Random Shortest Paths: Non-Euclidean Instances for Metric Optimization Problems

Karl Bringmann[1], Christian Engels[2], Bodo Manthey[3], B. V. Raghavendra Rao[4]

[1] Max Planck Institute for Informatics, kbringma@mpi-inf.mpg.de
[2] Saarland University, engels@cs.uni-saarland.de
[3] University of Twente, manthey@utwente.nl
[4] Indian Institute of Technology Madras, bvrr@cse.iitm.ac.in

Abstract. Probabilistic analysis for metric optimization problems has mostly been conducted on random Euclidean instances, but little is known about metric instances drawn from distributions other than the Euclidean.

This motivates our study of random metric instances for optimization problems obtained as follows: Every edge of a complete graph gets a weight drawn independently at random. The length of an edge is then the length of a shortest path (with respect to the weights drawn) that connects its two endpoints.

We prove structural properties of the random shortest path metrics generated in this way. Our main structural contribution is the construction of a good clustering. Then we apply these findings to analyze the approximation ratios of heuristics for matching, the traveling salesman problem (TSP), and the k-center problem, as well as the running-time of the 2-opt heuristic for the TSP. The bounds that we obtain are considerably better than the respective worst-case bounds. This suggests that random shortest path metrics are easy instances, similar to random Euclidean instances, albeit for completely different structural reasons.

1 Introduction

For large-scale optimization problems, finding optimal solutions within reasonable time is often impossible, because many such problems, like the traveling salesman problem (TSP), are NP-hard. Nevertheless, we often observe that simple heuristics succeed surprisingly quickly in finding close-to-optimal solutions. Many such heuristics perform well in practice but have a poor worst-case performance. In order to explain the performance of such heuristics, probabilistic analysis has proved to be a useful alternative to worst-case analysis. Probabilistic analysis of optimization problems has been conducted with respect to arbitrary instances (without the triangle inequality) [14, 22] or instances embedded in Euclidean space. In particular, the limiting behavior of various heuristics for many of the Euclidean optimization problems is known precisely [34].

However, the average-case performance of heuristics for general metric instances is not well understood. This lack of understanding can be explained by two reasons: First, independent random edge lengths (without the triangle inequality) and random geometric instances are relatively easy to handle from a technical point of view – the former because of the independence of the lengths, the latter because Euclidean space provides a structure that can be exploited. Second, analyzing heuristics on random metric spaces requires an understanding of random metric spaces in the first place. While Vershik [32] gave an analysis of a process for obtaining random metric spaces, using this directly to analyze algorithms seems difficult.

In order to initiate systematic research of heuristics on general metric spaces, we use the following model, proposed by Karp and Steele [23, Section 3.4]: Given an undirected complete graph, we draw edge weights independently at random. Then the length of an edge is the length of a shortest path connecting its endpoints. We call such instances *random shortest path metrics*.

This model is also known as *first-passage percolation*, and has been introduced by Broadbent and Hemmersley as a model for passage of fluid in a porous medium [6, 7]. More recently, it has also been used to model shortest paths in networks such as the internet [12]. The appealing feature of random shortest path metrics is their simplicity, which enables us to use them for the analysis of heuristics.

1.1 Known and Related Results

There has been significant study of random shortest path metrics or first-passage percolation. The expected length of an edge is known to be $\Theta(\log n/n)$ [9,20], and the same asymptotic bound holds also for the longest edge almost surely [17, 20]. This model has been used to analyze algorithms for computing shortest paths [15, 17, 27]. Kulkarni and Adlakha have developed algorithmic methods to compute distribution and moments of several optimization problems [24–26]. Beyond shortest path algorithms, random shortest path metrics have been applied only rarely to analyze algorithms. Dyer and Frieze, answering a question raised by Karp and Steele [23, Section 3.4], analyzed the patching heuristic for the asymmetric TSP (ATSP) in this model. They showed that it comes within a factor of $1 + o(1)$ of the optimal solution with high probability. Hassin and Zemel [17] applied their findings to the 1-center problem.

From a more structural point of view, first-passage percolation has been analyzed in the area of complex networks, where the hop-count (the number of edges on a shortest path) and the length of shortest path trees have been analyzed [19]. These properties have also been studied on random graphs with random edge weights [5,18]. More recently, Addario-Berry et. al. [1] showed that the number of edges in the longest of the shortest paths is $O(\log n)$ with high probability, and hence the shortest path trees have depth $O(\log n)$.

1.2 Our Results

As far as we are aware, simple heuristics such as greedy heuristics have not been studied in this model yet. Understanding the performance of such algorithms is particularly important as they are easy to implement and used in many applications.

We provide a probabilistic analysis of simple heuristics for optimization under random shortest path metrics. First, we provide structural properties of random shortest path metrics (Section 3). Our most important structural contribution is proving the existence of a good clustering (Lemma 3.8). Then we use these structural insights to analyze simple algorithms for minimum weight matching and the TSP to obtain better expected approximation ratios compared to the worst-case bounds. In particular, we show that the greedy algorithm for minimum-weight perfect matching (Theorem 4.2), the nearest-neighbor heuristic for the TSP (Theorem 4.3), and every insertion heuristic for the TSP (Theorem 4.4) achieve constant expected approximation ratios. We also analyze the 2-opt heuristic for the TSP and show that the expected number of 2-exchanges required before the termination of the algorithm is bounded by $O(n^8 \log^3 n)$ (Theorem 4.5). Investigating further the structural properties of random shortest path metrics, we then consider the k-center problem (Section 5), and show that the most trivial procedure of choosing k arbitrary vertices as k-centers yields a $1 + o(1)$ approximation in expectation, provided $k = O(n^{1-\varepsilon})$ for some $\varepsilon > 0$ (Theorem 5.2). Due to space limitations, most proofs are in the appendix.

2 Model and Notation

We consider undirected complete graphs $G = (V, E)$ without loops. First, we draw *edge weights* $w(e)$ independently at random according to the exponential distribution with parameter 1. (Exponential distributions are technically the easiest to handle because they are memoryless. However, our results hold also for other distributions, in particular for the uniform distribution on $[0, 1]$. We briefly discuss this in Section 6.) Second, let the *distances* or *lengths* $d : V \times V \to [0, \infty)$ be given by the lengths of the shortest paths between the vertices with respect to the weights thus drawn. In particular, we have $d(v, v) = 0$ for all $v \in V$, we have $d(u, v) = d(v, u)$ because G is undirected, and we have the triangle inequality: $d(u, v) \leq d(u, x) + d(x, v)$ for all $u, x, v \in V$. We call the complete graph with edge lengths d obtained from random weights w a *random shortest path metric*.

We use the following notation: Let $\Delta_{\max} = \max_{e \in E} d(e)$ denote the longest edge in the random shortest path metric. Let $N_\Delta^v = \{u \in V \mid d(u, v) \leq \Delta\}$ be the set of all nodes in a Δ-environment of v, and let $k_\Delta^v = |N_\Delta^v|$ the number of nodes around v in a Δ-environment. We denote the minimal Δ such that there are at least k nodes within a distance of Δ of v by Δ_k^v. Formally, we define $\Delta_k^v = \min\{\Delta \mid k_\Delta^v \geq k\}$. Note that $v \in N_\Delta^v$ for any $\Delta \geq 0$ because the distance of v to itself is 0. Consequently, we have $\Delta_1^v = 0$ and $k_0^v \geq 1$.

By $\mathrm{Exp}(\lambda)$, we denote the exponential distribution with parameter λ. By exp, we denote the exponential function. For $n \in \mathbb{N}$, let $[n] = \{1, \ldots, n\}$, and let $H_n = \sum_{i=1}^n 1/i$ be the n-th harmonic number.

3 Structural Properties of Shortest Path Metrics

3.1 Random Process

To understand random shortest path metrics, it is convenient to fix a starting vertex v and see how the lengths from v to the other vertices develop. In this way, we analyze the distribution of Δ_k^v.

The values Δ_k^v are generated by a simple birth process as follows. (The same process has been analyzed by Davis and Prieditis [9], Janson [20], and also in subsequent work.) For $k = 1$, we have $\Delta_k^v = 0$. For $k \geq 1$, we are looking for the closest vertex to any vertex in $N_{\Delta_k^v}^v$ in order to obtain Δ_{k+1}^v. This conditions all edges (u, x) with $u \in N_{\Delta_k^v}^v$ and $x \notin N_{\Delta_k^v}^v$ to be of length at least $\Delta_k^v - d(v, u)$. Otherwise, x would already be in $N_{\Delta_k^v}^v$. The set $N_{\Delta_k^v}^v$ contains k vertices. Thus, there are $k \cdot (n - k)$ connections from $N_{\Delta_k^v}^v$ to the rest of the graph. Consequently, the difference $\delta_k = \Delta_k^v - \Delta_{k-1}^v$ is distributed as the minimum of $k(n - k)$ exponential random variables (with parameter 1), or, equivalently, as an exponential random variable with parameter $k \cdot (n - k)$. We obtain that $\Delta_{k+1}^v = \sum_{i=1}^{k} \mathrm{Exp}\big(i \cdot (n - i)\big)$. (Note that the exponential distributions and the random variables $\delta_1, \ldots, \delta_n$ are independent.)

Exploiting linearity of expectation and that the expected value of $\mathrm{Exp}(\lambda)$ is $1/\lambda$ yields the following theorem.

Theorem 3.1. *For any $k \in [n]$ and any $v \in V$, we have $\mathbb{E}\big(\Delta_k^v\big) = \frac{1}{n} \cdot \big(H_{k-1} + H_{n-1} - H_{n-k}\big)$ and Δ_k^v is distributed as $\sum_{i=1}^{k-1} \mathrm{Exp}\big(i \cdot (n - i)\big)$.*

From this result, we can easily deduce two known results: averaging over k yields that the expected length of an edge is $\frac{H_{n-1}}{n-1} \approx \ln n/n$ [9,20]. By considering Δ_n^v, we obtain that the longest edge incident to a fixed vertex has an expected length of $2H_{n-1}/n \approx 2 \cdot \ln n/n$ [20]. For completeness, the length of the longest edge in the whole graph is roughly $3 \cdot \ln n/n$ [20].

3.2 Distribution of Δ_k^v

Let us now have a closer look at the distribution of Δ_k^v for fixed $v \in V$ and $k \in [n]$. Let F_k^v denote the cumulative distribution function (CDF) of Δ_k^v, i.e., $F_k^v(x) = \mathbb{P}(\Delta_k^v \leq x)$. A careful analysis of the distribution of a sum of exponential random variables yields the following two lemmas.

Lemma 3.2. *For every $\Delta \geq 0$, $v \in V$, and $k \in [n]$, we have*

$$\big(1 - \exp(-(n - k)\Delta)\big)^{k-1} \leq F_k^v(\Delta) \leq \big(1 - \exp(-n\Delta)\big)^{k-1}.$$

Proof. We have already seen that Δ_k^v is a sum of exponentially distributed random variables with parameters $\lambda_i = i(n - i) \in [(n - k)i, ni]$ for $i \in [k - 1]$. We approximate the parameters by ci for $c \in \{n - k, n\}$. The distribution with $c = n$ is stochastically dominated by the true distribution, which is in turn dominated by the distribution obtained for $c = n - k$.

We keep c as a parameter and obtain the following density function for the sum of exponentially distributed random variables with parameters $c, \dots, (k-1) \cdot c$ [30, p. 308ff]:

$$\sum_{i=1}^{k-1} \left(\prod_{j \in [k-1] \setminus \{i\}} \frac{j}{j-i} \right) \cdot ci \cdot \exp(-cix) = \sum_{i=1}^{k-1} \frac{\frac{(k-1)!}{i} \cdot (-1)^{i-1}}{(i-1)!(k-1-i)!} \cdot ci \cdot \exp(-cix)$$
$$= \sum_{i=1}^{k-1} \binom{k-1}{i} (-1)^{i-1} \cdot ci \cdot \exp(-cix).$$

Integrating plus the binomial theorem yields

$$\sum_{i=1}^{k-1} \binom{k-1}{i} \big(-\exp(-cix)\big)(-1)^{i-1} \cdot ci \cdot \exp(-cix) = \big(-\exp(-cx)+1\big)^{k-1} - 1.$$

Taking the difference of the function values at Δ and 0 yields $\big(1-\exp(-c\Delta)\big)^{k-1}$, which yields the bounds claimed by choosing $c = n-k$ and $c = n$. $\square$

Lemma 3.3. *Fix $\Delta \geq 0$ and a vertex $v \in V$. Then*

$$\big(1-\exp(-(n-k)\Delta)\big)^{k-1} \leq \mathbb{P}\big(k_\Delta^v \geq k\big) \leq \big(1-\exp(-n\Delta)\big)^{k-1}.$$

We can improve Lemma 3.2 slightly in order to obtain even closer lower and upper bounds. For $n, k \geq 2$, combining Lemmas 3.2 and 3.4 yields tight lower and upper bounds if we disregard the constants in the exponent, namely $F_k^v(\Delta) = \big(1-\exp(-\Theta(n\Delta))\big)^{\Theta(k)}$.

Lemma 3.4. *For all $v \in V$, $k \in [n]$, and $\Delta \geq 0$, we have $F_k^v(\Delta) \geq (1-\exp(-(n-1)\Delta/4))^{n-1}$ and $F_k^v(\Delta) \geq (1-\exp(-(n-1)\Delta/4))^{\frac{4}{3}(k-1)}$.*

3.3 Tail Bounds for k_Δ^v and $\Delta_{\max}$

Our first tail bound for k_Δ^v, which is the number of vertices within a distance of Δ of a given vertex v, follows directly from Lemma 3.2.

From this lemma we derive the following corollary, which is a crucial ingredient for the existence of good clusterings and, thus, for the analysis of the heuristics in the remainder of this paper.

Corollary 3.5. *Let $n \geq 5$ and fix $\Delta \geq 0$ and a vertex $v \in V$. Then we have*

$$\mathbb{P}\left(k_\Delta^v < \min\left\{\exp(\Delta n/5), \frac{n+1}{2}\right\}\right) \leq \exp(-\Delta n/5).$$

Corollary 3.5 is almost tight according to the following result.

Corollary 3.6. *Fix $\Delta \geq 0$, a vertex $v \in V$, and any $c > 1$. Then*

$$\mathbb{P}\big(k_\Delta^v \geq \exp(c\Delta n)\big) < \exp\big(-(c-1)\Delta n\big).$$

Janson [20] derived the following tail bound for the length $\Delta_{\max}$ of the longest edge. A qualitatively similar bound can be proved using Lemma 3.3 and can also be derived from Hassin and Zemel's analysis [17]. However, Janson's bound is stronger with respect to the constants in the exponent.

Lemma 3.7 (Janson [20, p. 352]). *For any fixed $c > 3$, we have $\mathbb{P}(\Delta_{\max} > c\ln(n)/n) \leq O(n^{3-c}\log^2 n)$.*

3.4 Stars and Clusters

In this section, we show our main structural contribution, which is a more global property of random shortest path metrics. We show that such instances can be divided into a small number of clusters of any given diameter.

From now on, let $\#(n, \Delta) = \min\{\exp(\Delta n/5), (n+1)/2\}$, as in Corollary 3.5. If the number k_Δ^v of vertices within a distance of Δ of v is at least $\#(n, \Delta)$, then we call the vertex v a *Δ-center*, and we call the set N_Δ^v of vertices within a distance of at most Δ of v (including v itself) the *Δ-star of v*. Otherwise, if $k_\Delta^v < \#(n, \Delta)$, we call the vertex v a *sparse Δ-center*. Any two vertices in the same Δ-star have a distance of at most 2Δ because of the triangle inequality. If Δ is clear from the context, then we also speak about centers and stars without parameter. We can bound, by Corollary 3.5, the expected number of sparse Δ-centers to be at most $O(n/\#(n, \Delta))$.

We want to partition the graph into a small number of clusters, each of diameter at most 6Δ. For this purpose, we put each sparse Δ-center in its own cluster (of size 1). Then the diameter of each such cluster is $0 \leq 6\Delta$ and the number of these clusters is expected to be at most $O(n/\#(n, \Delta))$.

We are left with the Δ-centers, which we cluster using the following algorithm: Consider an auxiliary graph whose vertices are all Δ-centers. We draw an edge between two Δ-centers u and v if $N_\Delta^u \cap N_\Delta^v \neq \emptyset$. Now consider any maximal independent set of this auxiliary graph (for instance, a greedy independent set), and let t be the number of its vertices. Then we form initial clusters $C_1', \ldots, C_t'$, each containing one of the Δ-stars corresponding to the vertices in the independent set. By the independence, all t Δ-stars are disjunct, which implies $t \leq n/\#(n, \Delta)$. The star of every remaining center v has at least one vertex (maybe v itself) in one of the C_i'. We add all remaining vertices of N_Δ^v to such a C_i' to form the final clusters $C_1, \ldots, C_t$. Now, the maximum distance within each C_i is at most 6Δ: Consider any two vertices $u, v \in C_i$. The distance of u towards its closest neighbor in the initial star C_i' is at most 2Δ. The same holds for v. Finally, the diameter of the initial star C_i' is also at most 2Δ.

With this partitioning, we have obtained the following structure: We have an expected number of $O(n/\#(n, \Delta))$ clusters of size 1 and diameter 0, and a number of $O(n/\#(n, \Delta))$ clusters, each of size at least $\#(n, \Delta)$ and diameter at most 6Δ. Thus, we have $O(n/\#(n, \Delta)) = O(1 + n/\exp(\Delta n/5))$ clusters in total. We summarize these findings in the following lemma. It will be the crucial ingredient for bounding the expected approximation ratios of the greedy, nearest-neighbor, and insertion heuristics.

Lemma 3.8. *Consider a random shortest path metric and let $\Delta \geq 0$. If we partition the instance into clusters, each of diameter at most 6Δ, then the expected number of clusters needed is $O(1 + n/\exp(\Delta n/5))$.*

4 Analysis of Heuristics

In order to bound approximation ratios, we will exploit a simple upper bound on the probability that an optimal TSP tour or matching has a length of at most c for some small constant c (Lemma B.1). Note that the expected lengths of the minimum-length perfect matching and the optimal TSP are $\Theta(1)$ even without taking shortest paths [14, 33]. Thus, both the optimal TSP and the optimal matching have an expected length of $O(1)$ for random shortest path metrics.

4.1 Greedy Heuristic for Minimum-Length Perfect Matching

Finding minimum-length perfect matchings in metric instances is the first problem that we consider. This problem has been widely considered in the past and has applications in, e.g., optimizing the speed of mechanical plotters [28,31]. The worst-case running-time of $O(n^3)$ for finding an optimal matching is prohibitive if the number n of points is large. Thus, simple heuristics are often used, with the greedy heuristic being probably the simplest one: at every step, choose an edge of minimum length incident to the unmatched vertices and add it to the partial matching. Let GREEDY denote the cost of the matching output by this greedy matching heuristic, and let MM denote the optimum value of the minimum weight matching. The worst-case approximation ratio for greedy matching on metric instances is $\Theta(n^{\log_2(3/2)})$ [28], where $\log_2(3/2) \approx 0.58$. In the case of Euclidean instances, the greedy algorithm has an approximation ratio of $O(1)$ with high probability on random instances [3]. For independent random edge weights (without the triangle inequality), the expected weight of the matching computed by the greedy algorithm is $\Theta(\log n)$ [10] whereas the optimal matching has a weight of $\Theta(1)$ with high probability, which gives an $O(\log n)$ approximation ratio.

We show that greedy matching finds a matching of constant expected length on random shortest path metrics. The proof is similar to the ones of Theorems 4.3 and 4.4, and we include it as an example.

Theorem 4.1. $\mathbb{E}[\mathsf{GREEDY}] = O(1)$.

Proof. Set $\Delta_i = i/n$ for $i \in \{0, 1, \ldots, \log n\}$. We divide the run of GREEDY in phases as follows: We say that GREEDY is in phase i if the lengths of the edges it inserts are in the interval $(6\Delta_{i-1}, 6\Delta_i]$. Lemma 3.7 allows to show that the expected sum of all lengths of edges longer than $6\Delta_{O(\log n)}$ is $o(1)$, so we can ignore them.

Since the lengths of the edges that GREEDY adds increases monotonically, GREEDY goes through phases i with increasing i (while a phase can be empty). We now estimate the contribution of phase i to the matching computed by

GREEDY. Using Lemma 3.8, after phase $i-1$, we can find a clustering into clusters of diameter at most $6\Delta_{i-1}$ using an expected number of $O(1+n/e^{(i-1)/5})$ clusters. Each such cluster can have at most one unmatched vertex. Thus, we have to add at most $O(1+n/e^{(i-1)/5})$ edges in phase i, each of length at most $6\Delta_i$. Thus, the contribution of phase i is $O(\Delta_i(1+n/e^{(i-1)/5}))$ in expectation. Summing over all phases yields the desired bound:

$$\mathbb{E}[\mathsf{GREEDY}] = o(1) + \sum_{i=1}^{\log n} O\big(\frac{i}{e^{(i-1)/5}} + \frac{i}{n}\big) = O(1).$$

□

Careful analysis allows us to even bound the expected approximation ratio.

Theorem 4.2. *The greedy algorithm for minimum-length matching has constant approximation ratio on random shortest path metrics, i.e.,*

$$\mathbb{E}\left[\frac{\mathsf{GREEDY}}{\mathsf{MM}}\right] \in O(1).$$

4.2 Nearest-Neighbor algorithm for the TSP

A greedy analogue for the traveling salesman problem (TSP) is the *nearest neighbor* heuristic: Start with a vertex v as the current vertex, and at every iteration choose the nearest yet unvisited neighbor u of the current vertex as the next vertex in the tour and move to the next iteration with the new vertex u as the current vertex. Let NN denote both the nearest-neighbor heuristic itself and the cost of the tour computed by it. Let TSP denote the cost of an optimal tour. The nearest-neighbor heuristic NN achieves a worst-case ratio of $O(\log n)$ for metric instances and also an average-case ratio (for independent, non-metric edge lengths) of $O(\log n)$ [2]. We show that NN achieves a constant approximation ratio on random shortest path instances. The proof is similar to the ones of Theorems 4.1 and 4.2.

Theorem 4.3. $\mathbb{E}[\mathsf{NN}] = O(1)$ *and* $\mathbb{E}\left[\frac{\mathsf{NN}}{\mathsf{TSP}}\right] \in O(1)$.

4.3 Insertion Heuristics

An insertion heuristic for the TSP is an algorithm that starts with an initial tour on a few vertices and extends this tour iteratively by adding the remaining vertices. In every iteration, a vertex is chosen according to some rule, and this vertex is inserted at the place in the current tour where it increases the total tour length the least. Certain insertion heuristics such as nearest neighbor insertion (which is different from the nearest neighbor algorithm from the previous section) are known to achieve constant approximation ratios [29]. The random insertion algorithm, where the next vertex is chosen uniformly at random from the remaining vertices, has a worst-case approximation ratio of

$\Omega(\log\log n/\log\log\log n)$, and there are insertion heuristics with a worst-case approximation ratio of $\Omega(\log n/\log\log n)$ [4].

For random shortest path metrics, we show that any insertion heuristic produces a tour whose length is expected to be within a constant factor of the optimal tour. This result holds irrespective of which insertion strategy we actually use. It holds even in the (admittedly a bit unrealistic) scenario, where an adversary specifies the order in which the vertices have to be inserted after the random instance is drawn.

Theorem 4.4. *The expected cost of the TSP tour obtained with any insertion heuristics is bounded from above by $O(1)$. This holds even against an adaptive adversary, i.e., if an adversary chooses the order in which the vertices are inserted after the edge weights are drawn.*

Furthermore, the expected approximation ratio of any insertion heuristic is also $O(1)$.

4.4 Running-Time of 2-Opt for the TSP

The 2-opt heuristic for the TSP starts with an initial tour and successively improves the tour by so-called 2-exchanges until no further refinement is possible. In a 2-exchange, a pair of edges $e_1 = \{u, v\}$ and $e_2 = \{x, y\}$ are replaced by a pair of edges $f_1 = \{u, y\}$ and $f_2 = \{x, v\}$ to get a shorter tour. The 2-opt heuristic is easy to implement and widely used. In practice, it usually converges quite quickly to close-to-optimal solutions [21]. However, its worst-case running-time is exponential [13]. To explain 2-opt's performance on geometric instances, Englert et al. [13] have proved that the number of iterations that 2-opt needs is bounded by a polynomial in a smoothed input model for geometric instances. Also for random shortest path metrics, the expected number of iterations that 2-opt needs is bounded by a polynomial. The proof is similar to Englert et al.'s analysis [13].

Theorem 4.5. *The expected number of iterations that 2-opt needs to find a local optimum is bounded by $O(n^8 \log^3 n)$.*

5 k-Center

In the (metric) k-center problem, we are given a finite metric space (V, d) and should pick k points $U \subseteq V$ such that $\sum_{v \in V} \min_{u \in U} d(v, u)$ is minimized. We call the set U a k-center. Gonzalez [16] gave a simple 2-approximation for this problem and showed that finding a $(2 - \varepsilon)$-approximation is NP-hard.

In this section, we consider the k-center problem in the setting of random shortest path metrics. In particular we examine the approximation ratio of the algorithm $\mathsf{TRIVIAL}$, which picks k points independent of the metric space, e.g., $U = \{1, \ldots, k\}$, or k random points in V. We show that $\mathsf{TRIVIAL}$ yields a $(1 + o(1))$-approximation for $k = O(n^{1-\varepsilon})$. This can be seen as an algorithmic result since it improves upon the worst-case approximation factor of 2, but it is essentially a structural result on random shortest path metrics. It means that any set of k points is, with high probability, a very good k-center, which gives some knowledge about the topology of random shortest path metrics. For larger, but not too large k, i.e., $k \leq (1-\varepsilon)n$, TRIVIAL still yields an $O(1)$-approximation.

The main insight comes from generalizing the growth process described in Section ?? Fixing $U = \{v_1, \ldots, v_k\} \subseteq V$ we sort the vertices $V \setminus U$ by their distance to U in ascending order, calling the resulting order $v_{k+1}, \ldots, v_n$. Now we consider $\delta_i = d(v_{i+1}, U) - d(v_i, U)$ for $k \leq i < n$. These random variables are generated by an easy growth process analogously to Section ??, which shows that the δ_i are independent and $\delta_i \sim \mathrm{Exp}(i(n-i))$. Since the cost of U as a k-center can be expressed using the δ_i's and since $a\,\mathrm{Exp}(1) \sim \mathrm{Exp}(1/a)$, we have $\mathsf{cost}(U) = \sum_{i=k}^{n-1} (n-i) \cdot \delta_i \sim \sum_{i=k}^{n-1} (n-i) \cdot \mathrm{Exp}(i(n-i)) \sim \sum_{i=k}^{n-1} \mathrm{Exp}(i)$. From this, we can read off the expected cost of U immediately, and thus the expected cost of TRIVIAL.

Lemma 5.1. *Fix $U \subseteq V$ of size k. We have $\mathbb{E}[\mathsf{TRIVIAL}] = \mathbb{E}[\mathsf{cost}(U)] = H_{n-1} - H_{k-1} = \ln(n/k) + \Theta(1)$.*

By closely examining the random variable $\sum_{i=k}^{n-1} \mathrm{Exp}(i)$, we can show good tail bounds for the probability that the cost of U is lower than expected. Together with the union bound this yields tail bounds for the optimal k-center CENTER, which implies the following theorem. In this theorem, the approximation ratio becomes $1 + O\big(\frac{\ln\ln(n)}{\ln(n)}\big)$ for $k = O(n^{1-\varepsilon})$.

Theorem 5.2. *Let $k \leq (1-\varepsilon)n$ for some constant $\varepsilon > 0$. Then $\mathbb{E}\left[\frac{\mathsf{TRIVIAL}}{\mathsf{CENTER}}\right] = O(1)$. If we even have $k \leq cn$ for some sufficiently small constant $c \in (0,1)$, then $\mathbb{E}\left[\frac{\mathsf{TRIVIAL}}{\mathsf{CENTER}}\right] = 1 + O\big(\frac{\ln\ln(n/k)}{\ln(n/k)}\big)$.*

6 Remarks and Open Problems

The results of this paper carry over to the case of edge weights drawn according to the uniform distribution on the interval $[0,1]$. The analysis remains basically identical, but gets technically a bit more difficult because we lose the memorylessness of the exponential distribution. The intuition is that, because the longest edge has a length of $O(\log n/n) = o(1)$, only the behavior of the distribution in a small, shrinking interval $[0, o(1)]$ is relevant. Essentially, if the probability that an edge weight is smaller than t is $t + o(t)$, then our results carry over. We refer to Janson's coupling argument [20] for more details.

To conclude the paper, let us list the open problems that we consider most interesting:

1. While the distribution of edge lengths in asymmetric instances does not differ much from the symmetric case, an obstacle in the application of asymmetric random shortest path metrics seems to be the lack of clusters of small diameter (see Section 3). Is there an asymmetric counterpart for this?

2. Is it possible to prove even an $1 + o(1)$ (like Dyer and Frieze [11] for the patching algorithm) approximation ratio for any of the simple heuristics that we analyzed?
3. What is the approximation ratio of 2-opt in random shortest path metrics? In the worst case, it is $O(\sqrt{n})$ [8]. For edge lengths drawn uniformly at random from the interval $[0, 1]$ without taking shortest paths, the expected approximation ratio is $O(\sqrt{n} \cdot \log^{3/2} n)$ [**?**]. For d-dimensional geometric instances, the smoothed approximation ratio is $O(\phi^{1/d})$ [13], where ϕ is the perturbation parameter.

   We easily get an approximation ratio of $O(\log n)$ based on the two facts that the length of the optimal tour is $\Theta(1)$ with high probability and that $\Delta_{\max} = O(\log n/n)$ with high probability. Can we prove that the expected ratio of 2-opt is $o(\log n)$?

References

1. Louigi Addario-Berry, Nicolas Broutin, and Gábor Lugosi. The longest minimum-weight path in a complete graph. *Combin. Probab. Comput.*, 19(1):1–19, 2010.
2. Giorgio Ausiello, Pierluigi Crescenzi, Giorgio Gambosi, Viggo Kann, Alberto Marchetti-Spaccamela, and Marco Protasi. *Complexity and Approximation*. Springer, 1999.
3. David Avis, Burgess Davis, and J. Michael Steele. Probabilistic analysis of a greedy heuristic for Euclidean matching. *Probab. Engrg. Inform. Sci.*, 2:143–156, 1988.
4. Yossi Azar. Lower bounds for insertion methods for TSP. *Combin. Probab. Comput.*, 3:285–292, 1994.
5. Shankar Bhamidi, Remco van der Hofstad, and Gerard Hooghiemstra. First passage percolation on the Erdős-Rényi random graph. *Combin. Probab. Comput.*, 20(5):683–707, 2011.
6. Nathaniel D. Blair-Stahn. First passage percolation and competition models. arXiv:1005.0649v1 [math.PR], 2010.
7. S. R. Broadbent and J.M. Hemmersley. Percolation processes. I. Crystals and mazes. *Proceedings of the Cambridge Philosophical Society*, 53(3):629–641, 1957.
8. Barun Chandra, Howard J. Karloff, and Craig A. Tovey. New results on the old k-opt algorithm for the traveling salesman problem. *SIAM J. Comput.*, 28(6):1998–2029, 1999.
9. Robert Davis and Armand Prieditis. The expected length of a shortest path. *Inform. Process. Lett.*, 46(3):135–141, 1993.
10. Martin Dyer, Alan Frieze, and Boris Pittel. The average performance of the greedy matching algorithm. *Ann. Appl. Probab.*, 3(2):526–552, 1993.
11. Martin E. Dyer and Alan M. Frieze. On patching algorithms for random asymmetric travelling salesman problems. *Math. Program.*, 46:361–378, 1990.
12. Maren Eckhoff, Jesse Goodman, Remco van der Hofstad, and Francesca R. Nardi. Short paths for first passage percolation on complete graphs. arXiv:1211.4569v1 [math.PR], 2012.
13. C. Engels and B. Manthey. Average-case approximation ratio of the 2-opt algorithm for the TSP. *Oper. Res. Lett.*, 37(2):83–84, 2009.
14. Matthias Englert, Heiko Röglin, and Berthold Vöcking. Worst case and probabilistic analysis of the 2-opt algorithm for the TSP. In *Proc. of the 18th Ann. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pp. 1295–1304. SIAM, 2007.

15. Alan M. Frieze. On random symmetric travelling salesman problems. *Math. Oper. Res.*, 29(4):878–890, 2004.
16. Alan M. Frieze and G. R. Grimmett. The shortest-path problem for graphs with random arc-lengths. *Discrete Appl. Math.*, 10:57–77, 1985.
17. Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoret. Comput. Sci.*, 38:293–306, 1985.
18. Refael Hassin and Eitan Zemel. On shortest paths in graphs with random weights. *Math. Oper. Res.*, 10(4):557–564, 1985.
19. Remco van der Hofstad, Gerard Hooghiemstra, and Piet van Mieghem. First passage percolation on the random graph. *Probab. Engrg. Inform. Sci.*, 15(2):225–237, 2001.
20. Remco van der Hofstad, Gerard Hooghiemstra, and Piet van Mieghem. Size and weight of shortest path trees with exponential link weights. *Combin. Probab. Comput.*, 15(6):903–926, 2006.
21. Svante Janson. One, two, three times $\log n/n$ for paths in a complete graph with edge weights. *Combin. Probab. Comput.*, 8(4):347–361, 1999.
22. David S. Johnson and Lyle A. McGeoch. Experimental analysis of heuristics for the STSP. In Gregory Gutin and Abraham P. Punnen, eds., *The Traveling Salesman Problem and its Variations*. Kluwer, 2002.
23. Richard M. Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Math. Oper. Res.*, 2(3):209–224, 1977.
24. Richard M. Karp and J. Michael Steele. Probabilistic analysis of heuristics. In E. L. Lawler et al., eds., *The Traveling Salesman Problem*, pp. 181–205. Wiley, 1985.
25. V. G. Kulkarni and V. G. Adlakha. Maximum flow in planar networks in exponentially distributed arc capacities. *Comm. Statist. Stochastic Models*, 1(3):263–289, 1985.
26. Vidyadhar G. Kulkarni. Shortest paths in networks with exponentially distributed arc lengths. *Networks*, 16(3):255–274, 1986.
27. Vidyadhar G. Kulkarni. Minimal spanning trees in undirected networks with exponentially distributed arc weights. *Networks*, 18(2):111–124, 1988.
28. Yuval Peres, Dmitry Sotnikov, Benny Sudakov, and Uri Zwick. All-pairs shortest paths in $o(n^2)$ time with high probability. In *Proc. of the 51st Ann. IEEE Symp. on Foundations of Computer Science (FOCS)*, pp. 663–672. IEEE, 2010.
29. Edward M. Reingold and Robert Endre Tarjan. On a greedy heuristic for complete matching. *SIAM J. Comput.*, 10(4):676–681, 1981.
30. Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis II. An analysis of several heuristics for the traveling salesman problem. *SIAM J. Comput.*, 6(3):563–581, 1977.
31. Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 2010.
32. Kenneth J. Supowit, David A. Plaisted, and Edward M. Reingold. Heuristics for weighted perfect matching. In *Proc. of the 12th Ann. ACM Symp. on Theory of Computing (STOC)*, pp. 398–419. ACM, 1980.
33. A. M. Vershik. Random metric spaces and universality. *Russian Math. Surveys*, 59(2):259–295, 2004.
34. David W. Walkup. On the expected value of a random assignment problem. *SIAM J. Comput.*, 8(3):440–442, 1979.
35. Joseph E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Springer, 1998.

A Proofs of Section 3

Theorem 3.1. For any $k \in [n]$ and any $v \in V$,

$$\mathbb{E}\big(\Delta_k^v\big) = \frac{1}{n} \cdot \big(H_{k-1} + H_{n-1} - H_{n-k}\big).$$

Proof. The proof is by induction on k. For $k = 1$, we have $\Delta_k^v = 0$ and $H_{k-1} + H_{n-1} - H_{n-k} = H_0 + H_{n-1} - H_{n-1} = 0$. Now assume that the lemma holds for k for some $k \geq 1$.

The additional distance $\Delta_{k+1}^v - \Delta_k^v$ is an exponentially distributed random variable with parameter $k \cdot (n-k)$, because it is the minimum of $k \cdot (n-k)$ exponentially distributed random variables. Thus, its expected value is $\frac{1}{k \cdot (n-k)}$. (Here, it is crucial that we use exponential distributions. Knowing Δ_k^v and the k vertices in $N_{\Delta_k^v}^v$ puts a restriction on the edge from vertices in $N_{\Delta_k^v}^v$ to vertices outside. However, because of the "memorylessness" of the exponential distribution, conditioning on a certain minimum weight still leaves us with an exponentially distributed random variable.)

Plugging in the induction hypothesis yields

$$\begin{aligned}
\mathbb{E}\big(\Delta_{k+1}^v\big) = \mathbb{E}\big(\Delta_k^v\big) + \frac{1}{k \cdot (n-k)} &= \frac{1}{n} \cdot \big(H_{k-1} + H_{n-1} - H_{n-k}\big) + \frac{1}{k \cdot (n-k)} \\
&= \frac{1}{n} \cdot \big(H_{k-1} + H_{n-1} - H_{n-k}\big) + \frac{1}{nk} + \frac{1}{n \cdot (n-k)} \\
&= \frac{1}{n} \cdot \big(H_k + H_{n-1} - H_{n-(k+1)}\big).
\end{aligned}$$

□

Lemma 3.4. For all $v \in V$, $k \in [n]$, and $\Delta \geq 0$, we have

$$F_k^v(\Delta) \geq (1 - \exp(-(n-1)\Delta/4))^{n-1}$$

and

$$F_k^v(\Delta) \geq (1 - \exp(-(n-1)\Delta/4))^{\frac{4}{3}(k-1)}.$$

Proof. As Δ_k^v is monotonically increasing in k, we have $F_k^v(\Delta) \geq F_{k+1}^v(\Delta)$ for all k. Thus, we have to prove the claim only for $k = n$. In this case, $\Delta_n^v = \sum_{i=1}^{n-1} \mathrm{Exp}(\lambda_i)$, with $\lambda_i = i(n-i) = \lambda_{n-i}$. Setting $m = \lceil n/2 \rceil$ and exploiting the symmetry around m yields

$$\Delta_n^v \preceq \sum_{i=1}^{m} \mathrm{Exp}(\lambda_i) + \sum_{i=1}^{m} \mathrm{Exp}(\lambda_i) = \Delta_m^v + \Delta_m^v.$$

Here, "$\preceq$" means stochastic dominance, "$=$" means equal distribution, and "$+$" means adding up two independent random variables. Hence,

$$F_n^v(\Delta) = \mathbb{P}[\Delta_n^v \leq \Delta] \geq \mathbb{P}[\Delta_m^v + \Delta_m^v \leq \Delta] \geq \mathbb{P}[\Delta_m^v \leq \Delta/2]^2.$$

By Lemma 3.2, and using $m \le (n+1)/2$, this is bounded by

$$F_n^v(\Delta) \ge (1 - \exp(-(n-m)\Delta/2))^{2(m-1)} \ge (1 - \exp(-(n-1)\Delta/4))^{n-1}.$$

For the second inequality, we use the first inequality of Lemma 3.4 for $k-1 \ge \frac{3}{4}(n-1)$ and Lemma 3.2 for $k-1 < \frac{3}{4}(n-1)$ as then $n-k \ge (n-1)/4$. $\square$

Lemma 3.3. Fix $\Delta \ge 0$ and a vertex $v \in V$. Then

$$\bigl(1 - \exp(-(n-k)\Delta)\bigr)^{k-1} \le \mathbb{P}\bigl(k_\Delta^v \ge k\bigr) \le \bigl(1 - \exp(-n\Delta)\bigr)^{k-1}.$$

Proof. We have $k_\Delta^v \ge k$ if and only if $\Delta_k^v \le \Delta$. Thus, the probability that $k_\Delta^v \ge k$ is sandwiched between $(1-\exp(-(n-k)\Delta))^{k-1}$ and $(1-\exp(-n\Delta))^{k-1}$ by Lemma 3.2. $\square$

Corollary 3.5. Let $n \ge 5$ and fix $\Delta \ge 0$ and a vertex $v \in V$. Then

$$\mathbb{P}\left(k_\Delta^v < \min\left\{\exp\left(\Delta n/5\right), \tfrac{n+1}{2}\right\}\right) \le \exp\left(-\Delta n/5\right).$$

Proof. Lemma 3.3 yields

$$\begin{aligned}\mathbb{P}\left(k_\Delta^v < \min\left\{\exp\left(\Delta \tfrac{n-1}{4}\right), \tfrac{n+1}{2}\right\}\right) &\le 1 - \left(1 - \exp\left(-\tfrac{n-1}{2}\Delta\right)\right)^{\exp(\Delta(n-1)/4)} \\ &\le \exp\left(-\Delta \tfrac{n-1}{4}\right),\end{aligned}$$

where the last inequality follows from $(1-x)^y \ge 1 - xy$ for $y \ge 1$, $x \ge 0$. Using $(n-1)/4 \ge n/5$ for $n \ge 5$ yields the simplified statement of the lemma (this is not really tight for larger n, but it does not change the asymptotics of our results). $\square$

Corollary 3.6. Fix $\Delta \ge 0$, a vertex $v \in V$, and any $c > 1$. Then

$$\mathbb{P}\bigl(k_\Delta^v \ge \exp(c\Delta n)\bigr) < \exp\bigl(-(c-1)\Delta n\bigr).$$

Proof. Lemma 3.3 with $k = c\Delta n$ yields

$$\mathbb{P}\bigl(k_\Delta^v \ge \exp(c\Delta n)\bigr) \le \bigl(1 - \exp(-n\Delta)\bigr)^{\exp(c\Delta n)-1}.$$

Using $1 + x \le e^x$ we get

$$\mathbb{P}\bigl(k_\Delta^v \ge \exp(c\Delta n)\bigr) \le \exp\left(\exp(-n\Delta) - \exp\bigl((c-1)\cdot \Delta n\bigr)\right).$$

Now, we bound $\exp(-n\Delta) \le 1$ and $\exp\bigl((c-1)\cdot\Delta n\bigr) \ge 1 + (c-1)\cdot \Delta n$, which yields the inequality claimed. $\square$

B Proofs of Section 4

We will use the following tail bound to estimate the approximation ratios of the greedy heuristic for matching as well as the nearest-neighbor and insertion heuristics for the TSP.

Lemma B.1. *Let S be the sum of the lightest $n/2$ edge weights drawn according to independent exponential distributions with parameter* 1 *(no shortest paths). Then $\mathbb{P}(S \leq c) \leq (2e^2c)^{n/2}$ for all $c \leq 1$.*

Furthermore, $\mathsf{TSP} \geq \mathsf{MM} \geq S$, where TSP and MM denote the length of the shortest TSP tour and the minimum-weight perfect matching, respectively, in the corresponding shortest path metric.

Proof. Let $k = n/2$. If there are q possible choice for k edges, then we have $\mathbb{P}(S \leq c) \leq q \cdot c^k/k!$ by a union bound over the q possible choice and an induction over k. We have $q \leq (n^2/2)^k/k!$.

Now we use $k^k \geq k! \geq (k/e)^k$. This yields an upper bound of

$$\begin{aligned}\mathbb{P}(S \leq c) &\leq \frac{n^{2k}e^k}{2^k k^k} \cdot \frac{c^k e^k}{k^k} = \frac{n^{2k}(e^2c/2)^k}{k^{2k}} = \frac{n^n(e^2c/2)^{n/2}}{(n/2)^n} \\ &= (e^2c/2)^{n/2}4^{n/2} = (2e^2c)^{n/2}.\end{aligned}$$

What remains to be proved is $\mathsf{TSP} \geq \mathsf{MM} \geq S$. The first inequality is trivial. For the second inequality, consider a minimum-weight perfect matching in a random shortest path metric. We replace every edge by the corresponding paths. If we disregard multiple edges, then we are still left with at least $n/2$ edges whose length is not shortened by taking shortest paths. The sum of the weights of these $n/2$ edges is at most MM and at least S. □

Theorem 4.2. The greedy algorithm for minimum weight matching has constant approximation ratio under the random-weights shortest path model, i.e., $\mathbb{E}\left[\frac{\mathsf{GREEDY}}{\mathsf{MM}}\right] \in O(1)$.

Proof. The worst-case approximation ratio of GREEDY for minimum-weight perfect matching is $n^{\log_2(3/2)}$ [28]. Let $c > 0$ be some constant to be specified later on. Then the approximation ratio of GREEDY on random shortest path instances is $\mathbb{E}\left[\frac{\mathsf{GREEDY}}{\mathsf{MM}}\right] \leq \mathbb{E}\left[\frac{\mathsf{GREEDY}}{c}\right] + \mathbb{P}(\mathsf{MM} < c) \cdot n^{\log_2(3/2)}$. By Theorem 4.1, the first term is $O(1)$. Choosing $c \leq 1/23$ and applying Lemma B.1 shows that the second term is $o(1)$. □

Theorem 4.3. $\mathbb{E}[\mathsf{NN}] = O(1)$.

Proof. The proof is similar to the proof of Theorem 4.2. Let $a_i = c \cdot \exp(i)$ for $i \in \mathbb{N}$, and let $\Delta_i = a_i/n$. Let $Q = O(\log n/n)$ be sufficiently large.

Let $C_{i,1}, \ldots, C_{i,\ell_i}$ denote the clusters obtained with parameter Δ_i as in the discussion preceding Lemma 3.8. We refer to these clusters as the *i-clusters.* Let v be any vertex. We call v *bad at i*, if v is in some i-cluster and NN chooses an edge of weight larger than $6\Delta_i$ for leaving v. Hence, if v is bad at i, then the next vertex lies outside of the cluster to which v belongs. (Note that v is not bad at i if the outgoing edge at v leads to a neighbor outside of the cluster of v, but has a length of less than $6\Delta_i$.)

In the following, let the costs of a vertex v be the length of the outgoing edge chosen with v as the current vertex. Thus, the length of the tour produced by NN equals the sum of costs over all vertices.

Claim. The expected number of vertices with costs in the range $(6\Delta_{i-1}, 6\Delta_i]$ is at most $O(n/\exp(a_{i-1}/5))$.

Proof (of Claim). Suppose that the costs of the neighbor chosen by NN for a vertex v is in the interval $(6\Delta_{i-1}, 6\Delta_i]$. Then v is either bad at $i-1$ or a single vertex cluster with respect to Δ_{i-1}. The event that v is bad at $i-1$ happens for at most one vertex in a Δ_{i-1} cluster. By Lemma 3.8, the number of Δ_{i-1} clusters is at most $O(n/\exp(a_{i-1}/5))$. □

If $\Delta_{\max} \le Q$, then it suffices to consider i for $i \le O(\log n)$. If $\Delta_{\max} > Q$, then we bound the value of the tour produced by NN by $n\Delta_{\max}$. This failure event, however, contributes only $o(1)$ to the expected value by Lemma 3.7. For the case $\Delta_{\max} \le Q$, the contribution to the expected length of the NN tour is bounded from above by

$$\sum_{i=1}^{O(\log n)} 6\Delta_i \cdot O\left(\frac{n}{\exp(a_{i-1}/5)}\right) = O(1).$$

The proof of the constant expected approximation ratio is similar to the proof of Theorem 4.2. □

Theorem 4.4. The expected cost of the TSP tour obtained with any insertion heuristics is bounded from above by $O(1)$. This holds even against an adaptive adversary, i.e., if an adversary chooses the order in which the vertices are inserted after the edge weights are drawn.

Proof. Let $a_1, \ldots, a_\ell$, and $\Delta_1, \ldots, \Delta_\ell$ be chosen as in Theorem 4.1, where $\ell = O(\log n)$. Choose $Q = O(\log n/n)$ sufficiently large and assume that $\Delta_{\max} \le Q$. If $\Delta_{\max} > Q$, then we bound the length of the tour produced by $n \cdot \Delta_{\max}$. This contributes only $o(1)$ to the expected value of length of the tour produced by Lemma 3.7.

Suppose we have a partial tour τ and v is the vertex that we have to insert next. If τ has a vertex u such that v and u are in a common Δ_i-cluster, then the triangle inequality implies that the costs of inserting v into τ is at most $12\Delta_i$. (The diameter of a cluster is at most $6\Delta_i$. Thus, going from u to v and back costs at most $12\Delta_i$. We add this to τ and take a shortcut to obtain a TSP tour.) For each i, there is at most one vertex for each Δ_i-cluster whose insertion could possibly cost more than $12\Delta_i$ (namely, the first vertex of that cluster). Thus, the number of vertices whose insertion would incur a cost in the range $(12\Delta_{i-1}, 12\Delta_i]$ is at most $O(\frac{n}{\exp(a_{i-1}/5)})$ in expectation, since all of them have cost more than $12\Delta_{i-1}$. Now, summing up the expected costs for all i, we obtain that the costs of the tour obtained by an insertion heuristic is bounded from above by

$$\begin{aligned}\sum_{i=1}^{\ell} O\left(\frac{n}{\exp(a_{i-1}/5)} \cdot \frac{a_i}{n}\right) &= \sum_{i=1}^{\ell} O\left(\frac{n}{\exp(a_{i-1}/5)} \cdot \frac{a_{i-1}/5}{n}\right) \\ &\leq O(1) + O(1) \cdot \int_{1}^{\ell+1} xe^{-x} \mathrm{d}x \\ &\leq O\left(1 + \frac{2}{e} - \frac{\log n}{e \cdot n} - \frac{1}{e \cdot n}\right) = O(1).\end{aligned}$$

Note that the above argument is independent of the choice of the vertex v being inserted at any step.

The proof is similar to the proof of Theorem 4.2. $\square$

Theorem 4.5. The expected number of iterations that 2-opt needs to find a local optimum is bounded by $O(n^8 \log^3 n)$.

Proof. The proof is similar to the analysis of 2-opt by Englert et al. [13]. Consider a 2-exchange where edges e_1 and e_2 are replaced by edges f_1 and f_2. Here, $e_1 = \{u, v\}$, $e_2 = \{x, y\}$, $f_1 = \{u, y\}$ and $f_2 = \{x, v\}$. The improvement obtained from this exchange is given by $\Delta(e_1, e_2, f_1, f_2) = d(u, v) + d(x, y) - d(u, y) - d(x, v)$.

We estimate the probability $\Pr[\Delta(e_1, e_2, f_1, f_2) \in (0, \varepsilon]]$ of the event that the improvement is at most ε for some $\varepsilon > 0$. The distances $d(u, v)$, $d(x, y)$, $d(u, y)$, and $d(x, v)$ correspond to shortest paths with respect to the edge weights w from which we obtained d. Hence, we can write them as sums of edge weights, assuming for now that they are fixed. This means that we can rewrite the improvement as

$$\begin{aligned}\Delta(e_1, e_2, f_1, f_2) = w(e_{1,1}) + \ldots + w(e_{1,k_1}) + w(e_{2,1}) + \ldots + w(e_{2,k_2}) \\ - w(f_{1,1}) - \ldots - w(f_{1,m_1}) - w(f_{2,1}) - \ldots - w(f_{2,m_2}). \qquad (1)\end{aligned}$$

If replacing e_1 and e_2 by f_1 and f_2 is indeed a 2-exchange, then $\Delta(e_1, e_2, f_1, f_2) > 0$. This means that there exists at least one edge whose contribution is not canceled in (2).

Now suppose that all edge weights except for one edge e are fixed by an adversary. Then $\Delta(e_1, e_2, f_1, f_2) \in (0, \varepsilon]$ only if $w(e)$ assumes a weight in an interval of length ε. Thus, given the choice of edges in (2), the probability that $\Delta(e_1, e_2, f_1, f_2) \in (0, \varepsilon]$ is bounded from above by ε. To obtain an upper bound for $\Pr\big[\Delta(e_1, e_2, f_1, f_2) \in (0, \varepsilon]\big]$, the first idea might be that we have to take a union bound over the choices for the four shortest paths involved. Unfortunately, the number of possible paths is exponential. However, we assumed that the weights of all but one edge are fixed and that this edge e does not cancel out.

This leaves us with way fewer possibilities: We can just take the union bound over the choice of e in the graph. This leaves us with $O(n^2)$ possibilities.

Let $\delta_{\min}$ be the minimum gain made by any 2-exchange. Since there are at most n^4 different 2-exchanges, we have $\Pr[\delta_{\min} \in (0, \varepsilon]] \leq O(n^6\varepsilon)$. The initial tour has a length of at most $n\Delta_{\max}$. Let T be the number of iterations that 2-opt takes. Then $T \leq n\Delta_{\max}/\delta_{\min}$. Now, $T > x$ implies $n\Delta_{\max}/\delta_{\min} > x$. The event $\Delta_{\max}/\delta_{\min} > x/n$ is contained in the union of the events $\Delta_{\max} > \log x \ln n/n$, and $\delta_{\min} < \ln n \cdot \log x/x$. The first happens with a probability of at most $n^{-\Omega(\log(x))}$ by Lemma 3.7. The second happens with a probability of at most $O(n^6 \log(x)/x)$. Thus, we obtain

$$\Pr[T > x] \leq n^{-\Omega(\log(x))} + O(n^6 \ln n \cdot \log(x)/x).$$

Since the number of iterations is at most $n!$, we obtain an upper bound of

$$\mathbb{E}[T] \leq \sum_{x=1}^{n!} \left(n^{-\Omega(\log(x))} + O(n^6 \ln n \log(x)/x) \right).$$

The sum of the $n^{-\Omega(\log(x))}$ is negligible. The sum of the $O(n^6 \ln n \log(x)/x)$ contributes $O(n^6 \ln n \log(n!)^2) = O(n^8 \log^3 n)$. $\square$

C Proofs of Section 5

Lemma 5.1. Fix $U \subseteq V$ of size k. We have

$$\mathbb{E}[\mathsf{TRIVIAL}] = \mathbb{E}[\mathsf{cost}(U)] = H_{n-1} - H_{k-1} = \ln(n/k) + \Theta(1).$$

Proof. We have $\mathbb{E}[\mathsf{cost}(U)] = \sum_{i=k}^{n-1} \mathbb{E}[\mathrm{Exp}(i)] = \sum_{i=k}^{n-1} \frac{1}{i} = H_{n-1} - H_{k-1}$. Using $H_n = \ln(n) + \Theta(1)$ yields the last equality. $\square$

Lemma C.1. *Let $c > 0$ be sufficiently large, and let $k \leq c'n$ for $c' = c'(c) > 0$ be sufficiently small. Then* $\Pr\left[\mathsf{CENTER} < \ln\binom{n}{k} - \ln\ln\binom{n}{k} - \ln c\right] = n^{-\Omega(c)}$.

Proof. Fix $U \subseteq V$ of size k and consider $\mathsf{cost}(U) \sim \sum_{i=k}^{n-1} \mathrm{Exp}(i)$. This sum of exponentially distributed random variables has the following probability density function [30, p. 308ff]:

$$\begin{aligned}
f(x) &= \left(\prod_{i=k}^{n-1} i\right) \cdot \sum_{j=k}^{n-1} \frac{\exp(-jx)}{\prod_{k \leq \ell < n, \ell \neq j} (\ell - j)} \\
&= \sum_{j=k}^{n-1} e^{-jx} \frac{(n-1)!(-1)^{j-k}}{(k-1)!(n-1-j)!(j-k)!} \\
&= \sum_{j=k}^{n-1} e^{-jx} \binom{n-1-k}{j-k} \binom{n-1}{k} k \cdot (-1)^{j-k} \\
&= k\binom{n-1}{k} e^{-kx} \sum_{j=0}^{n-1-k} \binom{n-1-k}{j} (-1)^j e^{-jx} \\
&= k\binom{n-1}{k} e^{-kx} (1 - e^{-x})^{n-1-k}.
\end{aligned}$$

In the following we set $m := n-1$ to shorten notation. We now want to upper bound $f(x)$ at $x = \ln\left(\frac{m}{ak}\right)$ for a large enough a with $1 \le a \le m/k$ (such an a exists since k is small enough). Plugging in this particular x and bounding $\binom{m}{k} \le m^k e^k / k^k$ yields

$$f(x) = k\binom{m}{k}\frac{a^k k^k (m-ak)^{m-k}}{m^m} \le k(ea)^k \left(1 - \frac{ak}{m}\right)^{m-k}.$$

Using $1 + x \le e^x$ and $m - k = \Omega(m)$, so that $(m-k)/m = \Omega(1)$, yields

$$f(x) \le k(ea)^k \exp(-\Omega(ak)).$$

Since a is large enough, the first two factors are lower order terms that we can hide by the Ω. Thus, we can simplify this further to

$$f(x) \le \exp(-\Omega(ak)).$$

Rearranging this using $a = \frac{m}{k}e^{-x}$ yields

$$f(x) = \exp(-\Omega(m\exp(-x)),$$

which holds for any $x \in [0, \ln\left(\frac{m}{\alpha k}\right)]$ for any sufficiently large $\alpha \ge 1$.

Now we can bound the probability that $\mathsf{cost}(U) \sim \sum_{i=k}^{n-1} \mathrm{Exp}(i)$ is less than $\ln\left(\frac{m}{\alpha k}\right)$. This probability is equal to

$$\begin{aligned}
\int_0^{\ln\left(\frac{m}{\alpha k}\right)} f(x)\mathrm{d}x &= \int_0^{\ln\left(\frac{m}{\alpha k}\right)} f\left(\ln\left(\frac{m}{\alpha k}\right) - x\right)\mathrm{d}x \\
&= \int_0^{\ln\left(\frac{m}{\alpha k}\right)} \exp(-\Omega(\alpha k \exp(x)))\mathrm{d}x \\
&\le \int_0^{\infty} \exp(-\Omega(\alpha k(1+x)))\mathrm{d}x \le \exp(-\Omega(\alpha k))
\end{aligned}$$

since $\int_0^\infty \exp(-\Omega(\alpha k x))\mathrm{d}x = O(1/(\alpha k)) \le 1$ as α is sufficiently large.

In order for CENTER to be less than $\ln\left(\frac{m}{\alpha k}\right)$, one of the subsets $U \subseteq V$ of size k has to have cost less than $\ln\left(\frac{m}{\alpha k}\right)$. We bound the probability of the latter using the Union Bound and get

$$\begin{aligned}
\Pr\left[\mathsf{CENTER} < \ln\left(\frac{m}{\alpha k}\right)\right] &= \Pr\left[\exists U \subseteq V, |U| = k\colon\ \mathsf{cost}(U) < \ln\left(\frac{m}{\alpha k}\right)\right] \\
&\le \binom{n}{k}\Pr\left[\mathsf{cost}(U) < \ln\left(\frac{m}{\alpha k}\right)\right] \\
&\le \binom{n}{k}\exp(-\Omega(\alpha k)).
\end{aligned}$$

By setting $\alpha = \beta\ln\left(\frac{n}{k}\right)$ for sufficiently large $\beta \ge 1$, we fulfill all conditions on α. This yields

$$\Pr\left[\mathsf{CENTER} < \ln\left(\frac{n}{k}\right) - \ln\ln\left(\frac{n}{k}\right) - \ln\beta\right] \le \left(\frac{en}{k}\right)^k \left(\frac{n}{k}\right)^{-\Omega(\beta k)}.$$

Now, since k is sufficiently smaller than n, we have $\frac{en}{k} \leq (\frac{n}{k})^2$. Thus, for β large enough, the right hand side simplifies to $(\frac{n}{k})^{-\Omega(\beta k)}$. Since k is at least 1 and small enough, we have $(\frac{n}{k})^k \geq n$, so that the probability is bounded by $n^{-\Omega(\beta)}$, which finishes the proof. $\square$

To bound the expected value of the quotient $\mathsf{TRIVIAL} / \mathsf{CENTER}$ we further need to bound the probabilities that $\mathsf{TRIVIAL}$ is much too large or CENTER is much too small. This is achieved by the following two lemmas.

Lemma C.2. *Let $k \leq (1-\varepsilon)n$ for some constant $\varepsilon > 0$. Then, for any $c > 0$, we have* $\Pr[\mathsf{CENTER} < c] = O(c)^{\Omega(n)}$.

Proof. Let $w_e \sim \mathrm{Exp}(1)$ be the edge weight we sample for edge e, before we take shortest paths to form the random shortest path metric G. Since $n-k$ vertices have to be connected to the k-center, the cost of the k-center is the sum of $n-k$ shortest path lengths. Thus, the cost of the minimal k-center is at least the sum of the smallest $n-k$ edge weights w_e. If we sort the edge weights w_e in ascending order to $\hat{w}_1, \ldots, \hat{w}_{\binom{n}{2}}$, then the cost of CENTER is at least

$$\sum_{i=1}^{n-k} \hat{w}_i \geq \frac{n-k}{2} \cdot \hat{w}_{(n-k)/2} = \Omega(n) \cdot \hat{w}_{\Omega(n)}.$$

For any $m = \Omega(n)$, $c > 0$, and $N = \binom{n}{2}$, a union bound yields

$$\Pr[\hat{w}_m \leq c/n] \leq \left(\frac{c}{n}\right)^m \binom{N}{m}.$$

Using a weak version of Stirling's formula, we have $\binom{N}{m} \leq \frac{N^m e^m}{m^m} \leq \frac{n^{2m} O(1)^m}{n^m}$ (since $m = \Omega(n)$), which yields

$$\Pr[\hat{w}_m \leq c/n] \leq O(c)^m = O(c)^{\Omega(n)}.$$

Hence, we have

$$\Pr[\mathsf{CENTER} < c] \leq \Pr\left[\hat{w}_{(n-k)/2} < \frac{2c}{n-k}\right] \leq O(c)^{\Omega(n)}.$$

$\square$

Lemma C.3. *For any $c \geq 3$, we have* $\Pr[\mathsf{TRIVIAL} > n^c] \leq \exp(-n^{c/3})$.

Proof. First, observe that the tail bound $\Pr[\mathsf{TRIVIAL} > c \cdot \log n] \leq n^{3-c} \cdot \log^2 n$ simply follows from Lemma 3.7, as $\mathsf{TRIVIAL} > c \cdot \log n \implies \Delta_{max} > c(\log n)/n$. If w_e is the weight of edge e before taking shortest paths, we can bound very roughly $\mathsf{TRIVIAL} \leq n \max_e\{w_e\}$. This yields

$$\begin{aligned}\Pr[\mathsf{TRIVIAL} \leq n^c] &\geq \Pr\left[\max_e\{w_e\} \leq n^{c-1}\right] \geq \Pr\left[\max_{1 \leq i \leq n^2} \mathrm{Exp}(1) \leq n^{c-1}\right] \\ &= \Pr\left[\mathrm{Exp}(1) \leq n^{c-1}\right]^{n^2} = \left(1 - \exp(-n^{c-1})\right)^{n^2} \\ &\geq 1 - n^2 \exp(-n^{c-1}) \geq 1 - \exp(-n^{c-2}) \geq 1 - \exp(-n^{c/3}).\end{aligned}$$

$\square$

Theorem 5.2. Let $k \leq (1-\varepsilon)n$ for some constant $\varepsilon > 0$. Then $\mathbb{E}\left[\frac{\mathsf{TRIVIAL}}{\mathsf{CENTER}}\right] = O(1)$. If we even have $k \leq cn$ for some sufficiently small constant $c \in (0,1)$, then $\mathbb{E}\left[\frac{\mathsf{TRIVIAL}}{\mathsf{CENTER}}\right] = 1 + O\left(\frac{\ln\ln(n/k)}{\ln(n/k)}\right)$.

Proof. Let $T = \mathsf{TRIVIAL}$ and $C = \mathsf{CENTER}$ for short. We have for any $m \geq 0$

$$\mathbb{E}\left[\frac{T}{C}\right] \leq \mathbb{E}\left[\frac{T}{m}\right] + \Pr[C < m] \cdot \mathbb{E}\left[\frac{T}{C}\middle| C < m\right]. \tag{2}$$

Case 1 – $k \leq c'n$, c' sufficiently small: Using Lemma C.1, we can pick $c > 0$ such that

$$\Pr\left[C < \ln\left(\frac{n}{k}\right) - \ln\ln\left(\frac{n}{k}\right) - \ln c\right] \leq n^{-7}.$$

Set $m = \ln\left(\frac{n}{k}\right) - \ln\ln\left(\frac{n}{k}\right) - \ln c$. Then, by Lemma 5.1, we have

$$\mathbb{E}\left[\frac{T}{m}\right] \leq \frac{\ln(n/k) + O(1)}{m} \leq 1 + O\left(\frac{\ln\ln(n/k)}{\ln(n/k)}\right).$$

We show that the second summand of inequality (1) is $O(1/n)$ in the current situation, which shows the claim. We have

$$\begin{aligned}
\Pr[C < m] \cdot \mathbb{E}\left[\frac{T}{C}\middle| C < m\right] &= \Pr[C < m] \cdot \int_0^\infty \Pr[T/C \geq x \mid C < m]\mathrm{d}x \\
&\leq \Pr[C < m] \cdot \left(n^6 + \int_{n^6}^\infty \Pr[T/C \geq x \mid C < m]\mathrm{d}x\right) \\
&\leq n^{-1} + \int_{n^6}^\infty \Pr[T/C \geq x \text{ and } C < m]\mathrm{d}x \\
&\leq n^{-1} + \int_{n^6}^\infty \Pr[T/C \geq x]\mathrm{d}x \\
&\leq n^{-1} + \int_{n^6}^\infty \max\left\{\Pr[T \geq \sqrt{x}], \Pr[C \leq 1/\sqrt{x}]\right\}\mathrm{d}x,
\end{aligned}$$

since if $T/C \geq x$ then $T \geq \sqrt{x}$ or $C \leq 1/\sqrt{x}$. Using Lemmas C.2 and C.3, this yields

$$\begin{aligned}
\Pr[C < m] \cdot \mathbb{E}\left[\frac{T}{C}\middle| C < m\right] &\leq n^{-1} + \int_{n^6}^\infty \max\{\exp(-x^{1/6}), O(1/\sqrt{x})^{\Omega(n)}\}\mathrm{d}x \\
&= O(1/n).
\end{aligned}$$

This shows the second claim.

Case 2 – $c'n < k \leq (1-\varepsilon)n$: We repeat the proof above, now choosing m to be a sufficiently small constant. Then $\Pr[C < m] = O(m)^{\Omega(n)} \leq O(n^{-7})$ by Lemma C.2 and we have

$$\mathbb{E}\left[\frac{T}{m}\right] = \frac{\ln(n/k) + O(1)}{m} = O(1),$$

since $k > c'n$. Together with the first case, this shows the first claim. $\square$