Metric random matchings with applications

Ching-Lueh Chang †‡

March 27, 2017

Abstract

Let \((\{1, 2, \ldots, n\}, d)\) be a metric space. We analyze the expected value and the variance of \(\sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i))\) for a uniformly random permutation \(\pi\) of \(\{1, 2, \ldots, n\}\), leading to the following results:

• Consider the problem of finding a point in \(\{1, 2, \ldots, n\}\) with the minimum sum of distances to all points. We show that this problem has a randomized algorithm that (1) always outputs a \((2 + \epsilon)\)-approximate solution in expected \(O(n/\epsilon^2)\) time and that (2) inherits Indyk’s [9, 10] algorithm to output a \((1 + \epsilon)\)-approximate solution in \(O(n/\epsilon^2)\) time with probability \(\Omega(1)\), where \(\epsilon \in (0, 1)\).

• The average distance in \((\{1, 2, \ldots, n\}, d)\) can be approximated in \(O(n/\epsilon)\) time to within a multiplicative factor in \([1/2 - \epsilon, 1]\) with probability \(1/2 + \Omega(1)\), where \(\epsilon > 0\).

• Assume \(d\) to be a graph metric. Then the average distance in \((\{1, 2, \ldots, n\}, d)\) can be approximated in \(O(n)\) time to within a multiplicative factor in \([1 - \epsilon, 1 + \epsilon]\) with probability \(1/2 + \Omega(1)\), where \(\epsilon = \omega(1/n^{1/4})\).

1 Introduction

A metric space is a nonempty set \(M\) endowed with a metric, i.e., a function \(d: M \times M \rightarrow [0, \infty)\) such that

- \(d(x, y) = 0\) if and only if \(x = y\) (identity of indiscernibles),
- \(d(x, y) = d(y, x)\) (symmetry), and

---

* A preliminary version of this paper appears in https://arxiv.org/abs/1702.03106.
† Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan.
‡ Supported in part by the Ministry of Science and Technology of Taiwan under grant 105-2221-E-155-047-.
\begin{itemize}
  \item \(d(x, y) + d(y, z) \geq d(x, z)\) (triangle inequality)
\end{itemize}

for all \(x, y, z \in M\) [14].

For all \(n \in \mathbb{Z}^+\), define \([n] \equiv \{1, 2, \ldots, n\}\). Given \(n \in \mathbb{Z}^+\) and oracle access to a metric \(d:\ [n] \times [n] \to [0, \infty )\), METRIC 1-MEDIAN asks for \(\arg\min_{y \in [n]} \sum_{x \in [n]} d(y, x)\), breaking ties arbitrarily. It generalizes the classical median selection on the real line and has a brute-force \(\Theta(n^2)\)-time algorithm. More generally, METRIC \(k\)-MEDIAN asks for \(c_1, c_2, \ldots, c_k \in [n]\) minimizing \(\sum_{x \in [n]} \min_{i=1}^k d(x, c_i)\). Because \(d(\cdot, \cdot)\) defines \(\binom{n}{2}\) nonzero distances, only \(o(n^2)\)-time algorithms are said to run in sublinear time [9]. For all \(\alpha \geq 1\), an \(\alpha\)-approximate 1-median is a point \(p \in [n]\) satisfying

\[
\sum_{x \in [n]} d(p, x) \leq \alpha \cdot \min_{y \in [n]} \sum_{x \in [n]} d(y, x).
\]

For all \(\epsilon > 0\), METRIC 1-MEDIAN has a Monte Carlo \((1 + \epsilon)\)-approximation \(O(n/\epsilon^2)\)-time algorithm [9, 10]. Guha et al. [8] show that METRIC \(k\)-MEDIAN has a Monte Carlo, \(O(\exp(O(1/\epsilon)))\)-approximation, \(O(nk \log n)\)-time, \(O(n^\epsilon)\)-space and one-pass algorithm for all small \(k\) as well as a deterministic, \(O(\exp(O(1/\epsilon)))\)-approximation, \(O(n^{1+\epsilon})\)-time, \(O(n^\epsilon)\)-space and one-pass algorithm. Given \(n\) points in \(\mathbb{R}^D\) with \(D \geq 1\), the Monte Carlo algorithms of Kumar et al. [11] find a \((1+\epsilon)\)-approximate 1-median in \(O(D \cdot \exp(1/\epsilon^{O(1)}))\) time and a \((1+\epsilon)\)-approximate solution to METRIC \(k\)-MEDIAN in \(O(Dn \cdot \exp((k/\epsilon)^{O(1)}))\) time. All randomized \(O(1)\)-approximation algorithms for METRIC \(k\)-MEDIAN take \(\Omega(nk)\) time [8, 12]. Chang [3] shows that METRIC 1-MEDIAN has a deterministic, \((2h)\)-approximation, \(O(hn^{1+1/h})\)-time and nonadaptive algorithm for all constants \(h \in \mathbb{Z}^+ \setminus \{1\}\), generalizing the results of Chang [2] and Wu [16]. On the other hand, he disproves the existence of deterministic \((2h - \epsilon)\)-approximation \(O(n^{1+1/(h-1)/h})\)-time algorithms for all constants \(h \in \mathbb{Z}^+ \setminus \{1\}\) and \(\epsilon > 0\) [4, 5].

In social network analysis, the closeness centrality of a point \(v\) is the reciprocal of the average distance from \(v\) to all points [15]. So METRIC 1-MEDIAN asks for a point with the maximum closeness centrality. Given oracle access to a graph metric, the Monte-Carlo algorithms of Goldreich and Ron [7] and Eppstein and Wang [6] estimate the closeness centrality of a given point and those of all points, respectively.

All known sublinear-time algorithms for METRIC 1-MEDIAN are either deterministic or Monte Carlo, the latter having a positive probability of failure. For example, Indyk’s Monte Carlo \((1 + \epsilon)\)-approximation algorithm outputs with a positive probability a solution without approximation guarantees. In contrast, we show that METRIC 1-MEDIAN has a randomized algorithm that \textit{always} outputs a \((2 + \epsilon)\)-approximate solution in expected \(O(n/\epsilon^2)\) time for all \(\epsilon \in (0, 1)\). So, excluding the known deterministic algorithms (which are Las Vegas only in the degenerate sense), this paper gives the \textit{first} Las Vegas approximation algorithm for METRIC 1-MEDIAN with an expected sublinear running time. Note that deterministic sublinear-time algorithms for METRIC 1-MEDIAN can be \(4\)-approximate
but not \((4 - \epsilon)\)-approximate for any constant \(\epsilon > 0\) \([2, 5]\). So our approximation ratio of \(2 + \epsilon\) beats that of any deterministic sublinear-time algorithm. Inheriting Indyk’s algorithm, our algorithm outputs a \((1 + \epsilon)\)-approximate 1-median in \(O(n/\epsilon^2)\) time with probability \(\Omega(1)\) for all \(\epsilon \in (0, 1)\).

Indyk \([9, 10]\) gives a Monte-Carlo \(O(n/\epsilon^{3.5})\)-time algorithm that approximates the average distance in any metric space \([(n),d]\) to within a multiplicative factor in \([1/2 - \epsilon, 1]\), for all \(\epsilon > 0\). Barhum, Goldreich and Shraibman \([1]\) improve Indyk’s time complexity of \(O(n/\epsilon^{3.5})\) to \(O(n/\epsilon^2)\). This paper gives a Monte-Carlo \(O(n/\epsilon)\)-time algorithm that approximates the average distance in \([(n),d]\) to within a multiplicative factor in \([1 - \epsilon, 1 + \epsilon]\), for all \(\epsilon > 0\). For all \(\epsilon = \omega(1/n^{1/4})\), we present a Monte-Carlo \(O(n)\)-time algorithm approximating the average distance of any graph metric to within a multiplicative factor in \([1 - \epsilon, 1 + \epsilon]\). But for general metrics, we do not know whether the \(O(n/\epsilon^2)\) running time of Barhum, Goldreich and Shraibman can be improved to \(O(n/\epsilon^2 - \Omega(1))\).

2 Definitions and preliminaries

For a metric space \([(n),d]\),

\[
\bar{r} \equiv \frac{1}{n^2} \sum_{x,y \in [n]} d(x,y),
\]

(1)

\[
p^* \equiv \arg\min_{p \in [n]} \sum_{x \in [n]} d(p,x),
\]

(2)

breaking ties arbitrarily in equation (2). So \(\bar{r}\) is the average distance in \([(n),d]\), and \(p^*\) is a 1-median.

An algorithm \(A\) with oracle access to \(d\): \([n] \times [n] \to [0, \infty)\) is denoted by \(A^d\) and may query \(d\) on any \((x, y) \in [n] \times [n]\) for \(d(x, y)\). In this paper, all Landau symbols (such as \(O(\cdot), o(\cdot), \Omega(\cdot)\) and \(\omega(\cdot)\)) are w.r.t. \(n\). The following result is due to Indyk.

**Fact 1** \([9, 10]\). For all \(\epsilon > 0\), metric 1-median has a Monte Carlo \((1 + \epsilon)\)-approximation \(O(n/\epsilon^2)\)-time algorithm with a failure probability of at most \(1/e\).

Henceforth, denote Indyk’s algorithm in Fact 1 by **Indyk median**. It is given \(n \in \mathbb{Z}^+, \epsilon > 0\) and oracle access to a metric \(d\): \([n] \times [n] \to [0, \infty)\). The following fact on estimating the average distance is due to Barhum, Goldreich and Shraibman.

**Fact 2** \([1]\). Given \(n \in \mathbb{Z}^+, \epsilon > 0\) and oracle access to a metric \(d\): \([n] \times [n] \to [0, \infty)\), a real number in \([(1 - \epsilon)\bar{r}, (1 + \epsilon)\bar{r}]\) can be found in \(O(n/\epsilon^2)\) time with probability at least \(1/2 + \Omega(1)\).

**Chebyshev’s inequality** \([13]\). Let \(X\) be a random variable with a finite expected value and a finite nonzero variance. Then for all \(k \geq 1\),

\[
\Pr \left[ |X - \mathbb{E}[X]| \geq k\sqrt{\text{var}(X)} \right] \leq \frac{1}{k^2}.
\]
1: while true do
2: \[ z \leftarrow \text{Indyk median}^d(n, \varepsilon/8); \]
3: \[ \text{Pick independent and uniformly random permutations } \pi_1, \pi_2, \ldots, \pi_{80[1/\varepsilon]} : [n] \rightarrow [n]; \]
4: \[ \text{if there exists } j \in [\lceil 1/\varepsilon \rceil] \text{ satisfying } \sum_{x \in [n]} d(z, x) \leq (2 + \varepsilon) \sum_{i=1}^{\lceil n/2 \rceil} d(\pi_j(2i - 1), \pi_j(2i)) \text{ then} \]
5: \[ \text{return } z; \]
6: end if
7: end while

Figure 1: Algorithm Las Vegas median with oracle access to a metric \( d: [n] \times [n] \rightarrow [0, \infty) \) and with inputs \( n \in \mathbb{Z}^+ \) and \( \varepsilon \in (0, 1) \)

3 Las Vegas approximation for metric 1-median selection

This section presents a randomized algorithm that always outputs a \((2 + \varepsilon)\)-approximate 1-median, where \( \varepsilon \in (0, 1) \). Clearly,

\[
\sum_{x \in [n]} d(p^*, x) \overset{(2)}{=} \min_{p \in [n]} \sum_{x \in [n]} d(p, x) \leq \frac{1}{n} \sum_{p \in [n]} \sum_{x \in [n]} d(p, x) \overset{(1)}{=} n\bar{r}. \tag{3}
\]

For each permutation \( \pi : [n] \rightarrow [n] \),

\[
\sum_{i=1}^{\lceil n/2 \rceil} d(\pi(2i - 1), \pi(2i)) \leq \sum_{i=1}^{\lceil n/2 \rceil} d(p^*, \pi(2i - 1)) + d(p^*, \pi(2i)) \leq \sum_{x \in [n]} d(p^*, x), \tag{4}
\]

where the first and the second inequalities follow from the triangle inequality and the injectivity of \( \pi \).

Lemma 3. When line 5 of Las Vegas median in Fig. 1 is run, \( z \) is a \((2 + \varepsilon)\)-approximate 1-median.

Proof. The condition in line 4 of Las Vegas median implies

\[
\sum_{x \in [n]} d(z, x) \overset{(4)}{\leq} (2 + \varepsilon) \sum_{x \in [n]} d(p^*, x) \overset{(2)}{=} (2 + \varepsilon) \min_{p \in [n]} \sum_{x \in [n]} d(p, x).
\]

So when line 5 is run, it returns a \((2 + \varepsilon)\)-approximate 1-median. \qed

Inequalities (3)–(4) yield the following.

Lemma 4. For each permutation \( \pi : [n] \rightarrow [n] \),

\[
\sum_{i=1}^{\lceil n/2 \rceil} d(\pi(2i - 1), \pi(2i)) \leq n\bar{r}.
\]
Lemma 5. For a uniformly random permutation \( \pi : [n] \rightarrow [n] \),

\[
E_{\pi} \left[ \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right] = \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{n\bar{r}}{n-1}.
\]

Proof. For each \( i \in [\lfloor n/2 \rfloor] \), \( \{\pi(2i-1), \pi(2i)\} \) is a uniformly random size-2 subset of \([n]\), implying

\[
E_{\pi} \left[ d(\pi(2i-1), \pi(2i)) \right] = \frac{1}{n \cdot (n-1)} \cdot \sum_{\text{distinct } x, y \in [n]} d(x, y) = \frac{1}{n \cdot (n-1)} \cdot \sum_{x, y \in [n]} d(x, y) \quad \text{(1)}
\]

where the second equality follows from the identity of indiscernibles. Finally, use the linearity of expectation. \( \square \)

Lemma 6. For all \( \epsilon \in (0, 1) \) and in each iteration of the while loop of Las Vegas median,

\[
\Pr \left[ \exists j \in \left[ 80 \cdot \left\lceil \frac{1}{\epsilon} \right\rceil \right], \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi_j(2i-1), \pi_j(2i)) \geq \left( \frac{1}{2} - \frac{\epsilon}{8} \right) n\bar{r} \right] \geq 0.9, \quad (5)
\]

where the probability is taken over \( \pi_1, \pi_2, \ldots, \pi_{80\lceil 1/\epsilon \rceil} \) in line 3 of Las Vegas median.

Proof. Let \( \pi : [n] \rightarrow [n] \) be a uniformly random permutation and

\[
\alpha = \Pr_{\pi} \left[ \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \geq \left( \frac{1}{2} - \frac{\epsilon}{8} \right) n\bar{r} \right]. \quad (6)
\]

So by Lemma 4,

\[
E_{\pi} \left[ \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right] \leq \alpha n\bar{r} + (1 - \alpha) \left( \frac{1}{2} - \frac{\epsilon}{8} \right) n\bar{r}.
\]

This and Lemma 5 imply \( \alpha \geq \epsilon/8 \). So the left-hand side of inequality (5) is at least \( 1 - (1 - \epsilon/8)^{80\lceil 1/\epsilon \rceil} \geq 0.9 \). \( \square \)
Lemma 7. For all \( \epsilon \in (0, 1) \) and in each iteration of the while loop of Las Vegas median,

\[
\Pr \left[ \left( \sum_{x \in [n]} d(z, x) \leq \left( 1 + \frac{\epsilon}{8} \right) \min_{p \in [n]} \sum_{x \in [n]} d(p, x) \right)
\land \left( \exists j \in \left[ 80 \cdot \left[ \frac{1}{\epsilon} \right] \right], \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi_j (2i - 1), \pi_j (2i)) \geq \left( \frac{1}{2} - \frac{\epsilon}{8} \right) n \bar{r} \right)
\land \left( \exists j \in \left[ 80 \cdot \left[ \frac{1}{\epsilon} \right] \right], \sum_{x \in [n]} d(z, x) \leq (2 + \epsilon) \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi_j (2i - 1), \pi_j (2i)) \right) \right] 
= \frac{1}{2} + \Omega(1),
\]

where the probability is taken over \( \pi_1, \pi_2, \ldots, \pi_{80[1/\epsilon]} \) and the random coin tosses of Indyk median.

Proof. By Fact 1 and line 2 of Las Vegas median, the first condition within \( \Pr[\cdot] \) in equation (7) holds with probability at least \( 1 - 1/e \) over the random coin tosses of Indyk median. By Lemma 6, the second condition holds with probability at least \( 0.9 \) over \( \pi_1, \pi_2, \ldots, \pi_{80[1/\epsilon]} \). In summary, the first two conditions hold simultaneously with probability at least \( (1 - 1/e) \cdot 0.9 = 1/2 + \Omega(1) \) (note that the random coin tosses of Indyk median are independent of \( \pi_1, \pi_2, \ldots, \pi_{80[1/\epsilon]} \)). Finally, the first two conditions together imply the third by inequality (3) and the easy fact that

\[
\left( 1 + \frac{\epsilon}{8} \right) \leq (2 + \epsilon) \left( \frac{1}{2} - \frac{\epsilon}{8} \right).
\]

\[\square\]

Theorem 8. For all \( \epsilon \in (0, 1) \), Metric 1-Median has a randomized algorithm that (1) always outputs a \( (2 + \epsilon) \)-approximate solution in expected \( O(n/\epsilon^2) \) time and (2) outputs a \( (1 + \epsilon) \)-approximate solution in \( O(n/\epsilon^2) \) time with probability \( \Omega(1) \).

Proof. By Lemma 7, each execution of lines 4–5 of Las Vegas median returns with probability \( 1/2 + \Omega(1) \). So the expected number of iterations is \( O(1) \). By Fact 1, line 2 takes \( O(n/\epsilon^2) \) time. Line 3 takes \( 80[1/\epsilon] \cdot O(n) \) time by the Knuth shuffle. Clearly, lines 4–5 take \( O(n/\epsilon) \) time. In summary, the expected running time of Las Vegas median is \( O(1) \cdot O(n/\epsilon^2) = O(n/\epsilon^2) \). To prevent Las Vegas median from running forever, find a 1-median by brute force (which obviously takes \( O(n^2) \) time) after \( n^2 \) steps of computation. By Lemma 3, Las Vegas median is \( (2 + \epsilon) \)-approximate.
By Lemma 7, z is \((1 + \epsilon/8)\)-approximate and is also returned in line 5 with probability \(\Omega(1)\) in the first (in fact, any) iteration. Finally, the previous paragraph has shown each iteration to take \(O(n/\epsilon^2)\) time. \(\square\)

By Fact 1, indyk median satisfies condition (2) in Theorem 8. But it does not satisfy condition (1).

We now justify the optimality of the ratio of \(2 + \epsilon\) in Theorem 8. Let \(A\) be a randomized algorithm that always outputs a \((2 - \epsilon)\)-approximate 1-median. Furthermore, denote by \(p \in [n]\) (resp., \(Q \subseteq [n] \times [n]\)) the output (resp., the set of queries as unordered pairs) of \(A^{d_1}(n)\), where \(d_1\) is the discrete metric (i.e., \(d_1(x,y) = 1\) and \(d_1(x,x) = 0\) for all distinct \(x, y \in [n]\)). Without loss of generality, assume \((p,y) \in Q\) for all \(y \in [n] \setminus \{p\}\) by adding dummy queries. So the queries in \(Q\) witness that

\[
\sum_{y \in [n] \setminus \{p\}} d_1(p,y) = n - 1. \tag{8}
\]

Assume without loss of generality that \(A\) never queries for the distance from a point to itself.

In the sequel, consider the case that \(|Q| < \epsilon \cdot (n-1)^2/8\). By the averaging argument, there exists a point \(\hat{p} \in [n] \setminus \{p\}\) involved in at most \(2 \cdot |Q|/(n-1)\) queries in \(Q\) (note that each query involves two points). Because every function \(f: [n] \times [n] \to [0, \infty)\) with

\[
\{f(x,y) \mid (x,y \in [n]) \land (x \neq y)\} \subseteq \left\{\frac{1}{2}, 1\right\}
\]

satisfies the triangle inequality, \(A\) cannot exclude the possibility that \(d_1(\hat{p},y) = 1/2\) for all \(y \in [n] \setminus \{\hat{p}\}\) satisfying \((\hat{p},y) \notin Q\). In summary, \(A\) cannot rule out the case that

\[
\sum_{y \in [n]} d_1(\hat{p},y) \leq \frac{2 \cdot |Q|}{n-1} \cdot 1 + \left(n - 1 - \frac{2 \cdot |Q|}{n-1}\right) \cdot \frac{1}{2} < \left(\frac{1}{2} + \frac{\epsilon}{8}\right) \cdot (n-1). \tag{9}
\]

Equations (8)–(9) contradict the guarantee that \(p\) is \((2 - \epsilon)\)-approximate. Consequently, the case that \(|Q| < \epsilon \cdot (n-1)^2/8\) should never happen. The next theorem summarizes the above.

**Theorem 9.** Metric 1-median has no randomized algorithm that always outputs a \((2 - \epsilon)\)-approximate solution and that makes fewer than \(\epsilon \cdot (n-1)^2/8\) queries with a positive probability given oracle access to the discrete metric, for any constant \(\epsilon \in (0, 1)\).

Lemmas 4 and 6 yield the following estimation of the average distance.
Theorem 10. Given \( n \in \mathbb{Z}^+ \), \( \epsilon > 0 \) and oracle access to a metric \( d: [n] \times [n] \to [0, \infty) \), a real number in \( [(1/2 - \epsilon)\bar{r}, \bar{r}] \) can be found in \( O(n/\epsilon) \) time with probability \( 1/2 + \Omega(1) \).

Proof. By Lemmas 4 and 6,
\[
\frac{1}{n} \cdot \max_{j \in [80 \cdot \lceil 1/\epsilon \rceil]} \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi_j(2i-1), \pi_j(2i)) \in \left[ \left( \frac{1}{2} - \frac{\epsilon}{8} \right) \bar{r}, \bar{r} \right] \tag{10}
\]
with probability \( 1/2 + \Omega(1) \). The Knuth shuffle picks \( \pi_1, \pi_2, \ldots, \pi_{80\lceil 1/\epsilon \rceil} \) in \( 80\lceil 1/\epsilon \rceil \cdot O(n) \) time. Then the left-hand side of relation (10) can be calculated in \( O(n/\epsilon) \) time.

Note that the estimation of the average distance in Theorem 10 has only one-sided error. The time complexity (resp., approximation ratio) in Theorem 10 is better (resp., worse) than that in Fact 2.

4 Estimating the average distance of a graph metric

Throughout this section, take any \( \epsilon = \omega(1/n^{1/4}) \) less than a small constant, e.g., \( \epsilon = 10^{-100} \). Define
\[
\delta \equiv \frac{\epsilon^2}{10^{10}}, \tag{11}
\]
\[
r \equiv \frac{1}{n} \cdot \sum_{x \in [n]} d(p^*, x), \tag{12}
\]
where \( p^* \) is as in equation (2). As \( \epsilon = \omega(1/n^{1/4}) \), \( \delta = \omega(1/\sqrt{n}) \) by equation (11).

Lemma 11. \( \bar{r} \leq 2r \).

Proof. By equation (1) and the triangle inequality,
\[
\bar{r} \leq \frac{1}{n^2} \cdot \sum_{x,y \in [n]} \left( d(p^*, x) + d(p^*, y) \right)
\]
\[
= \frac{1}{n^2} \cdot \frac{n}{n} \cdot \left( \sum_{x \in [n]} d(p^*, x) + \sum_{y \in [n]} d(p^*, y) \right)
\]
\[
= \frac{2}{n} \cdot \sum_{x \in [n]} d(p^*, x). \tag{13}
\]
Equations (12)–(13) complete the proof. \( \square \)
1: Pick a uniformly random permutation \( \pi : [n] \to [n] \);
2: return \( \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \cdot 2/n \);

Figure 2: Algorithm average distance with oracle access to a metric \( d : [n] \times [n] \to [0, \infty) \) and with inputs \( n \in \mathbb{Z}^+ \) and \( \epsilon = \omega(1/n^{1/4}) \).

As in line 1 of average distance in Fig. 2, let \( \pi : [n] \to [n] \) be a uniformly random permutation. Clearly,

\[
\mathbb{E}_{\pi} \left[ \left( \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \right)^2 \right] \tag{14}
\]

\[= \mathbb{E}_{\pi} \left[ \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)) \cdot \sum_{j=1}^{\lfloor n/2 \rfloor} d(\pi(2j-1), \pi(2j)) \right] \]

\[= \sum_{\text{distinct } i, j = 1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d(\pi(2i-1), \pi(2i)) \cdot d(\pi(2j-1), \pi(2j)) \right] + \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d^2(\pi(2i-1), \pi(2i)) \right], \tag{15}
\]

where the last equality follows from the linearity of expectation and the separation of pairs \((i, j)\) according to whether \(i = j\). The next three lemmas analyze the variance of

\[\sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i-1), \pi(2i)).\]

**Lemma 12.**

\[
\sum_{\text{distinct } i, j = 1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d(\pi(2i-1), \pi(2i)) \cdot d(\pi(2j-1), \pi(2j)) \right] \leq \frac{1}{4} \cdot \left( 1 + O \left( \frac{1}{n} \right) \right) n^2 \bar{r}^2.
\]

**Proof.** Pick any distinct \(i, j \in [\lfloor n/2 \rfloor]\). Clearly,

\[\{ \pi(2i-1), \pi(2i), \pi(2j-1), \pi(2j) \}\]

is a uniformly random size-4 subset of \([n]\). So

\[
\mathbb{E}_{\pi} \left[ d(\pi(2i-1), \pi(2i)) \cdot d(\pi(2j-1), \pi(2j)) \right]
\]

\[= \frac{1}{n \cdot (n-1) \cdot (n-2) \cdot (n-3)} \cdot \sum_{\text{distinct } u, v, x, y \in [n]} d(u, v) \cdot d(x, y).
\]
Clearly,
\[
\sum_{\text{distinct } u, v, x, y \in [n]} d(u, v) \cdot d(x, y) \leq \sum_{u, v \in [n]} d(u, v) \cdot \sum_{x, y \in [n]} d(x, y) = \left( \sum_{x, y \in [n]} d(x, y) \right)^2.
\]

In summary,
\[
\sum_{i, j = 1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d(\pi(2i - 1), \pi(2i)) \cdot d(\pi(2j - 1), \pi(2j)) \right] \\
\leq \left[ \frac{n}{2} \right] \left( \left[ \frac{n}{2} \right] - 1 \right) \cdot \frac{1}{n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3)} \cdot \left( \sum_{x, y \in [n]} d(x, y) \right)^2.
\]

This and equation (1) complete the proof. \( \square \)

Define
\[
\Delta \equiv \max_{x, y \in [n]} d(x, y)
\]
to be the diameter of \(([n], d)\).

**Lemma 13.** If
\[
\delta nr \geq \Delta
\]
then
\[
\sum_{i = 1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d^2(\pi(2i - 1), \pi(2i)) \right] \leq \left( \frac{1}{2} + O \left( \frac{1}{n} \right) \right) \left( \delta n^2 r^2 + \delta^2 nr^2 \right). \tag{17}
\]

**Proof.** Clearly, \( \{ \pi(2i - 1), \pi(2i) \} \) is a uniformly random size-2 subset of \([n]\) for each \( i \in \lfloor n/2 \rfloor \). Therefore,
\[
\sum_{i = 1}^{\lfloor n/2 \rfloor} \mathbb{E}_{\pi} \left[ d^2(\pi(2i - 1), \pi(2i)) \right] = \sum_{i = 1}^{\lfloor n/2 \rfloor} \frac{1}{n \cdot (n - 1)} \cdot \sum_{\text{distinct } x, y \in [n]} d^2(x, y) \tag{18}
\]
\[
\leq \sum_{i = 1}^{\lfloor n/2 \rfloor} \frac{1}{n \cdot (n - 1)} \cdot \sum_{x, y \in [n]} d^2(x, y) = \left[ \frac{n}{2} \right] \cdot \frac{1}{n \cdot (n - 1)} \cdot \sum_{x, y \in [n]} d^2(x, y).
\]
By inequality (16),
\[ d(x, y) \leq \delta nr \] (19)

for all \( x, y \in [n] \).

By equations (1) and (18)–(19), the left-hand side of inequality (17) cannot exceed the optimal value of the following problem, called MAX SQUARE SUM:

Find \( d_{x,y} \in \mathbb{R} \) for all \( x, y \in [n] \) to maximize
\[ \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n - 1)} \cdot \sum_{x,y\in[n]} d_{x,y}^2 \] (20)

subject to
\[ \frac{1}{n^2} \cdot \sum_{x,y\in[n]} d_{x,y} = \bar{r}, \] (21)
\[ \forall x, y \in [n], \; 0 \leq d_{x,y} \leq \delta nr. \] (22)

Above, constraint (21) (resp., (22)) mimics equation (1) (resp., inequality (19) and the non-negativeness of distances). Appendix A bounds the optimal value of MAX SQUARE SUM from above by
\[ \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n - 1)} \cdot \left( \left\lfloor \frac{n\bar{r}}{\delta r} \right\rfloor + 1 \right) \cdot (\delta nr)^2. \]

This evaluates to be at most
\[ \left( \frac{1}{2} + O \left( \frac{1}{n} \right) \right) \left( \delta n^2 r\bar{r} + \delta^2 nr^2 \right). \]

Recall that the variance of any random variable \( X \) equals \( \text{E}[X^2] - (\text{E}[X])^2 \).

**Lemma 14.** If \( \delta nr \geq \Delta \), then
\[ \var_{\pi} \left( \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i - 1), \pi(2i)) \right) \leq (1 + o(1)) \delta n^2 r^2. \]

**Proof.** By equations (14)–(15) and Lemmas 12–13,
\[
\text{E}_{\pi} \left[ \left( \sum_{i=1}^{\lfloor n/2 \rfloor} d(\pi(2i - 1), \pi(2i)) \right)^2 \right] \\
\leq \frac{1}{4} \cdot \left( 1 + O \left( \frac{1}{n} \right) \right) n^2 r^2 + \left( \frac{1}{2} + O \left( \frac{1}{n} \right) \right) \left( \delta n^2 r\bar{r} + \delta^2 nr^2 \right).
\]
This and Lemma 5 imply
\[
\var_{\pi} \left( \sum_{i=1}^{[n/2]} d(\pi(2i-1),\pi(2i)) \right) 
\leq O\left( \frac{1}{n} \right) \cdot n^2 \bar{r}^2 + \left( \frac{1}{2} + O\left( \frac{1}{n} \right) \right) (\delta n^2 r\bar{r} + \delta^2 n\bar{r}^2).
\]

Finally, invoke Lemma 11 and recall that \( \delta = \omega(1/\sqrt{n}) \).

Lemma 15. If \( \delta nr \geq \Delta \), then
\[
\Pr_{\pi} \left[ \left| \left( \sum_{i=1}^{[n/2]} d(\pi(2i-1),\pi(2i)) \right) - \frac{1}{2} \cdot \left( 1 \pm O\left( \frac{1}{n} \right) \right) n\bar{r} \right| \geq k\sqrt{(1+o(1))\delta nr} \right] \leq \frac{1}{k^2}
\]
for all \( k > 1 \).

Proof. Use Chebyshev’s inequality and Lemmas 5 and 14.

Lemma 16. If \( \delta nr \geq \Delta \), then
\[
\Pr_{\pi} \left[ \left| \left( \sum_{i=1}^{[n/2]} d(\pi(2i-1),\pi(2i)) \right) - \frac{1}{2} \cdot \left( 1 \pm O\left( \frac{1}{n} \right) \right) n\bar{r} \right| \geq k\sqrt{(1+o(1))\delta nr} \right] \leq \frac{1}{k^2}
\]
for all \( k > 1 \).

Proof. By inequalities (3) and (12),
\[
r \leq \bar{r}.
\]

This and Lemma 15 complete the proof.

We now arrive at an efficient estimation of the average distance on a graph.

Theorem 17. Given \( n \in \mathbb{Z}^+ \), \( \epsilon = \omega(1/n^{1/4}) \) and oracle access to a graph metric \( d: [n] \times [n] \to \mathbb{N} \), a real number in \( [(1-\epsilon)\bar{r},(1+\epsilon)\bar{r}] \) can be found in \( O(n) \) time with probability \( 1/2 + \Omega(1) \).

Proof. Let \( G = ([n],E) \) be an undirected unweighted graph inducing the distance function \( d \). Then pick \( x,y \in [n] \) with \( d(x,y) = \Delta \), i.e., \( (x,y) \) is a furthest pair of vertices of \( G \). Find a simple shortest \( x-y \) path, denoted \( (v_0 = x, v_1, \ldots, v_\Delta = y) \), in \( G \). By equation (12),
\[
r \geq \frac{1}{n} \cdot \sum_{i=0}^{\Delta} d(p^*,v_i).
\]
Now,
\[
\sum_{i=0}^{\Delta} d(p^*, v_i) = \frac{1}{2} \cdot \sum_{i=0}^{\Delta} d(p^*, v_i) + d(p^*, v_{\Delta-i}) \geq \frac{1}{2} \cdot \sum_{i=0}^{\Delta} d(v_i, v_{\Delta-i}) = \frac{1}{2} \cdot \sum_{i=0}^{\Delta} |\Delta - 2i| \geq \frac{\Delta^2}{4},
\] (24)
where the first inequality (resp., the second equality) follows from the triangle inequality (resp., \((v_0, v_1, \ldots, v_\Delta)\) being a shortest \(v_0-v_\Delta\) path).\(^1\) By inequalities (23)–(24),
\[
nr \geq \frac{\Delta^2}{4}.
\] (25)
Because \(d\) is a graph metric, \(d(x, y) \geq 1\) for all distinct \(x, y \in [n]\). So by equation (12),
\[
r \geq \frac{1}{n} \cdot \sum_{x \in [n] \setminus \{p^*\}} 1 \geq \frac{1}{2}
\] (26)
for all \(n \geq 2\).

By inequalities (25)–(26),
\[
\delta nr \geq \delta \cdot \max \left\{ \frac{\Delta^2}{4}, \frac{n}{2} \right\}.
\]
So
\[
\delta nr \geq \Delta
\] (27)
for all sufficiently large \(n\).\(^2\) By equation (11),
\[
3 \sqrt{(1 + o(1))} \delta \leq 0.1 \epsilon
\] (28)
for all sufficiently large \(n\). By inequalities (27)–(28), Lemma 16 with \(k = 3\) and recalling that \(\epsilon = \omega(1/n^{1/4})\),
\[
\Pr_{\pi} \left[ \sum_{i=1}^{[n/2]} d(\pi(2i - 1), \pi(2i)) \in \left[ \left( \frac{1}{2} - \frac{\epsilon}{2} \right) n\bar{r}, \left( \frac{1}{2} + \frac{\epsilon}{2} \right) n\bar{r} \right] \right] \geq 1 - \frac{1}{9}
\] (29)
for all sufficiently large \(n\). Consequently, the output of line 2 of **average distance** in Fig. 2 is in \([(1 - \epsilon)\bar{r}, (1 + \epsilon)\bar{r}]\) with probability \(1/2 + \Omega(1)\). Line 1 takes \(O(n)\) time by the Knuth shuffle. Clearly, line 2 also takes \(O(n)\) time.

The time complexity of \(O(n)\) in Theorem 17 is independent of \(\epsilon\). But for general metrics, we do not know whether the time complexity of \(O(n/\epsilon^2)\) in Fact 2 can be improved to \(O(n/\epsilon^2 - o(1))\).

\(^1\)It is easy to verify that \(\sum_{i=0}^{\Delta} |\Delta - 2i| = (\Delta + 2)\Delta/2\) if \(\Delta \equiv 0 \pmod{2}\) and \(\sum_{i=0}^{\Delta} |\Delta - 2i| = (\Delta + 1)^2/2\) otherwise.

\(^2\)If \(\Delta \geq 4/\delta\), then \(\delta \Delta^2/4 \geq \Delta\). Otherwise, \(\delta n/2 \geq \Delta\) for all \(n > 8/\delta^2\). Finally, recall that \(\delta = \omega(1/\sqrt{n})\).
A Analyzing MAX SQUARE SUM

MAX SQUARE SUM has an optimal solution, denoted \( \{ \tilde{d}_{x,y} \in \mathbb{R} \}_{x,y \in [n]} \), because its feasible solutions (i.e., those satisfying constraints (21)–(22)) form a closed and bounded subset of \( \mathbb{R}^{(n^2)} \). (Recall from elementary mathematical analysis that a continuous real-valued function on a closed and bounded subset of \( \mathbb{R}^k \) has a maximum value, where \( k < \infty \).) Note that \( \{ \tilde{d}_{x,y} \in \mathbb{R} \}_{x,y \in [n]} \) must be feasible to MAX SQUARE SUM. Below is a consequence of constraint (21).

Lemma A.1.

\[
\left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} = \delta nr\} \right| \leq \left\lfloor \frac{n \bar{r}}{\delta r} \right\rfloor. \tag{30}
\]

Proof. Clearly,

\[
n^2 \bar{r} \sum_{x,y \in [n]} \tilde{d}_{x,y} \geq \left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} = \delta nr\} \right| \cdot \delta nr.
\]

Furthermore, the left-hand side of inequality (30) is an integer. \( \square \)

Lemma A.2.

\[
\left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} > 0\} \right| \leq \left\lfloor \frac{n \bar{r}}{\delta r} \right\rfloor + 1.
\]

Proof. Assume otherwise. Then

\[
\begin{align*}
&\left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} > 0 \land \tilde{d}_{x,y} \neq \delta nr\} \right| \\
\geq & \left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} > 0\} \right| - \left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} = \delta nr\} \right| \\
\geq & \left\lfloor \frac{n \bar{r}}{\delta r} \right\rfloor + 2 - \left| \{(x, y) \in [n] \times [n] \mid \tilde{d}_{x,y} = \delta nr\} \right| \\
\geq & 2.
\end{align*}
\]

(31)

So by constraint (22) (and the feasibility of \( \{ \tilde{d}_{x,y} \}_{x,y \in [n]} \) to MAX SQUARE SUM),

\[
\left| \{(x, y) \in [n] \times [n] \mid 0 < \tilde{d}_{x,y} < \delta nr\} \right| \geq 2.
\]

Consequently, there exist distinct \( (x', y'), (x'', y'') \in [n] \times [n] \) satisfying

\[
0 < \tilde{d}_{x',y'} < \tilde{d}_{x'',y''} < \delta nr. \tag{31}
\]

By symmetry, assume \( \tilde{d}_{x',y'} \geq \tilde{d}_{x'',y''} \). By inequality (31), there exists a small real number \( \beta > 0 \) such that increasing \( \tilde{d}_{x',y'} \) by \( \beta \) and simultaneously decreasing

14
\( \tilde{d}_{x',y''} \) by \( \beta \) will preserve constraints (21)–(22). I.e., the solution \( \{\hat{d}_{x,y} \in \mathbb{R}\}_{x,y \in [n]} \) defined below is feasible to MAX SQUARE SUM:

\[
\hat{d}_{x,y} = \begin{cases} 
\tilde{d}_{x',y'} + \beta, & \text{if } (x,y) = (x',y'), \\
\tilde{d}_{x'',y''} - \beta, & \text{if } (x,y) = (x'',y''), \\
\tilde{d}_{x,y}, & \text{otherwise}.
\end{cases}
\]

(32)

Clearly, objective (20) w.r.t. \( \{\hat{d}_{x,y} \}_{x,y \in [n]} \) exceeds that w.r.t. \( \{\tilde{d}_{x,y} \}_{x,y \in [n]} \) by

\[
\left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \sum_{x,y \in [n]} \left( \hat{d}_{x,y}^2 - \tilde{d}_{x,y}^2 \right)
\]

\[
= \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \left( (\tilde{d}_{x',y'} + \beta)^2 + (\tilde{d}_{x'',y''} - \beta)^2 - \tilde{d}_{x',y'}^2 - \tilde{d}_{x'',y''}^2 \right)
\]

\[
= \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \left( 2\beta \tilde{d}_{x',y'} - 2\beta \tilde{d}_{x'',y''} + 2\beta^2 \right)
\]

\[
> 0,
\]

where the inequality holds because \( \tilde{d}_{x',y'} \geq \tilde{d}_{x'',y''} \) and \( \beta > 0 \).

In summary, \( \{\hat{d}_{x,y} \}_{x,y \in [n]} \) is a feasible solution to MAX SQUARE SUM achieving a greater objective (20) than the optimal solution \( \{\tilde{d}_{x,y} \}_{x,y \in [n]} \) does, a contradiction.

We now bound the optimal value of MAX SQUARE SUM.

**Theorem A.3.** The optimal value of MAX SQUARE SUM is at most

\[
\left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \left( \left\lceil \frac{n^r}{\delta r} \right\rceil + 1 \right) \cdot (\delta nr)^2
\]

**Proof.** W.r.t. the optimal (and thus feasible) solution \( \{\tilde{d}_{x,y} \}_{x,y \in [n]} \), objective (20) equals

\[
\left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \sum_{x,y \in [n]} \chi \left[ \tilde{d}_{x,y} \neq 0 \right] \cdot \tilde{d}_{x,y}^2
\]

\[
\leq \left\lfloor \frac{n}{2} \right\rfloor \cdot \frac{1}{n \cdot (n-1)} \cdot \sum_{x,y \in [n]} \chi \left[ \tilde{d}_{x,y} > 0 \right] \cdot (\delta nr)^2,
\]

where \( \chi[P] = 1 \) if \( P \) is true and \( \chi[P] = 0 \) otherwise, for any predicate \( P \). Now invoke Lemma A.2. \( \square \)
References

[1] K. Barhum, O. Goldreich, and A. Shraibman. On approximating the average distance between points. In Proceedings of the 10th International Workshop on Approximation and the 11th International Workshop on Randomization, and Combinatorial Optimization, pages 296–310, 2007.

[2] C.-L. Chang. Deterministic sublinear-time approximations for metric 1-median selection. Information Processing Letters, 113(8):288–292, 2013.

[3] C.-L. Chang. A deterministic sublinear-time nonadaptive algorithm for metric 1-median selection. Theoretical Computer Science, 602:149–157, 2015.

[4] C.-L. Chang. Metric 1-median selection: Query complexity vs. approximation ratio. In Proceedings of the 22nd International Computing and Combinatorics Conference, pages 131–142, Ho Chi Minh City, Vietnam, 2016. Full version at https://arxiv.org/abs/1509.05662.

[5] C.-L. Chang. A lower bound for metric 1-median selection. Journal of Computer and System Sciences, 84:44–51, 2017.

[6] D. Eppstein and J. Wang. Fast approximation of centrality. Journal of Graph Algorithms and Applications, 8(1):39–45, 2004.

[7] O. Goldreich and D. Ron. Approximating average parameters of graphs. Random Structures & Algorithms, 32(4):473–493, 2008.

[8] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O’Callaghan. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering, 15(3):515–528, 2003.

[9] P. Indyk. Sublinear time algorithms for metric space problems. In Proceedings of the 31st Annual ACM Symposium on Theory of Computing, pages 428–434, 1999.

[10] P. Indyk. High-dimensional computational geometry. PhD thesis, Stanford University, 2000.

[11] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. Journal of the ACM, 57(2):5, 2010.

[12] R. R. Mettu and C. G. Plaxton. Optimal time bounds for approximate clustering. Machine Learning, 56(1–3):35–60, 2004.

[13] R. Motwani and P. Raghavan. Randomized Algorithms. Cambridge University Press, Cambridge, UK, 1995.
[14] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[15] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[16] B.-Y. Wu. On approximating metric 1-median in sublinear time. *Information Processing Letters*, 114(4):163–166, 2014.