Approximate Distance Oracles with Improved Stretch for Sparse Graphs

Liam Roditty and Roei Tov^(✉)

Bar Ilan University, Ramat Gan, Israel

Abstract. Thorup and Zwick [19] introduced the notion of approximate distance oracles, a data structure that produces for an n-vertices, m-edges weighted undirected graph $G = (V, E)$, distance estimations in *constant* query time. They presented a distance oracle of size $O(kn^{1+1/k})$ that given a pair of vertices $u, v \in V$ at distance $d(u, v)$ produces in $O(k)$ time an estimation that is bounded by $(2k - 1)d(u, v)$, i.e., a $(2k - 1)$-multiplicative approximation (stretch). Thorup and Zwick [19] presented also a lower bound based on the girth conjecture of Erdős.

For sparse unweighted graphs (i.e., $m = \tilde{O}(n)$) the lower bound does not apply. Pătraşcu and Roditty [10] used the sparsity of the graph and obtained a distance oracle that uses $\tilde{O}(n^{5/3})$ space, has $O(1)$ query time and a stretch of 2. Pătraşcu et al. [11] presented infinity many distance oracles with fractional stretch factors that for graphs with $m = \tilde{O}(n)$ converge exactly to the integral stretch factors and the corresponding space bound of Thorup and Zwick.

It is not known, however, whether graph sparsity can help to get a stretch which is better than $(2k - 1)$ using only $\tilde{O}(kn^{1+1/k})$ space. In this paper we answer this open question and prove a separation between sparse and dense graphs by showing that using sparsity it is possible to obtain better stretch/space tradeoffs than those of Thorup and Zwick. We show that for every $k \geq 2$ there is a distance oracle of size $O(knm^{1/k} \log n)$ that produces in $O(k)$ time an estimation $d^*(u, v)$ that satisfies $d(u, v) \leq d^*(u, v) \leq (2k - 1)d(u, v) - 4$, for $k > 2$, and $d(u, v) \leq d^*(u, v) \leq 3d(u, v) - 2$, for $k = 2$.

Another contribution of this paper is a refined stretch analysis of Thorup and Zwick distance oracles that allows us to obtain a better understanding of this important data structure. We present simple conditions for every $w \in V$ that characterizes the exact scenarios in which every query that involves w produces an estimation of stretch strictly better than $2k - 1$, even in the case of dense graphs. We complement this contribution with an experiment on real world graphs. The main finding in the experiment is that different real world graphs are likely to satisfy the required conditions and hence the stretch of Thorup and Zwick distance oracles is much better than its worst case bound in these real world graphs.

Keywords: Graph algorithms · Approximate shortest paths · Approximate distance oracles

C.-Y. Chen et al. (Eds.): COCOON 2021, LNCS 13025, pp. 89–100, 2021.
https://doi.org/10.1007/978-3-030-89543-3_8

1 Introduction

An approximate distance oracle is a data structure that is required to produce distance estimations in *constant* query time. Thorup and Zwick [19] showed that given an undirected weighted graph $G = (V, E)$ with m edges and n vertices and an integer $k \geq 1$, there is a data structure of size $O(kn^{1+1/k})$ that for every pair of vertices $u, v \in V$ returns in $O(k)$ time an estimation $\hat{d}(u, v)$ which is a $(2k - 1)$ multiplicative approximation (stretch) of $d(u, v)$, that is, $d(u, v) \leq \hat{d}(u, v) \leq (2k-1)d(u, v)$, where $d(u, v)$ is the length of the shortest path between u and v in G.

Thorup and Zwick [19] presented also a lower bound based on the girth conjecture of Erdős[1]. More specifically, they proved that, for every $k \geq 1$, if there is a graph of $\Omega(n^{1+1/k})$ edges whose girth is $2k + 2$ then any distance oracle with stretch $t \leq 2k$, requires $\Omega(n^{1+1/k})$ bits on some input. A careful examination of their proof reveals that it relies on the stretch of the estimation for vertex pairs $u, v \in V$ for which $(u, v) \in E$, that is, $d(u, v) = 1$. Therefore, it still might be possible to obtain a data structure with constant query time and a stretch better than $2k-1$ using $O(kn^{1+1/k})$ space, for vertex pairs $u, v \in V$ that satisfy $d(u, v) \geq 2$, or for graphs with $m = o(n^{1+1/k})$, that is, sparse graphs[2].

We present a new distance oracle for unweighted undirected graphs, that uses $O(knm^{1/k} \log n)$ space and provides in $O(k)$ query time an estimation $d^*(u, v)$ that satisfies $d(u, v) \leq d^*(u, v) \leq (2k - 1)d(u, v) - 4$, for every $k > 2$, and $d(u, v) \leq d^*(u, v) \leq 3d(u, v) - 2$, for $k = 2$. This implies that for sparse graphs with $m = \tilde{O}(n)$[3] our new distance oracle uses the same space as Thorup and Zwick's distance oracle (up to poly-logarithmic factors) and produces in $O(k)$ time an estimation of strictly better stretch than the stretch of Thorup and Zwick's distance oracle. Sparse graphs with $m = \tilde{O}(n)$ edges are very interesting both from the practical perspective and the theoretical perspective.

From the practical perspective, it is important to note that many real world graphs are sparse and $m = \tilde{O}(n)$. This is usually the case in social networks and in many other types on networks[4].

From the theoretical perspective, Pătraşcu, Roditty and Thorup [11] proved a conditional lower bound for the case of sparse graphs with $m = \tilde{O}(n)$, based on a set intersection hardness conjecture. They showed that for any $\ell > 1$, a distance oracle that for every pair of vertices at distance $\ell + 1$, provides in constant query time an estimation strictly smaller than $3(\ell + 1) - 2$ requires $\tilde{\Omega}(n^{1+\frac{1}{2-1/\ell}})$ space. Notice that for $k = 2$ our distance oracle has an estimation that is at most $3d(u, v) - 2$, for every $u, v \in V$ and uses $\tilde{O}(n^{1.5})$ space for sparse graphs with

[1] The girth is the length of the shortest cycle in an unweighted graph.

[2] A trivial way to get a smaller space for sparse graphs is to simply save the graph and answer any query in $O(m)$ time by doing BFS, this however, violates the additional requirement for distance oracles of a constant or almost a constant query time.

[3] Throughout the paper we will use the $\tilde{O}(\cdot)$ notation to hide small poly-logarithmic factors.

[4] See for more examples https://snap.stanford.edu/index.html.

$m = \tilde{O}(n)$. It follows from [11] that bounding the estimation by a value strictly smaller than $3d(u,v) - 2$ requires $\tilde{\Omega}(n^{1.5+\varepsilon})$ space, where $\varepsilon > 0$.

Pătraşcu et al. [11] showed also that there are infinitely many distance oracles for sparse graphs with fractional stretch factors. Their distance oracles converge exactly to the integral stretch factors and the corresponding space bound of Thorup-Zwick distance oracles. Our new construction implies that for space $\tilde{O}(km^{1+1/k})$ a stretch that is strictly better than the corresponding integral stretch of $2k - 1$ is possible.

The implications of our new distance oracles are not restricted only for sparse graphs with $m = \tilde{O}(n)$. Consider graphs with $m \in [n, o(n^{1+1/k})]$ edges. A natural question is whether a distance oracle for such graphs requires $\Omega(n^{1+1/k})$ for stretch $2k - 1$. The girth based approach, as in the lower bound of Thorup and Zwick [19], is not possible here since we can store the entire graph. This implies that for vertex pairs $u, v \in V$ with $d(u,v) = 1$, we can store the exact distance. Our new distance oracle rules out also the option to use pairs of vertices $u, v \in V$ for which $d(u,v) = 2$, as a possible source of hardness for a possible lower bound. If we construct our new distance oracle with parameter $k + 1$ then the space required is in the range $[n, o(n^{1+1/k})]$ and for every pair of vertices $u, v \in V$, for which $d(u,v) = 2$, the estimation is at most $(2(k+1) - 1)2 - 4 = (2k - 1)2$, and therefore, when $d(u,v) = 2$ the stretch is at most $2k - 1$.

The distance oracles of Thorup and Zwick, beside being an important data structure on their own, are also extremely useful as a tool in many applications. They were a crucial building block in several important dynamic graph algorithms along the last decade (e.g., [2,7,8,16]). They also play a pivotal role in designing distance labeling and compact routing schemes as was already shown by Thorup and Zwick [18] and in subsequent works (e.g., [1,3,13,14]). Distance oracles were also implemented and tested (e.g., [6,12]) and found useful on real world graphs. Therefore, any further understanding that we gain on the basic properties of distance oracles is of great interest.

We obtain our new distance oracle by a careful combination of a variant of Thorup and Zwick distance oracles with a new idea that interplays between a hitting set of vertices and a hitting set of edges to overcome a certain hard case that is relatively common in analysis of algorithms of shortest paths. Therefore, our new approach is of independent interest, as it might be found useful in other closely related problems.

Motivated by our theoretical finding, another contribution that we make in this paper is a refined analysis of the stretch of Thorup and Zwick distance oracles. At the base of the distance oracles there is an hierarchy of vertex sets $A_0, A_1, \ldots, A_k$, where $A_0 = V$, $A_k = \emptyset$ and A_i is formed by picking each vertex of A_{i-1}, independently, with some probability p. For every $u \in V$ the distance $d(u, A_i)$ between u and A_i is computed and saved. We introduce a simple parameter, called the *average distance*, which is roughly defined[5] for every $i \in [1, k-1]$ as the distance between u and A_i divided by i, that is $d(u, A_i)/i$. Our refined analysis characterizes several cases in which the stretch is strictly better than

[5] In the formal definition we take the ceiling of the average distance.

$2k-1$ using only the average distance, which can be easily computed using the current information saved with the distance oracle. Roughly speaking, if there exist $i, j \in [1, k-1]$ such that $i \neq j$ and $d(u, A_i)/i \neq d(u, A_j)/j$, then the stretch is strictly better than $2k-1$ for every distance query that includes the vertex u.

Based on similar ideas we also show that if $D(u) = \{\Delta_1, \dots, \Delta_\ell\}$ is the set of all possible distances of $u \in V$ with other vertices in the graph then there is at most one value $\Delta \in D(u)$ for which the stretch of the distance estimation is exactly $2k-1$, that is, only for vertices v that satisfy $d(u, v) = \Delta$ it might be that $\hat{d}(u, v) = (2k-1)d(u, v)$.

We complement the refined stretch analysis by conducting a small experiment on real world graphs. In the experiment we check how frequent are the cases that allow for a better stretch in these real world graphs. Interestingly, these cases are quite frequent and thus in many cases the actual stretch is much better than the worst case stretch bound.

1.1 Related Work

Since their introduction by Thorup and Zwick [19] distance oracles were studied by many researchers. Chechik [4,5], presented a $(2k-1)$-stretch distance oracle with $O(1)$ query time and $O(n^{1+1/k})$ space. (See also [9,20].)

Pătraşcu and Roditty [10] showed a distance oracle for weighted undirected graphs with stretch 2 and size $O(n^{4/3}m^{1/3})$. For $m = o(n^2)$, this distance oracle has $o(n^2)$ size and stretch 2. Pătraşcu, Roditty and Thorup [11] showed for every integer $k \geq 0$ and $\ell > 0$ distance oracles, that use $\tilde{O}(m^{1+1/(k\pm1/\ell)})$ space and answer distance query in $O(k+\ell)$ time with stretch $2k+1\pm2/\ell$. Sommer, Verbin, and Yu [17] provided a lower bound in the cell probe model. They showed that there are sparse graphs for which constant stretch and query time requires $m^{1+\Omega(1)}$ space[6].

Due to lack of space, we refer the reader to the full version of this paper [15] for the rest of the related work section.

1.2 Paper Organization

In the next section we present some necessary preliminaries, the distance oracles of Thorup-Zwick and a standard variant of it, that is required in order to obtain our new distance oracle. In Sect. 3 we present our new distance oracles. In Sect. 4 we present our refined stretch analysis for Thorup-Zwick distance oracles. In Sect. 5 we present some concluding remarks and open problems. Due to lack of space, we omit here some of the proofs of Sect. 2 and the technical part of Sect. 4. We refer the reader to [15] for the full version of this paper. Also, in [15] we present the experiment that we have conducted on real world graphs. In the experiment we examine how frequent are the cases that are characterized in our refined stretch analysis from Sect. 4.

[6] Using current techniques of cell probe lower bounds we cannot hope for more specific tradeoff since it is not possible to separate asymptotically the query times of data structures of size $m^{1.99}$ and $m^{1.01}$ for input size m.

2 Preliminaries and Previous Work

Let $G = (V, E)$ be an n-vertices m-edges undirected unweighted graph. For every $u, v \in V$, let $d(u, v)$ be the length of the shortest path between u and v. Let $N(u)$ be the vertices that are neighbours of u and let $deg(u) = |N(u)|$ be the degree of u.

For every set $A \subseteq V$, let $p_A(u)$ be the closest vertex to u from A, that is $p_A(u) := \arg\min_{v \in A}(d(u, v))$, where ties are broken in favor of the vertex with a smaller identifier, and let $d(u, A) = d(u, p_A(u))$. Notice that it follows from this definition that if v is on a shortest path between u and $p_A(u)$, then $p_A(u) = p_A(v)$. For a set $E' \subseteq E$ let $V(E') = \{u \mid (u, v) \in E'\}$. Let $N(u, s, A)$ be the s closest vertices to u from the set A.

Let $B(u, r) = \{v \in V \mid d(u, v) < r\}$ and let $B(u, r, X) = \{v \in X \mid d(u, v) < r\}$, where $X \subseteq V$. Let $L(u, r) = \{v \in V \mid d(u, v) = r\}$.

The following Lemma is a standard tool in the area of approximate shortest paths and we provide it here for completeness.

Lemma 1. (e.g. Lemma 3.6 in [19]). *Let U be a set of size u. Let $Q_1, \ldots, Q_n \subseteq U$. If $|Q_i| \geq s$, for every $1 \leq i \leq n$ then a hitting set A of size $\tilde{O}(u/s)$ such that $Q_i \cap A \neq \emptyset$ can be found with a deterministic algorithm in $O(u + \sum_{i=1}^{n} |Q_i|)$ time.*

2.1 The Distance Oracle of Thorup and Zwick

In their seminal paper Thorup and Zwick [19] showed that there is a data structure of size $O(kn^{1+1/k})$ that returns a $(2k - 1)$ multiplicative approximation (stretch) of the distances of an undirected weighted graph in $O(k)$ time. Let $k \geq 1$ and let $A_0, A_1, \ldots, A_k$ be sets of vertices, such that $A_0 = V$, $A_k = \emptyset$ and A_i is a subset of A_{i-1} of size at most $\tilde{O}(|A_{i-1}|/s)$ that hits for every $v \in V$ the set $N(v, s, A_{i-1})$, where s is a parameter. The set A_i is computed using Lemma 1. For every $u \in V$, let $p_i(u) = p_{A_i}(u)$ and $\ell_i(u) = d(u, A_i) = d(u, p_i(u))$. We set $p_0(u)$ to u, $p_k(u)$ to be null and $\ell_k(u)$ to ∞.

For every $0 \leq i \leq k - 1$, let $B_i(u) = B(u, \ell_{i+1}(u), A_i)$. The *bunch* of $u \in V$ is $B(u) = \cup_{i=0}^{k-1} B_i(u)$.

The information saved in the distance oracle for every $u \in V$ is $B(u) = \cup_{i=0}^{k-1} B_i(u)$, the value of $d(u, v)$, for every $v \in B(u)$, in a 2-level hash table and the vertex $p_i(u)$, where $0 \leq i \leq k$.

Thorup and Zwick proved the following:

Lemma 2. *[Theorem 3.7 [19]]. For every $u \in V$ and $i \in [0, k - 2]$, the size of $B_i(u)$ is at most s and the size of $B_{k-1}(u)$ is $\tilde{O}(n/s^{k-1})$.*

Setting $s = n^{1/k} c \log n$ yields the desired size bound $O(kn^{1+1/k})$. The query algorithm $dist(u, v)$ of the distance oracle is presented in [15]. We look for the smallest even i such that $p_i(u) \in B_i(v)$ or $p_{i+1}(v) \in B_{i+1}(u)$. Since both $p_{k-1}(u) \in B_{k-1}(v)$ and $p_{k-1}(v) \in B_{k-1}(u)$ the algorithm always stops. Let $f(u, v)$ be the largest value that i reached to during the run of $dist(u, v)$. In other words, $f(u,v)$ is the largest value such that for every even $j < f(u,v)$, it holds that $p_j(u) \notin B_j(v)$ and for every odd $j < f(u,v)$ it holds that $p_j(v) \notin B_j(u)$. Since $dist(u,v)$ always stops it follows that $f(u,v) \leq k-1$.

To bound the stretch we first prove the following Lemma that is implicit in [19]. We prove it explicitly in [15] since we use it in our proofs

Lemma 3. *For every even $i \leq f(u,v)$ it holds that $\ell_i(u) \leq i \cdot d(u,v)$ and for every odd $i \leq f(u,v)$ it holds that $\ell_i(v) \leq i \cdot d(u,v)$.*

We proceed with the following useful observation on Thorup-Zwick distance oracle that we will use later on. Consider the set A_{i-j}, where i and j are even and $0 \leq j < i \leq f(u,v)$. From Lemma 3 it follows that $\ell_{i-j}(u) \leq (i-j) \cdot d(u,v)$ and $\ell_i(u) \leq i \cdot d(u,v)$. But what if we have a bound for $\ell_{i-j}(u)$ that is better than $(i-j) \cdot d(u,v)$, can we use it to obtain a better bound for $\ell_i(u)$? In the next Lemma we present a generalization of Lemma 3 and show that this is indeed possible. The proof is given in [15].

Lemma 4. *For every even $i \leq f(u,v)$: (i) $\ell_i(u) \leq \ell_{i-j}(u) + j \cdot d(u,v)$, for every even $j \leq i$, and (ii) $\ell_i(u) \leq \ell_{i-j}(v) + j \cdot d(u,v)$, for every odd $j \leq i$.*

For every odd $i \leq f(u,v)$: (i) $\ell_i(v) \leq \ell_{i-j}(u) + j \cdot d(u,v)$, for every even $j \leq i$, and (ii) $\ell_i(v) \leq \ell_{i-j}(v) + j \cdot d(u,v)$, for every odd $j \leq i$.

We finish the description of Thorup-Zwick distance oracle with a bound on $dist(u,v)$.

Lemma 5. *$dist(u,v)$ outputs an estimation that is bounded by $2\ell_{f(u,v)}(u) + d(u,v) \leq (2f(u,v)+1)d(u,v) \leq (2k-1)d(u,v)$, for even $f(u,v)$ and by $2\ell_{f(u,v)}(v) + d(u,v) \leq (2f(u,v)+1)d(u,v) \leq (2k-1)d(u,v)$, for odd $f(u,v)$.*

Proof. Let $i = f(u,v)$ be even. The algorithm returns $d(u,p_i(u)) + d(v,p_i(u))$. Using the triangle inequality we get $d(u,p_i(u)) + d(v,p_i(u)) \leq 2\ell_i(u) + d(u,v)$. From Lemma 3 we have $\ell_i(u) \leq i \cdot d(u,v)$ and since $i \leq k-1$ we get $d(u,p_i(u)) + d(v,p_i(u)) \leq (2i+1)d(u,v) \leq (2k-1)d(u,v)$. For the case that $f(u,v)$ is odd the proof is the same with u and v switching their roles.

2.2 A Standard Variant of the Distance Oracle of Thorup and Zwick

In order to obtain the new distance oracle we are using a slightly different but relatively standard variant of the distance oracle of Thorup and Zwick (e.g. [5]), which we present below.

In this variant we also save in the distance oracle the exact distance for every pair $\langle u,v \rangle \in A_{k/2} \times A_{k/2-1}$, when k is even, and every pair $\langle u,v \rangle \in A_{(k-1)/2} \times A_{(k-1)/2}$ when k is odd. In both cases the space remains $O(kn^{1+1/k} \log n)$, since $|A_{k/2}| \cdot |A_{k/2-1}| = O(kn^{1+1/k} \log n)$, when k is even and $|A_{(k-1)/2}| \cdot |A_{(k-1)/2}| = O(kn^{1+1/k} \log n)$, when k is odd.

The query will work as follows. Let $u, v \in V$. Let $f = \min(f(u,v), f(v,u))$. If $f \leq \lfloor k/2 \rfloor$ then we output $\min(dist(u,v), dist(v,u))$. If $f > \lfloor k/2 \rfloor$ then we output $\min\big(\ell_{k/2}(u) + d(p_{k/2}(u), p_{k/2-1}(v)) + \ell_{k/2-1}(v), \ell_{k/2}(v) + d(p_{k/2}(v), p_{k/2-1}(u)) + \ell_{k/2-1}(u)\big)$, for an even k, and $\ell_{(k-1)/2}(u) + d(p_{(k-1)/2}(u),$ $p_{(k-1)/2}(v)) + \ell_{(k-1)/2}(v)$, for an odd k.

In the next Lemma we establish an upper bound on the query output when $f > \lfloor k/2 \rfloor$.

Lemma 6. *When $f > \lfloor k/2 \rfloor$ the query algorithm described above returns an estimation that is at most* $\min(2\ell_{k/2}(u) + 2\ell_{k/2-1}(v) + d(u,v), 2\ell_{k/2}(v) + 2\ell_{k/2-1}(u) + d(u,v))$*, when k is even and at most* $2\ell_{(k-1)/2}(u) + 2\ell_{(k-1)/2}(v) + d(u,v)$*, when k is odd.*

Proof. Let $a = \ell_{k/2}(u) + d(p_{k/2}(u), p_{k/2-1}(v)) + \ell_{k/2-1}(v)$. Let $b = \ell_{k/2}(v) + d(p_{k/2}(v), p_{k/2-1}(u)) + \ell_{k/2-1}(u)$. Let $A = 2\ell_{k/2}(u) + 2\ell_{k/2-1}(v) + d(u,v)$ and let $B = 2\ell_{k/2}(v) + 2\ell_{k/2-1}(u) + d(u,v)$. For even k, the query returns $\min\big(a, b\big)$. We show that this value is at most $\min(A, B)$.

Using the triangle inequality we get that $d(p_{k/2}(u), p_{k/2-1}(v)) \leq \ell_{k/2}(u) + d(u,v) + \ell_{k/2-1}(v)$. Therefore, $a \leq A$. Similarly, we get that $d(p_{k/2}(v), p_{k/2-1}(u)) \leq \ell_{k/2}(v) + d(u,v) + \ell_{k/2-1}(u)$. Therefore, $b \leq B$. Adding it all together we get that $\min(a,b) \leq \min(A,B)$, as required.

When k is odd, the query returns $\ell_{(k-1)/2}(u) + d(p_{(k-1)/2}(u), p_{(k-1)/2}(v)) + \ell_{(k-1)/2}(v) \leq \ell_{(k-1)/2}(u) + (\ell_{(k-1)/2}(u) + d(u,v) + \ell_{(k-1)/2}(v)) + \ell_{(k-1)/2}(v) = 2\ell_{(k-1)/2}(u) + 2\ell_{(k-1)/2}(v) + d(u,v)$.

It is relatively straightforward to prove that the estimation produced by the updated query algorithm has $2k-1$ stretch by combining Lemma 6 with Lemma 3.

Throughout the paper we will refer to this variant of Thorup-Zwick distance oracle as the standard variant of Thorup-Zwick distance oracle.

3 Distance Oracles with Improved Stretch

In this section we present our new distance oracle construction. We combine between two ideas. The first idea is to interplay between a hitting set of vertices and a hitting set of edges. This allows us to obtain, in some cases, a better bound on $\ell_1(u)$, for every $u \in V$. Consider a pair of vertices $u, v \in V$ such that $d(u,v) = \Delta$. In Thorup and Zwick distance oracles if $v \notin B_0(u)$ then it follows that $\ell_1(u) \leq \Delta$ and this bound is used, among other bounds, to bound the estimation. In our distance oracles we will have to use $\ell_1(u)$ to bound the estimation only in the case that $\ell_1(u) \leq \Delta - 1$. Our second idea is that in order to amplify the affect of this better bound we can use the standard variant of Thorup and Zwick distance oracles, presented in Sect. 2.2, since it allows to combine in the bound of the estimation both $\ell_1(u)$ and $\ell_1(v)$ in the case that both $\ell_1(u) \leq \Delta - 1$ and $\ell_1(v) \leq \Delta - 1$.

We now prove the following Theorem:

Theorem 1. *Let $G = (V, E)$ be an n-vertices m-edges undirected unweighted graph. For every $k > 2$ there is a distance oracle that uses $O(knm^{1/k} \log n)$ space and for every pair of vertices $u, v \in V$ returns in $O(k)$ time an estimation $d^*(u, v)$ such that:*

$$d(u, v) \leq d^*(u, v) \leq (2k - 1)d(u, v) - 4.$$

For $k = 2$, the estimation $d^*(u, v)$ satisfies: $d(u, v) \leq d^*(u, v) \leq 3d(u, v) - 2$.

Proof. Our new distance oracle is constructed as follows. Let $s = m^{1/k} c \log n$. We start with the set A_1 that will be the union of two sets, A_1^{v} and A_1^{e}. The set $A_1^{\mathrm{v}} \subseteq V$ is a hitting set of size $\tilde{O}(m/s)$ of the sets $N(v, s, V)$, for every $v \in V$, computed using Lemma 1.

The set A_1^{e} is computed as follows. We first compute for every $u \in V$ the set $L(u, d(u, A_1^{\mathrm{v}}))$. Let $V^H = \{u \mid |L(u, d(u, A_1^{\mathrm{v}}))| \geq s\}$. For every $u \in V^H$ let $E^H(u) = \{(x, y) \in E \mid x \in L(u, d(u, A_1^{\mathrm{v}}) - 1) \wedge y \in L(u, d(u, A_1^{\mathrm{v}}))\}$, that is, all the edges with one endpoint at distance $d(u, A_1^{\mathrm{v}}) - 1$ from u and another endpoint at distance $d(u, A_1^{\mathrm{v}})$ from u. Consider now the sets $E^H(u)$, for every $u \in V^H$. Each such set contains at least s edges and there are at most n such sets. Thus, we can apply Lemma 1 to compute a hitting set $E^H \subseteq E$ of size $\tilde{O}(m/s)$. Let $A_1^{\mathrm{e}} = V(E^H)$. We set A_1 to $A_1^{\mathrm{v}} \cup A_1^{\mathrm{e}}$.

We now proceed with the sets $A_2, \ldots, A_{k-1}$ as in the distance oracle of Thorup and Zwick, that is, A_i is a subset of A_i of size at most $\tilde{O}(|A_{i-1}|/s)$ that hits for every $v \in V$ the set $N(v, s, A_{i-1})$. The set A_k is empty.

We use the sets $V = A_0, A_1, \ldots, A_k$ to construct the standard variant of the distance oracle. The special way we used to compute the set A_1 allows us to prove the following crucial Lemma:

Lemma 7. $\sum_{u \in V} |L(u, \ell_1(u))| = \tilde{O}(nm^{1/k})$.

Proof. Assume, towards a contradiction, that there exists $u \in V$ such that $|L(u, \ell_1(u))| > s$. Since $A_1 = A_1^{\mathrm{v}} \cup A_1^{\mathrm{e}}$ we have $\ell_1(u) = \min(d(u, A_1^{\mathrm{v}}), d(u, A_1^{\mathrm{e}}))$. It cannot be that $\ell_1(u) = d(u, A_1^{\mathrm{v}})$ because this implies that $|L(u, d(u, A_1^{\mathrm{v}}))| > s$ and $u \in V^H$. In such a case, an edge (x, y) from $E^H(u)$ is in E^H and $x \in A_1^{\mathrm{e}}$ is added to A_1. Since $d(u, A_1^{\mathrm{e}}) \leq d(u, x) = d(u, A_1^{\mathrm{v}}) - 1$ and $\ell_1(u) = \min(d(u, A_1^{\mathrm{v}}), d(u, A_1^{\mathrm{e}}))$ we get that it must be that $\ell_1(u) < d(u, A_1^{\mathrm{v}})$.

So we have $|L(u, \ell_1(u))| > s$ and $\ell_1(u) = d(u, A_1^{\mathrm{e}}) < d(u, A_1^{\mathrm{v}})$. The set A_1^{v} is a hitting set for the sets $N(v, s, V)$, for every $v \in V$. From Lemma 2 it follows that $|B(u, d(u, A_1^{\mathrm{v}}))| \leq s$. Since $\ell_1(u) = d(u, A_1^{\mathrm{e}}) < d(u, A_1^{\mathrm{v}})$ we get that $L(u, \ell_1(u)) \subseteq B(u, d(u, A_1^{\mathrm{v}}))$, a contradiction to the fact that $|L(u, \ell_1(u))| > s$. Thus, we get that $\sum_{u \in V} |L(u, \ell_1(u))| = s \cdot n = \tilde{O}(nm^{1/k})$, as required.

It follows from the above Lemma that we can save also the set $L(u, \ell_1(u))$, for every $u \in V$, in a 2-level hash table, without increasing the total size of the distance oracle.

Given a pair $u, v \in V$ the query works as follows. First, we check if $(u, v) \in E$ and if so return 1 and stop. Otherwise, we check if either $v \in L(u, \ell_1(u))$ or $u \in L(v, \ell_1(v))$ and if so return the exact distance and stop. If this is not the case we use the query of the standard variant of Thorup-Zwick distance oracle on u, v and on v, u and report the minimum of these two estimations.

Next, we analyze the stretch of the distance oracle. Let $u, v \in V$ and let $\varDelta = d(u, v)$. If $(u, v) \in E$ or $u \in B_0(v)$ or $v \in B_0(u)$ then the exact distance is returned. Therefore, we can assume that $(u, v) \notin E$, $u \notin B_0(v)$ and $v \notin B_0(u)$. Let $d(u', v) = d(u, v') = \varDelta - 1$, where $u' \in N(u)$ and $v' \in N(v)$. If $u' \in B_0(v)$ (respectively, $v' \in B_0(u)$) then $u \in L(v, \ell_1(v))$ (respectively, $v \in L(u, \ell_1(u))$) and the exact distance is returned. Therefore, we can assume also that $u' \notin B_0(v)$ and $v' \notin B_0(u)$. This implies that $\ell_1(v) \leq \varDelta - 1$ and $\ell_1(u) \leq \varDelta - 1$.

For $k = 2$ the standard variant of Thorup-Zwick distance oracle degenerates to the regular one since the additional distances stored are for pairs from $A_1 \times A_0$. The query returns $\ell_1(u) + d(v, p_1(u))$ which is bounded by $2\ell_1(u) + \varDelta$. Using the bound $\ell_1(u) \leq \varDelta - 1$ we get that the estimation is bounded by $3\varDelta - 2$, as required.

Consider now the case that $k \geq 3$. As we have checked whether $(u, v) \in E$, we can assume that $\varDelta \geq 2$. Let $f = \min\big(f(u, v), f(v, u)\big)$. In the case that $f \leq \lfloor k/2 \rfloor$ the query returns $\min(dist(u, v), dist(v, u))$. From Lemma 5 it follows that this estimation is bounded by $(2(k/2)+1)d(u, v) = (k+1)\varDelta \leq (2k-1)\varDelta - 4$ for even $k \geq 4$ and $\varDelta \geq 2$, and bounded by $(2((k-1)/2)+1)d(u, v) = k\varDelta \leq (2k-1)\varDelta - 4$ for odd $k \geq 3$ and $\varDelta \geq 2$.

For $f > \lfloor k/2 \rfloor$ the query returns $\min\big(\ell_{k/2}(u) + d(p_{k/2}(u), p_{k/2-1}(v)) + \ell_{k/2-1}(v), \ell_{k/2}(v) + d(p_{k/2}(v), p_{k/2-1}(u)) + \ell_{k/2-1}(u)\big)$, for an even k, and $\ell_{(k-1)/2}(u) + d(p_{(k-1)/2}(u), p_{(k-1)/2}(v)) + \ell_{(k-1)/2}(v)$, for an odd k.

Consider the case of an even k. Let $i = k/2$ and assume that i is even. It follows from Lemma 6 that $2\ell_i(u) + 2\ell_{i-1}(v) + d(u, v)$ is an upper bound for the estimation. From Lemma 4 we have $\ell_i(u) \leq \ell_1(v) + (i-1)\varDelta$ and $\ell_{i-1}(v) \leq \ell_1(u) + (i-2)\varDelta$. Thus, we get:

$$\begin{aligned} 2\ell_i(u) + 2\ell_{i-1}(v) + d(u, v) &\leq 2(\ell_1(v) + (i-1)\varDelta) + 2((\ell_1(u) + (i-2)\varDelta)) + \varDelta \\ &\leq 2\ell_1(u) + 2\ell_1(v) + 4i\varDelta - 5\varDelta \\ &\leq 4(\varDelta - 1) + 4(k/2)\varDelta - 5\varDelta \\ &\leq (2k-1)\varDelta - 4 \end{aligned}$$

Assume now that i is odd. It follows from Lemma 6 that $2\ell_i(v) + 2\ell_{i-1}(u) + d(u, v)$ is an upper bound for the estimation. From Lemma 4 we have $\ell_i(v) \leq \ell_1(v) + (i-1)\varDelta$ and $\ell_{i-1}(u) \leq \ell_1(v) + (i-2)\varDelta$. Thus, we get:

$$\begin{aligned} 2\ell_i(v) + 2\ell_{i-1}(u) + d(u, v) &\leq 4\ell_1(v) + 4i\varDelta - 5\varDelta \\ &\leq 4(\varDelta - 1) + 4(k/2)\varDelta - 5\varDelta \\ &\leq (2k-1)\varDelta - 4 \end{aligned}$$

Consider now the case that k is odd. Let $i = (k-1)/2$. It follows from Lemma 6 that $2\ell_i(u) + 2\ell_i(v) + d(u, v)$ is an upper bound for the estimation.

From Lemma 4 we have $\ell_i(v) \leq \ell_1(u) + (i-1)\Delta$ and $\ell_i(u) \leq \ell_1(v) + (i-1)\Delta$ if i is even or odd. Thus, we get:

$$\begin{aligned}
2\ell_i(u) + 2\ell_i(v) + d(u,v) &\leq 2(\ell_1(v) + (i-1)\Delta) + 2(\ell_1(u) + (i-1)\Delta) + \Delta \\
&\leq 4(\Delta - 1 + (i-1)\Delta) + \Delta \\
&\leq 4(i\Delta - 1) + \Delta \\
&\leq (2k-1)\Delta - 4
\end{aligned}$$

Remark. The hierarchal nature of the query algorithm that is based on the bunches induced by the sets $V = A_0, A_1, \ldots, A_k$ makes it tempting to try to apply the interplay between a hitting set of vertices and a hitting set of edges not only to A_1 but also to the sets $A_2, \ldots, A_k$. This however is not possible from the following reason. To obtain the improved bound on $\ell_1(u)$ we need that $p_{A_1}(u) \in A_1^{\mathrm{e}}$. Thus, in the next step of the query we need to check if $p_{A_1}(u) \in A_1^{\mathrm{e}}$ is in $B_2(v)$. To get a better bound now for $\ell_2(v)$ we need to be able to either save the vertices of A_1 that are at distance $\ell_2(v)$ from v, in case that there are at most s such vertices or to improve the bound on $\ell_2(v)$ by a tighter hitting set of size $\tilde{O}(m/s^2)$, if there are strictly more than s such vertices. However, in the later case, the fact that there are more than s vertices of A_1, which all might be vertices of A_1^{e}, at distance $\ell_2(v)$ does not imply that the number of edges with one endpoint at distance $\ell_2(v)-1$ from v and another endpoint at distance $\ell_2(v)$ from v is more than s^2. It might be that there are many edges (strictly more than s^2) with both endpoints at distance $\ell_2(v)$ from v. These edges can cause to strictly more than s vertices of A_1^{e} to be at distance $\ell_2(v)$ from v. On the other hand, hitting these set of edges might result with an edge whose both endpoints are at distance $\ell_2(v)$ and will not improve $\ell_2(v)$.

4 A Refined Stretch Analysis of Thorup-Zwick Distance Oracle

In this section we present several different conditions that can be easily checked and once fulfilled by the distance oracle of Thorup-Zwick guarantee that the estimation has a stretch which is strictly better than $2k-1$.

The main parameter that we use is the *average distance* between a vertex and the sets $A_1, \ldots, A_{k-1}$. We define the average distance between $u \in V$ and A_i to be $\bar{\ell}_i(u) = \lceil \ell_i(u)/i \rceil$, where $i \in [1, k-1]$.

Let $\hat{d}(u,v) = \min(dist(u,v), dist(v,u))$. We prove the following properties:

Property 1. Let $u \in V$. If $\bar{\ell}_i(u) \neq \bar{\ell}_j(u)$ for some $i,j \in [1,k-1]$ then for every $v \in V$ the stretch of $\hat{d}(u,v)$ is strictly better than $(2k-1)$.

Property 2. Let $u,v \in V$. If $\bar{\ell}_i(u) \neq \bar{\ell}_i(v)$ for some $i \in [1,k-1]$ then the stretch of $\hat{d}(u,v)$ is strictly better than $(2k-1)$.

Property 3. Let $u, v \in V$. If $\bar{\ell}_i(u) = \bar{\ell}_i(v) = q$, for every $i \in [1, k-1]$ and $d(u,v) \neq q$ then the stretch of $\hat{d}(u,v)$ is strictly better than $(2k-1)$.

Before we turn into the technical part of this section we discuss these properties. First notice to the nice relation between these properties. If the conditions of Property 1 do not hold then the conditions of Property 2 can still hold, and if the conditions of both Properties 1 and 2 do not hold then the conditions of Property 3 can still hold.

From the implementation perspective we can verify whether Property 1 and Property 2 hold using a simple computation that does not require the actual computation of the distance oracle itself. Moreover, if Property 1 does not hold then we have $\bar{\ell}_i(u) = \ell_1(u)$, for every $i \in [1, k-1]$, since $\bar{\ell}_1(u) = \ell_1(u)$. Thus, $\ell_1(u) - 1 \leq {}^{\ell_i(u)}/_i \leq \ell_1(u)$ and we get that $\ell_i(u) \in [i\ell_1(u) - i, i\ell_1(u)]$. In such a scenario the shortest paths tree of u has a relatively well defined structure in which $|B(u, \ell_1(u))| \leq n^{1/k}$ and for every $i \in [2, k-1]$ it holds that $|B(u, i\ell_1(u) - i)| \leq n^{i/k}$ and $n^{i/k} \leq |B(u, i\ell_1(u))|$. It is a plausible conjecture that such a well defined structure is not common. For the sake of completeness we do a small experiment on several different datasets of real world graphs to test how frequent these properties are. We elaborate more on this experiment in [15].

Due to lack of space, we omit the technical part of this section, which can be found in [15].

5 Concluding Remarks

In this paper we proved that for every $k \geq 2$ there is a distance oracle of size $O(knm^{1/k} \log n)$ that produces in $O(k)$ time an estimation $d^*(u,v)$ that satisfies $d(u,v) \leq d^*(u,v) \leq (2k-1)d(u,v) - 4$, for $k > 2$, and $d(u,v) \leq d^*(u,v) \leq 3d(u,v) - 2$, for $k = 2$.

An interesting open problem is whether it is possible to obtain a distance oracle with the same size and query time whose estimation $d^*(u,v)$ satisfies $d(u,v) \leq d^*(u,v) \leq (2k-1)d(u,v) - \Omega(k)$, for large enough k.

References

1. Abraham, I., Gavoille, C.: On approximate distance labels and routing schemes with affine stretch. In: Peleg, D. (ed.) DISC 2011. LNCS, vol. 6950, pp. 404–415. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-24100-0_39
2. Bernstein, A.: Fully dynamic (2 + epsilon) approximate all-pairs shortest paths with fast query and close to linear update time. In: 50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, Atlanta, Georgia, USA, 25–27 October 2009, pp. 693–702 (2009)
3. Chechik, S.: Compact routing schemes with improved stretch. In: ACM Symposium on Principles of Distributed Computing, PODC 2013, Montreal, QC, Canada, 22–24 July 2013, pp. 33–41 (2013)
4. Chechik, S.: Approximate distance oracles with constant query time. In: STOC (2014)

5. Chechik, S.: Approximate distance oracles with improved bounds. In: STOC (2015)
6. Chen, W., Sommer, C., Teng, S.-H., Wang, Y.: Compact routing in power-law graphs. In: Keidar, I. (ed.) DISC 2009. LNCS, vol. 5805, pp. 379–391. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04355-0_41
7. Henzinger, M., Krinninger, S., Nanongkai, D.: Dynamic approximate all-pairs shortest paths: breaking the O(mn) barrier and derandomization. SIAM J. Comput. **45**(3), 947–1006 (2016)
8. Lacki, J., Ocwieja, J., Pilipczuk, M., Sankowski, P., Zych, A.: The power of dynamic distance oracles: efficient dynamic algorithms for the Steiner tree. In: Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, 14–17 June 2015, pp. 11–20 (2015)
9. Mendel, M., Naor, A.: Ramsey partitions and proximity data structures. In: 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS 2006), 21–24 October 2006, Berkeley, California, USA, Proceedings, pp. 109–118 (2006)
10. Patrascu, M., Roditty, L.: Distance oracles beyond the Thorup-Zwick bound. SIAM J. Comput. **43**, 300–311 (2014)
11. Patrascu, M., Roditty, L., Thorup, M.: A new infinity of distance oracles for sparse graphs. In: FOCS (2012)
12. Qi, Z., Xiao, Y., Shao, B., Wang, H.: Toward a distance oracle for billion-node graphs. PVLDB **7**(1), 61–72 (2013)
13. Roditty, L., Tov, R.: New routing techniques and their applications. In: Proceedings of the 2015 ACM Symposium on Principles of Distributed Computing, PODC 2015, Donostia-San Sebastián, Spain, 21–23 July 2015, pp. 23–32 (2015)
14. Roditty, L., Tov, R.: Close to linear space routing schemes. Distrib. Comput. **29**(1), 65–74 (2015). https://doi.org/10.1007/s00446-015-0256-5
15. Roditty, L., Tov, R.: Approximate distance oracles with improved stretch for sparse graphs (2021). https://github.com/roei-tov/Approximate-Distance-Oracles-with-Improved-Stretch-for-Sparse-Graphs
16. Roditty, L., Zwick, U.: Dynamic approximate all-pairs shortest paths in undirected graphs. SIAM J. Comput. **41**(3), 670–683 (2012). https://doi.org/10.1137/090776573
17. Sommer, C., Verbin, E., Yu, W.: Distance oracles for sparse graphs. In: FOCS (2009)
18. Thorup, M., Zwick, U.: Compact routing schemes. In: SPAA, pp. 1–10 (2001)
19. Thorup, M., Zwick, U.: Approximate distance oracles. J. ACM **52**, 1–24 (2005)
20. Wulff-Nilsen, C.: Approximate distance oracles with improved query time. In: Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013, New Orleans, Louisiana, USA, 6–8 January 2013, pp. 539–549 (2013)