquely determined by fixing  $\vec{y}$  and  $\vec{z}$ . Thus in each triplet  $(\vec{x}, \vec{y}, \vec{z})$ , vertex  $y$  contributes by 1 to the size of  $S_x^*$  or  $S_z^*$ . Thus  $\sum_{v \in V} S_v^* \geq (s^\ell)^2 \cdot 2^{-\ell}$ , and taking into account Eq. (1), claim (iii) follows.  $\square$

*Proof of Lemma 2.2.* Finally, we prove the Lemma 2.2. We first consider the weighted graph  $H_{b,\ell}$  only. Intuitively, we show that the unique shortest path in  $H_{b,\ell}$  has the following structure: it climbs from level 0 to level  $\ell$  on its first  $\ell$  edges, reaching vertex  $v_{\ell,\vec{y}}$ , and then climbs next  $\ell$  edges to level  $2\ell$ ; the portions of the path from  $v_{0,\vec{x}}$  to  $v_{\ell,\vec{y}}$  and from  $v_{\ell,\vec{y}}$  to  $v_{2\ell,\vec{z}}$  have a point symmetry with respect to  $v_{\ell,\vec{y}}$ .

Let  $d(v_{0,\vec{x}}, v_{2\ell,\vec{z}})$  denote the weighted distance between the pair of nodes considered in the Lemma.

Consider any path  $P_{\vec{x},\vec{z}} = (v_{\vec{x}} \equiv v_{0,\vec{j}0}, v_{1,\vec{j}1}, \dots, v_{2\ell-1,\vec{j}2\ell-1}, v_{2\ell,\vec{j}2\ell} \equiv v_{\vec{z}})$  in  $H_{b,\ell}$  on  $2\ell$  edges. Denote  $\delta_i = j_i^i - j_i^{i-1}$  and for  $i \in [1, \ell]$  and  $\delta_i = j_{2\ell-i+1}^i - j_{2\ell-i+1}^{i-1}$  for  $i \in [\ell + 1, 2\ell]$ .

The following properties hold:

- The weighted length of any such  $P_{\vec{x},\vec{z}}$  is:

$$\sum_{i=1}^{2\ell} (A + \delta_i^2) \leq 2\ell A + 2\ell s^2 < (2\ell + 1)A.$$

- All  $v_{0,\vec{x}}v_{2\ell,\vec{z}}$  paths in  $H_{b,\ell}$  have at least  $2\ell$  edges.
- Any  $v_{0,\vec{x}}v_{2\ell,\vec{z}}$  path in  $H_{b,\ell}$  with at least  $2\ell + 1$  edges has weighted length at least  $(2\ell + 1)A$ , which is greater than that of  $P_{\vec{x},\vec{z}}$ .



- $P_{\vec{x}, \vec{z}}$  satisfies  $\delta_i + \delta_{2\ell+1-i} = z_i - x_i$ . It thus follows that the total length  $\sum_{i=1}^{\ell} (2A + \delta_i^2 + \delta_{2\ell+1-i}^2)$  is minimized when  $\delta_i = \delta_{2\ell+1-i} = \frac{1}{2}(z_i - x_i)$ , realized by an unique  $v_{0, \vec{x}} v_{2\ell, \vec{z}}$  path having point symmetry at  $v_{\ell, \vec{y}}$ .

This finishes the proof for  $H_{b, \ell}$ . To complete the proof for  $G_{b, \ell}$  we observe that for any  $u \in V_i$  and  $v \in V_j$  with  $i < j$ , we have  $\text{dist}_{G_{b, \ell}}(u, v) = \text{dist}_{H_{b, \ell}}(u, v)$ . This comes from the construction of  $G_{b, \ell}$  for  $j = i + 1$  and from the fact that any  $V_{i'}$  with  $i < i' < j$  is a vertex cut in  $G_{b, \ell}$  separating  $u$  and  $v$  for  $j > i + 1$ : any shortest path  $Q$  from  $u$  to  $v$  in  $G_{b, \ell}$  must pass through  $j - i - 1$  vertices in  $V_{i+1} \cup \dots \cup V_{j-1}$  corresponding to a path  $P$  in  $G_{b, \ell}$  such that  $|Q| = w(P)$ . The shortest path from  $v_{0, \vec{x}}$  to  $v_{2\ell, \vec{z}}$  in  $G_{b, \ell}$  is thus unique and passes through  $v_{\ell, (\vec{x} + \vec{z})/2}$  similarly as the shortest path in  $H_{b, \ell}$ , which concludes the proof.  $\square$

Setting both  $b = \ell = \sqrt{\log N}$  for some appropriately chosen  $N = n/2^{\Theta(\sqrt{\log n})}$  in Theorem 2.1 we finally obtain the main result of the Section.

**Theorem 1.1.** *Graphs of max-degree 3 require average hub size to be at least  $\frac{n}{2^{\Theta(\sqrt{\log n})}}$ , where  $n$  is the number of vertices.*

### 3 Lower bound for distance labeling

We now show that distance labeling in sparse graphs is no easier than solving Sum-Index. We essentially use the graph construction from Theorem 1.1, and the fact that the distance between nodes  $v_{0, \vec{x}}$  and  $v_{2\ell, \vec{z}}$  is sensitive to the presence (or absence) in the graph of the node  $v_{\ell, (\vec{x} + \vec{z})/2}$ .

**Theorem 1.6.** *Distance labeling in graphs on  $n$  vertices and max-degree 3 requires at least  $\frac{1}{2^{\Theta(\sqrt{\log n})}} \text{SUMINDEX}(n)$  bits per vertex.*

*Proof.* Let  $S$  be the binary vector of length  $m$  from Sum-Index Problem, with  $m$  to be determined later. We describe the strategies of Alice and Bob of constructing the messages  $M_a$  and  $M_b$  respectively. Consider graph  $G_{b, \ell}$  from Theorem 2.1. We construct  $G'_{b, \ell}$  by removing some vertices from layer  $V_{\ell}$  (together with all adjacent edges) from  $G_{b, \ell}$ . We denote the choice of whether to include or remove particular vertex  $v_{\ell, \vec{x}}$  as  $W(\vec{x})$ , and we defer how we decide those to later part of the proof.

**Observation 3.1.** *Let  $\vec{x}$  and  $\vec{z}$  be as in Lemma 2.2, that is for any  $k \in [1, \ell]$   $z_k - x_k$  is even. Then  $W((\vec{x} + \vec{z})/2)$  can be decoded based only on  $\vec{x}, \vec{z}$  and the length of the shortest path between  $v_{0, \vec{x}}$  and  $v_{2\ell, \vec{z}}$  in  $G'_{b, \ell}$ .*

We also note that  $G'$  has  $2^{b\ell} \cdot 2^{\Theta(b+\log \ell)}$  vertices and maximum degree of 3.

Recall we denoted  $s = 2^b$  the side-length of a layer in  $G$  (or  $G'$ ). We fix  $m = (s/2)^\ell$ . For a vector  $\vec{x} \in [s]^\ell$  we denote  $\text{repr}(\vec{x}) = (\sum_i x_i \cdot (s/2)^i) \bmod m$  as the integer value coming from treating coordinates of  $\vec{x}$  as digits in  $(s/2)$ -ary representation. Observe that while  $\text{repr}()$  is a bijection between  $[0, s/2 - 1]^\ell$  and  $[0, (s/2)^\ell - 1]$ , this is not the case for the whole space  $[0, s - 1]^\ell$  (in fact every value is in the image of  $2^\ell$  vectors). We fix the predicate  $W(\vec{x})$  as:  $W(\vec{x}) \equiv [S_{\text{repr}(\vec{x})} = 1]$ .

The protocol for Alice is as follow:

1. Alice constructs graph  $G'_{b, \ell}$  based on binary word  $S \in \{0, 1\}^m$ .
2. Alice constructs a *distance labeling* for  $G'_{b, \ell}$ .
3. Alice finds unique  $\vec{x} \in [0, s/2 - 1]^\ell$  such that  $\text{repr}(\vec{x}) = a$  (based on the  $(s/2)$ -ary representation of  $a$ ).

4. Alice sends to the referee the label of vertex  $v_{0,2\vec{x}}$  together with integer  $a$ .

Bob proceeds analogously, representing  $b$  as vector  $\vec{z}$ , constructing the same distance labeling of  $G'_{b,\ell}$  as Alice and sending the label of  $v_{2\ell,2\vec{z}}$  with integer  $b$ . The referee is then able to compute distance between Alice's and Bob's vertices, reconstructs vectors  $\vec{x}$  and  $\vec{z}$ , and by the Observation 3.1 knows the bit  $S_{\text{repr}(\vec{x}+\vec{z})} = S_{(a+b) \bmod m}$ , which follows from  $\text{repr}(\vec{x}+\vec{z}) = (\text{repr}(\vec{x}) + \text{repr}(\vec{z})) \bmod m = (a+b) \bmod m$ .

We thus reach the conclusion that distance labeling of sparse graph with  $2^{b\ell} \cdot 2^{\Theta(b+\log \ell)}$  requires at least  $\text{SUMINDEX}(2^{(b-1)\ell}) - b\ell$  bits per label. Fixing  $b = \ell = \sqrt{\log N}$  for some  $N = \frac{n}{2^{\Theta(\sqrt{\log n})}}$  finishes the proof.  $\square$

## 4 Upper bound

In this section we show how to take advantage of a structure of induced matchings in dense graphs to construct hubsets. We start by considering a case of graphs of constant maxdegree.

**Theorem 4.1.** *Any graph  $G = (V, E)$  on  $n$  vertices and maxdegree  $\Delta = \mathcal{O}(1)$  admits a hub labeling  $\{S_v\}$  of total size  $\sum_v |S_v| = \mathcal{O}(\frac{n^2}{\text{RS}(n)^{1/c}})$  for some constant  $c \leq 7$ .*

*Proof.* Let  $D$  be a parameter  $1 \leq D \leq \text{RS}(n)$  to be fixed later. For any pair of vertices  $u, v$  we denote  $H_{uv}$  the set of all valid hubs for this pair, that is  $H_{uv} \equiv \{x : \text{dist}(u, x) + \text{dist}(x, v) = \text{dist}(u, v)\}$ .

We start by formulating the following property for a set  $S \subseteq V$ : (\*) for two vertices  $u, v$  such that  $|H_{uv}| \geq D$ , we have  $S \cap H_{uv} \neq \emptyset$ . There exists a set  $S$  such that  $|S| = \mathcal{O}(\frac{n}{D} \log D)$  and (\*) is satisfied for all pairs  $uv$  except  $\frac{n^2}{D}$  many. Existence of such set can proven by probabilistic method (c.f. [BCE05]). Indeed, pick such set uniformly at random with  $|S| = \frac{n}{D} \ln D$ . Any pair  $u, v$  with at least  $D$  hubs is *not covered* with probability at most  $(1 - D/n)^{|S|} \leq 1/D$ , thus the expected number of uncovered pairs is at most  $n^2/D$ , and there is a selection of  $S$  with at most that many pairs not covered. From now on we focus on such  $S$ , and we denote by  $\{Q_v\}$  a family of sets such that if a pair  $uv$  is not covered by  $S$  and satisfies  $|H_{uv}| \geq D$ , then we put  $v \in Q_u$ . By the property of  $S$ , we have  $\sum_{v \in V} |Q_v| \leq n^2/D$ .

We now color all vertices of  $G$  with  $D^3$  colors, so that for each  $v \in V$ , the color  $c_v \in [1, D^3]$  is assigned independently and uniformly at random. Consider the following event: for a pair  $u, v$ , each vertex from  $H_{uv}$  is assigned a different color. If  $|H_{uv}| \leq D$ , then such an event happens with probability at least  $1 - 1/D$ , and the opposite event with probability at most  $1/D$ . If we define sets  $\{R_u\} = \{v : \text{there are } x, y \in H_{u,v} \text{ such that } c_x = c_y\}$ , then by simple computation of expected value, there is a choice of colors  $c_v$  such that  $\sum_v |R_v| \leq n^2/D$ . Thus we can afford for each  $v$  to store as hubs  $R_v$ , the vertices where coloring of potential hubs failed to assign unique colors.

We now deal with  $u, v$  such that  $|H_{uv}| \leq D$  and  $H_{uv}$  was properly colored using different colors. First, we iterate through  $a, b \geq 0$  such that  $1 \leq a + b \leq D$  and iterate  $h \in V$ . Consider bipartite graph  $(V, V, E_{a,b}^h)$ ,  $E_{a,b}^h \subseteq V \times V$ . For  $u, v$  we put  $(u, v) \in E_{a,b}^h$  if the following conditions hold: (i)  $|H_{uv}| \leq D$ , (ii) each vertex of  $H_{uv}$  was colored using distinct color, (iii)  $h \in H_{uv}$ ,  $\text{dist}(u, h) = a$  and  $\text{dist}(h, v) = b$ . We now use the following Lemma, with proof provided later:

**Lemma 4.2.** *Construct  $\{F_v\}$  as follow:  $a, b$  iterate so that  $1 \leq a + b \leq D$  and  $h$  iterate over  $V$ , and consider some minimum vertex cover  $(V_1, V_2)$ ,  $V_1, V_2 \subseteq V$  of  $(V, V, E_{a,b}^h)$ . If  $v \in V_1 \cup V_2$ , we put  $h$  into  $F_v$ . We then have  $\sum_v |F_v| = \mathcal{O}(D^5 \frac{n^2}{\text{RS}(n)})$ .*

For  $X \subseteq V$ , let  $N(X) = \{v : \exists x \in X \text{dist}(v, x) \leq 1\}$  denote neighborhood. We observe, that for any two vertices  $u, v$ , one of the following holds:

1. If  $H_{uv} \geq D$ , then either there exists  $h \in S$  such that  $h \in H_{uv}$  is a hub for  $u, v$ , or  $v \in Q_u$  is a hub for  $u, v$ .
2. If  $H_{uv} \leq D$  and there is a color conflict in  $H_{uv}$ , then  $v \in R_u$  is a hub for  $u, v$ .
3. If  $H_{uv} \leq D$  and there is no color conflict in  $H_{uv}$ , then for any  $h \in H_{uv}$ , there are  $a = \text{dist}(u, h)$  and  $b = \text{dist}(h, v)$  such that  $(u, v) \in E_{a,b}^h$ . Thus at least one of  $u, v$  is in the corresponding vertex cover, and so  $h \in F_u$  or  $h \in F_v$ . Since this holds for any  $h \in H_{uv}$ , and w.l.o.g.  $u \in F_u, v \in F_v$ , by induction along a  $uv$ -shortest path there is an edge  $(x, y) \in E$  such that  $x \in F_u$  and  $y \in F_v$  and  $x, y \in H_{uv}$ , guaranteeing that  $N(F_u) \cap N(F_v) \neq \emptyset$ .

We then set each vertex hubset  $H_v = S \cup Q_v \cup R_v \cup N(F_v)$ . A bound on size follows  $\sum_v |H_v| \leq n \cdot |S| + \sum_v |Q_v| + \sum_v |R_v| + (\Delta + 1) \sum_v |F_v| \leq \mathcal{O}(\frac{n^2}{D} \log D) + \mathcal{O}(\frac{n^2}{D}) + \mathcal{O}(D^5 \frac{n^2}{\text{RS}(n)})$  using Lemma 4.2. Setting  $D = \text{RS}(n)^{1/6}$  completes the proof.  $\square$

*Proof of Lemma 4.2.* Fix values  $a, b$  and  $h$  and vertex cover  $VC_{a,b}^h$  used in that iteration. By relation between maximum matching and minimum vertex cover, any maximal matching  $MM_{a,b}^h \subseteq E_{a,b}^h$  satisfies  $|VC_{a,b}^h| \leq 2|MM_{a,b}^h|$ . For each color  $c$ , consider bipartite graph  $G_{a,b}^c = \bigcup_{h:c_h=c} MM_{a,b}^h$ . We now show that  $MM_{a,b}^h$  is in fact an induced matching in  $G_{a,b}^c$ . It is enough to show the following: for any two distinct  $(u_1, v_1), (u_2, v_2) \in MM_{a,b}^h$ , we have  $(u_1, v_2) \notin G_{a,b}^c$  and  $(u_2, v_1) \notin G_{a,b}^c$ . Assume otherwise, that  $(u_1, v_2) \in MM_{a,b}^{h'}$  for some  $h' \neq h$  (we know we can exclude  $(u_1, v_2) \in MM_{a,b}^h$  due to  $MM_{a,b}^h$  being a matching). We know that  $\text{dist}(u_1, h) = a$  and  $\text{dist}(h, v_2) = b$ , and each vertex of  $H_{u_1 v_2}$  was colored using a different color, thus either  $\text{dist}(u_1, v_2) < a + b$  or  $c_{h'} \neq c_h$ , a contradiction.

This shows that  $G_{a,b}^c$  is a Ruzsa-Szemerédi graph (on  $2n$  vertices), with edge partition into induced matchings  $MM_{a,b}^h$  for  $h : c_h = c$ . It thus follows that  $|G_{a,b}^c| \leq \frac{(2n)^2}{\text{RS}(2n)} = \mathcal{O}(\frac{n^2}{\text{RS}(n)})$ .<sup>3</sup> We then obtain the following bound:

$$\begin{aligned}
\sum_{v \in V} |F_v| &\leq \sum_{1 \leq a+b \leq D} \sum_{h \in V} |VC_{a,b}^h| \\
&\leq \sum_{1 \leq a+b \leq D} \sum_{h \in V} 2|MM_{a,b}^h| \\
&\leq \sum_{1 \leq a+b \leq D} \sum_{c \in [1, D^3]} 2|G_{a,b}^c| \\
&= \mathcal{O}(D^5 \frac{n^2}{\text{RS}(n)}).
\end{aligned}$$

$\square$

Note that the construction used in proof of Theorem 4.1 generalizes to edges with  $\{0, 1\}$  weights since any path length  $c \leq D$  still decomposes in at most  $D$  sums  $c = a + b$  of two path lengths. We now note that if  $G$  is a graph of constant *average degree*, that is  $m/n = \mathcal{O}(1)$ , then we can reduce construction of hub labeling to constant *max degree* case as follow. First, we construct  $G'$  by subdividing any vertex  $v \in V(G)$  of degree  $\text{deg}(v)$  into  $\lceil \frac{\text{deg}(v)}{\lceil m/n \rceil} \rceil$  vertices of degree at most  $2 + \lceil m/n \rceil$ , using a path with weight-0 auxiliary edges for linking them (treating all the non-auxiliary edges as weight-1 edges). We have then  $|E(G')| = \mathcal{O}(m)$  and  $|V(G')| = \mathcal{O}(m)$ . We note that for any  $v' \in V(G')$  there is  $v \in V(G)$  that originated  $v'$ , and for any

<sup>3</sup>It is straightforward to show  $\text{RS}(2n) = \Theta(\text{RS}(n))$ .

$v \in V(G)$  we can pick one  $v' \in V(G')$  as a representative of  $v$  in  $G'$ . Having constructed hub labeling for  $G'$ , denote it  $\{H'_v\}$ , we construct a hubset of  $v \in V(G)$  by taking hubset of its representative in  $G'$  and projecting back each selected hub in  $G'$  to its original vertex in  $G$ . We thus have reached our main result on the side of upper bounds.

**Theorem 1.4.** *Any graph  $G = (V, E)$  on  $n$  vertices and  $\mathcal{O}(n)$  edges admits a hub labeling  $\{S_v\}$  of average size  $\frac{1}{n} \sum_v |S(v)| = \mathcal{O}(\frac{n}{RS(n)^{1/c}})$  for some constant  $c \leq 7$ .*

## References

- [ADF<sup>+</sup>16] Ittai Abraham, Daniel Delling, Amos Fiat, Andrew V. Goldberg, and Renato F. Werneck. Highway dimension and provably efficient shortest path algorithms. *J. ACM*, 63(5):41:1–41:26, 2016.
- [ADGW12] Ittai Abraham, Daniel Delling, Andrew V. Goldberg, and Renato F. Werneck. Hierarchical hub labelings for shortest paths. In *ESA*, pages 24–35, 2012.
- [ADKP16] Stephen Alstrup, Søren Dahlgaard, Mathias Bæk Tejs Knudsen, and Ely Porat. Sublinear distance labeling. In *ESA*, pages 5:1–5:15, 2016.
- [AG11] Ittai Abraham and Cyril Gavoille. On approximate distance labels and routing schemes with affine stretch. In *DISC*, pages 404–415, 2011.
- [AGHP16a] Stephen Alstrup, Cyril Gavoille, Esben Bstrup Halvorsen, and Holger Petersen. Simpler, faster and shorter labels for distances in graphs. In *SODA*, pages 338–350, 2016.
- [AGHP16b] Stephen Alstrup, Inge Li Gørtz, Esben Bstrup Halvorsen, and Ely Porat. Distance labeling schemes for trees. In *ICALP*, pages 132:1–132:16, 2016.
- [AGMW18] Amir Abboud, Paweł Gawrychowski, Shay Mozes, and Oren Weimann. Near-optimal compression for the planar graph metric. In *SODA*, pages 530–549, 2018.
- [Amb96] Andris Ambainis. Upper bounds on multiparty communication complexity of shifts. In *STACS*, pages 631–642, 1996.
- [AMS12] Noga Alon, Ankur Moitra, and Benny Sudakov. Nearly complete graphs decomposable into large induced matchings and their applications. In *STOC*, pages 1079–1090, 2012.
- [AS04] Noga Alon and Asaf Shapira. Testing subgraphs in directed graphs. *J. Comput. Syst. Sci.*, 69(3):354–382, 2004.
- [BCE05] Béla Bollobás, Don Coppersmith, and Michael Elkin. Sparse distance preservers and additive spanners. *SIAM Journal on Discrete Mathematics*, 19(4):1029–1055, 2005.
- [BD10] Reinhard Bauer and Daniel Delling. SHARC: Fast and robust unidirectional routing. *J. Exp. Algorithmics*, 14:4:2.4–4:2.29, January 2010.
- [Beh46] Felix A Behrend. On sets of integers which contain no three terms in arithmetical progression. *Proceedings of the National Academy of Sciences of the United States of America*, 32(12):331, 1946.

- [BGKL03] László Babai, Anna Gál, Peter G. Kimmel, and Satyanarayana V. Lokam. Communication complexity of simultaneous messages. *SIAM J. Comput.*, 33(1):137–166, 2003.
- [BKL95] László Babai, Peter G. Kimmel, and Satyanarayana V. Lokam. Simultaneous messages vs. communication. In *STACS*, pages 361–372, 1995.
- [BLM93] Yitzhak Birk, Nathan Linial, and Roy Meshulam. On the uniform-traffic capacity of single-hop interconnections employing shared directional multichannels. *IEEE Trans. Information Theory*, 39(1):186–191, 1993.
- [Bod17] Greg Bodwin. Linear size distance preservers. In *SODA*, pages 600–615, 2017.
- [CHKZ03] Edith Cohen, Eran Halperin, Haim Kaplan, and Uri Zwick. Reachability and distance queries via 2-hop labels. *SIAM J. Comput.*, 32(5):1338–1355, May 2003.
- [CP10] Hagai Cohen and Ely Porat. On the hardness of distance oracle for sparse graph. *CoRR*, abs/1006.1117, 2010.
- [Elk10] Michael Elkin. An improved construction of progression-free sets. In *SODA*, pages 886–905, 2010.
- [FGNW17] Ofer Freedman, Paweł Gawrychowski, Patrick K. Nicholson, and Oren Weimann. Optimal distance labeling schemes for trees. In *PODC*, pages 185–194, 2017.
- [Fox11] Jacob Fox. A new proof of the graph removal lemma. *Annals of Mathematics*, pages 561–579, 2011.
- [GKK12] Ashish Goel, Michael Kapralov, and Sanjeev Khanna. On the communication and streaming complexity of maximum bipartite matching. In *SODA*, pages 468–485, 2012.
- [GKU16] Paweł Gawrychowski, Adrian Kosowski, and Przemysław Uznański. Sublinear-space distance labeling using hubs. In *DISC*, pages 230–242, 2016.
- [GP72] R.L. Graham and H.O. Pollak. On embedding graphs in squashed cubes. In Y. Alavi, D.R. Lick, and A.T. White, editors, *Graph Theory and Applications*, volume 303 of *Lecture Notes in Mathematics*, pages 99–110. Springer Berlin Heidelberg, 1972.
- [GPPR04] Cyril Gavoille, David Peleg, Stéphane Pérennes, and Ran Raz. Distance labeling in graphs. *J. Algorithms*, 53(1):85–112, October 2004.
- [GU16] Paweł Gawrychowski and Przemysław Uznański. A note on distance labeling in planar graphs. *CoRR*, abs/1611.06529, 2016.
- [HG91] Johan Håstad and Mikael Goldmann. On the power of small-depth threshold circuits. *Computational Complexity*, 1:113–129, 1991.
- [HW03] Johan Håstad and Avi Wigderson. Simple analysis of graph tests for linearity and PCP. *Random Struct. Algorithms*, 22(2):139–160, 2003.
- [Kap13] Michael Kapralov. Better bounds for matchings in the streaming model. In *SODA*, pages 1679–1697, 2013.
- [KMS06] Ekkehard Köhler, Rolf H. Möhring, and Heiko Schilling. Fast point-to-point shortest path computations with arc-flags. In *9th DIMACS Implementation Challenge*, 2006.



- [Kon15] Christian Konrad. Maximum matching in turnstile streams. In *ESA*, pages 840–852, 2015.
- [NW93] Noam Nisan and Avi Wigderson. Rounds in communication complexity revisited. *SIAM J. Comput.*, 22(1):211–219, 1993.
- [Pel00] David Peleg. Proximity-preserving labeling schemes. *Journal of Graph Theory*, 33(3):167–176, 2000.
- [PRS97] Pavel Pudlák, Vojtech Rödl, and Jirí Sgall. Boolean circuits, tensor ranks, and communication complexity. *SIAM J. Comput.*, 26(3):605–633, 1997.
- [Pud94] Pavel Pudlák. Unexpected upper bounds on the complexity of some communication games. In *ICALP*, pages 1–10, 1994.
- [RS78] Imre Z. Ruzsa and Endre Szemerédi. Triple systems with no six points carrying three triangles. *Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai*, 18:939–945, 1978.
- [SVY09] Christian Sommer, Elad Verbin, and Wei Yu. Distance oracles for sparse graphs. In *FOCS*, pages 703–712, 2009.
- [Twi06] Andrew D Twigg. Compact forbidden-set routing. Technical report, University of Cambridge, Computer Laboratory, 2006.
- [WP11] Oren Weimann and David Peleg. A note on exact distance labeling. *Inf. Process. Lett.*, 111(14):671–673, 2011.