August 2016

MASTER THESIS — APPLIED MATHEMATICS

PROBABILISTIC ANALYSIS OF FACILITY LOCATION ON RANDOM SHORTEST PATH METRICS

Stefan Klootwijk

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)

Department of Discrete Mathematics and Mathematical Programming (DMMP)

Assessment committee:

Dr. B. Manthey
Prof. dr. M.J. Uetz
Dr. ir. W.R.W. Scheinhardt

—

UNIVERSITY OF TWENTE.

Abstract

The facility location problem is an $\mathcal{NP}$-hard optimization problem. Therefore, approximation algorithms are often used to solve large instances. In practical situations, these approximation algorithms perform better than their worst-case approximation ratios suggest. In order to explain this behavior, probabilistic analysis is a widely used tool. Most research on probabilistic analysis of $\mathcal{NP}$-hard optimization problems involving metric spaces, such as the facility location problem, has been focused on Euclidean instances. However, we would like to extend this knowledge to other, more general, metrics, since these provide a better resemblance to real-world instances.
In this thesis we investigate the facility location problem using random shortest path metrics. A random shortest path metric is constructed by drawing independent random edge weights for each edge in a complete graph and then setting the distance between each pair of vertices to the length of a shortest path between them (according to the drawn edge weights). We analyze some probabilistic properties for three simple procedures which give a solution to the facility location problem: opening all facilities, opening one arbitrary facility, and opening a certain number of arbitrary facilities (with that certain number only depending on the facility opening cost).
We show that, for any facility opening cost, at least one of these three procedures yields a $1 + o(1)$ approximation in expectation, unless $f \in \Theta(1/n)$. In the latter case we show that at least one of the three procedures yields an $O(1)$ approximation in expectation.

Contents

1 Introduction

Suppose that we are given the task of deciding in which cities a supermarket chain should open their distribution centers or warehouses in order to be able to provide each supermarket with their supplies. There are costs for each distribution center that is to be opened, and furthermore there are transportation (or connection) costs for delivering goods to each supermarket in a city without distribution center, which depend on the distance to the nearest city with a distribution center. Of course, the goal is to choose a set of cities in which a distribution center is to be opened, such that the total cost is minimized.

This problem is known as the (uncapacitated) facility location problem. Mathematically, it can be described as follows: given a (complete) graph $G = (V, E)$, facility opening cost f_i for each vertex $v_i \in V$ and a distance $d(u, v)$ between each pair of vertices $u, v \in V$, find a subset $U \subseteq V$ of vertices such that the total cost is minimized. Here, the total cost is given by the sum of the opening cost f_i for all vertices $u_i \in U$ and the sum of the 'connection' costs $d(u_v, v)$ for all vertices $v \in V \backslash U$, where u_v is a vertex in U which minimizes the distance $d(u, v)$ over all vertices $u \in U$.

The example above is just one of the real-world applications for this optimization problem. Other examples include deciding the optimal locations for stores, plants or bus depots.
Unfortunately the (uncapacitated) facility location problem is $\mathcal{NP}$-hard [2], as is the case with many such optimization problems, implying that for large-scale instances it is impossible to find an optimal solution within reasonable time. Therefore research on the facility location problem (and other $\mathcal{NP}$-hard problems) has been focused on different heuristics, ranging from straightforward to rather sophisticated, and their worst-case performances. However, it is observed that many such heuristics perform much better in practical situations than their worst-case performance suggests.

Indeed, worst-case analysis often tends to be too pessimistic. It does for example not take into consideration that some inputs leading to the worst-case instances (almost) never occur in practical situations. In order to overcome this gap between the worst-case performance and the actual performance in practical situations, probabilistic analysis is a widely used tool. Probabilistic analysis uses an assumption about a probabilistic distribution for the set with all possible instances. By doing so, it takes into account the observation that in practical situations instances of the problem are likely to (or sometimes even certainly) possess certain properties, for example the triangle inequality.

So far, probabilistic analysis of heuristics for optimization problems like the facility location problem has been focused on instances either using Euclidean space or on instances with independent (random) edge lengths (and thus not necessarily satisfying the triangle inequality in the latter case). This seems logical since these instances are relatively easy to handle from a mathematical point of view. For instances with independent (random) edge lengths, this independency allows the usage of some nice formulas from the probability theory. On the other hand, instances using Euclidean space have a nice structure which can be exploited to achieve good results.

However, we would like to apply probabilistic analysis to more general metric instances. In order to do so, in this thesis we will use the model of so-called random shortest path metrics, which have also used by Bringmann et al. [1], who initiated this research. The construction of random shortest path metrics starts with drawing random edge weights independently from an exponential distribution with parameter one, for all edges of an undirected complete graph. The distance between each pair of vertices is then given by the shortest path between them, according to the (random) edge weights. Due to this construction the distances between any two vertices are not independent anymore, making it more difficult to derive results when using this model.

As far as we are aware, only a few problems have currently been studied in this model, namely minimum-weight perfect matching, the traveling salesman problem and the k-median problem. So, there are still an awful lot of (combinatorial optimization) problems which have not been studied yet in this model. In order to extend the knowledge about the properties of random shortest path metrics and to study more problems in this model, we have started with applying probabilistic analysis to some of the most 'simple' heuristics for the facility location problem, namely just opening a fixed number of arbitrary facilities.

1.1 Related work

For general metric instances of the facility location problem, it is known that there cannot exist a polynomial-time approximation algorithm with an approximation ratio less than $\rho^* \approx 1.463$ (where ρ^* is the unique solution of $\rho^* = 1 + 2e^{-\rho^*}$), unless $\mathcal{P} = \mathcal{NP}$. This is a result due to Guha and Khuller [5] and Sviridenko (see [17, Sect. 4.4]). On the other hand, currently the best known worst-case approximation ratio for a polynomial-time algorithm is given by 1.488, due to Li [12].

Flaxman et al. studied random instances of the facility location problem using Euclidean space [4]. They expected to prove that the algorithm of Mahdian et al. [13], which had with 1.52 the best worst-case approximation ratio for a polynomial time algorithm at the time, would be asymptotically optimal under these circumstances, but instead showed that this was not the case for this sophisticated approximation algorithm, and they did the same thing for two related approximation algorithms of Jain et al. [9] (with worst-case approximation ratios 1.861 and 1.61, respectively). On the other hand, Flaxman et al. described a 'trivial heuristic' which is asymptotically optimal.

The model of random shortest path metrics is also known as first-passage percolation, introduced by Hammersley and Welsh as a model for fluid flow through a (random) porous medium [6, 8]. The 'standard' first-passage percolation models use $\mathbb{Z}^d$ as a graph, i.e., a (multidimensional) grid. But the complete graph, which we are interested in, has also been studied. Many structural properties, such as the number of edges on the shortest path between two arbitrary vertices [7], have been analyzed for first-passage percolation on the complete graph.

Although a lot of studies have been conducted on random shortest path metrics, or first-passage percolation, systematic research of the behavior of (simple) heuristics for optimization problems on random shortest path metrics has been initiated only recently, by Bringmann et al. [1]. They provide some structural properties of random shortest path metrics, including the existence of a good clustering. These properties are then used for a probabilistic analysis of simple algorithms for several optimization problems, including the minimum-weight perfect matching problem and the k-median problem.

1.2 Summary of results

The goal of this research is to extend the knowledge about the probabilistic behavior of (simple) heuristics for optimization problems using random shortest path metrics. In this thesis, we will do so by investigating the probabilistic properties of the three most trivial procedures which give a solution to the facility location problem: opening all facilities, opening one arbitrary facility, and opening a certain number of arbitrary facilities (with that certain number only depending on the facility opening cost). Due to the simple structure of these procedures, our results will essentially be structural results on random shortest path metrics.

We show that the most trivial procedure of opening a fixed number of arbitrary facilities (with that fixed number only depending on the facility opening cost f) yields a $1 + o(1)$ approximation in expectation, unless $f \in \Theta(1/n)$ (Theorems 6.5, 6.6, 6.11 and 6.16). If $f \in \Theta(1/n)$ then this procedure is shown to yield an $O(1)$ approximation in expectation (Theorems 6.5 and 6.16).

1.3 Structure of this thesis

The remainder of this thesis is structured as follows. In Section 2 we first introduce briefly some basic notation that will be used throughout this thesis. Then we give a detailed definition of the random shortest path metrics that will be used in this thesis, also providing some basic known results about these random shortest path metrics (Section 2.1). After that, we take a closer look at the (uncapacitated) facility location problem and make a connection between its solutions and solutions to the (related) k-median problem (Section 2.2). Then we introduce the three trivial heuristics for the (uncapacitated) facility location problem, whose behavior will be the main focus of this thesis (Section 2.3).

Sections 3–6 mainly consist of mathematical lemmas and theorems and their proofs, together providing the main results of this thesis. The long way to these results starts with some rather technical results, which are presented in Section 3. Then we take a look at the properties of the three trivial heuristics in Section 4. First we derive the probability distributions for the solutions corresponding to these heuristics, and provide (bounds for) the corresponding expected values (Section 4.1). Then we continue the journey towards the main results by deriving an asymptotic bound for an expression involving the cumulative distribution function of one of the heuristics (Section 4.2).
Section 5 is dedicated to the properties of the optimal solution to the (uncapacitated) facility location problem. In this section we derive various bounds and equalities involving either the cumulative probability distribution of the optimal solution or the expected value of the reciprocal of the optimal solution.

In Section 6 our main theorems are stated and proven, accompanied by some lemmas intended to improve the readability of the proofs of the main theorems. Section 6.1 deals with the heuristic which opens all facilities and shows that this heuristic yields either a constant approximation ratio in expectation or a $1 + o(1)$ approximation ratio in expectation when the facility opening cost is relatively small (Theorem 6.5).
Section 6.2 deals with the heuristic which opens only one arbitrary facility and shows that this heuristic yields either a constant approximation ratio in expectation or a $1 + o(1)$ approximation ratio in expectation when the facility opening cost is relatively large (Theorems 6.6 and 6.11).
Section 6.3 in the end deals with the heuristic which opens some arbitrary facilities and shows that this heuristic yields either a constant approximation ratio in expectation or a $1 + o(1)$ approximation ratio in expectation when the facility opening cost is neither relatively small nor relatively large.

Finally, in Section 7 we provide an overview for the main results from this thesis. From this overview it can be seen that for any facility opening cost at least one of the three trivial heuristics yields at least a constant approximation ratio in expectation. Moreover, in most cases this can be improved to a $1 + o(1)$ approximation ratio in expectation. We conclude this thesis with a short discussion and some final remarks in Section 8.

2 Model description and notation

In this section we describe the random shortest path metrics, as introduced by Bringmann et al. [1], and we will repeat some structural properties of these random shortest path metrics which are used in this thesis. After that, we will give a model description of the facility location problem and introduce three different, but very simple, trivial solutions for that problem. However, before we continue with the random shortest path metrics and the facility location problem, we introduce some general notation.

We use $X \sim P$ to denote that a random variable X is distributed using a probability distribution P. $\text{Exp}(\lambda)$ is being used to denote the exponential distribution with parameter λ. In particular, we use $X \sim \sum_{i=1}^{n} \text{Exp}(\lambda_i)$ to denote that X is the sum of n independent exponentially distributed random variables having parameters $\lambda_1, \ldots, \lambda_n$.
We use $\mathbb{E}[X]$, $\text{Var}(X)$ and $\text{Cov}(X, Y)$ to denote the expected value of a random variable X, the variance of a random variable X, and the covariance of two random variables X and Y, respectively. For $n \in \mathbb{N}$, we use $[n]$ as shorthand notation for $\{1, \ldots, n\}$, and furthermore we use H_n as shorthand notation for the nth harmonic number, i.e. $H_n = \sum_{i=1}^{n} 1/i$. Lastly, for $x \in \mathbb{R}$, we use $\lfloor x \rfloor$ and $\lceil x \rceil$ to denote the floor and ceiling of x, respectively, i.e. $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leq x\}$ and $\lceil x \rceil = \min\{n \in \mathbb{Z} \mid n \geq x\}$.

2.1 Random Shortest Path Metrics

Random shortest path metrics combine properties of both Euclidean space and arbitrary random instances (in which the triangle inequality does not necessarily hold). Loosely speaking, random shortest path metrics can be constructed as follows: given an undirected complete graph, draw edge weights for each edge independently and then define the distance between any two vertices as the total weight of the shortest path between them, measured with respect to the random weights [1]. Following Bringmann et al. [1], we will be using the exponential distribution for drawing the edge weights, since these are technically the easiest to handle due to their memorylessness*.

To be more precise, we define the random shortest path metric as follows. Consider an undirected complete graph $G = (V, E)$ on n vertices. For any edge $e \in E$, let $w(e) \sim \text{Exp}(1)$ independently. Now define the distances $d(u, v)$ between each pair of vertices $u, v \in V$ as the minimum total weight of a u, v-path in G, or, equivalently,

$$d(u, v) = \min_{u,v\text{-path } \mathcal{P}} \sum_{e \in \mathcal{P}} w(e).$$

Using this definition, it is easy to see that $d(v, v) = 0$ for all $v \in V$ and that $d(u, v) = d(v, u)$ for all $u, v \in V$, since any u, v-path $\mathcal{P}$ in G is also a v, u-path in G using exactly the same edges e. Furthermore, from this definition it can be seen that the triangle inequality holds: $d(u, v) \leq d(u, s) + d(s, v)$ for all $u, s, v \in V$, since any concatenation of a u, s-path $\mathcal{P}'$ in G and a s, v-path $\mathcal{P}''$ in G induces a u, v-path $\mathcal{P}$ in G (after removing any cycles that may have arisen due to the concatenation) which uses only edges which are present in either $\mathcal{P}'$ or $\mathcal{P}''$.

In order to give a rough idea about the basic structure of random shortest path metrics, we mention briefly three known results on the expected value of the distances $d(u, v)$ as n grows large. First of all, the expected distance between two arbitrary fixed vertices $u, v \in V$ is approximately $\ln(n)/n$ [1, 3, 10], i.e. $\mathbb{E}[d(u, v)] \approx \ln(n)/n$ as $n \to \infty$. Secondly, the expected maximum distance between

*A (continuous, non-negative) probability distribution of a random variable X is said to be memoryless if and only if $\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$ for all $s, t \geq 0$ [16, p.294] a fixed vertex $u \in V$ and any arbitrary vertex $v \in V$ is approximately $2\ln(n)/n$ [1, 10], i.e. $\mathbb{E}[\max_{v\in V} d(u,v)] \approx 2\ln(n)/n$ as $n \to \infty$. Lastly, the expected maximum distance between any two arbitrary vertices $u, v \in V$ is approximately $3\ln(n)/n$ [10], i.e. $\mathbb{E}[\max_{u,v\in V} d(u,v)] \approx 3\ln(n)/n$ as $n \to \infty$.

2.2 Facility Location Problem

In the (uncapacitated) facility location problem we are given a complete undirected graph $G = (V, E)$ on n vertices, distances $d : V \times V \to \mathbb{R}_{\geq 0}$ between each pair of vertices, and opening cost $f : V \to \mathbb{R}_{\geq 0}$. In this thesis we consider the facility location problem on random shortest path metrics, so the distances $d(u, v)$ are defined as described in Section 2.1. Moreover, we assume throughout this thesis that every vertex has the same opening cost f for a facility, where f may depend on the total number of vertices n (and is very likely to do so).
A solution to the facility location problem is given by a nonempty subset $U \subseteq V$, and the total cost of such a solution is given by

$$\mathsf{cost}(U) = f \cdot |U| + \sum_{v\in V\setminus U} \min_{u\in U} d(v,u). \tag{2.1}$$

The goal of the facility location problem is to find a nonempty subset $U \subseteq V$ such that $\mathsf{cost}(U)$ is minimal. This problem is $\mathcal{NP}$-hard [2]. If we let OPT denote the total cost of an optimal solution to the facility location problem, then it follows that

$$\mathsf{OPT} = \min_{\varnothing\neq U\subseteq V} \mathsf{cost}(U).$$

The facility location problem is closely related to the k-median problem, in which we have the additional requirement that $|U| = k$ for some constant k. Since the number of facilities that must be opened is fixed, the total cost of a solution U to the k-median problem is given by

$$\mathsf{cost}_k(U) = \sum_{v\in V\setminus U} \min_{u\in U} d(v,u).$$

Observe that it follows from the equations above that the equality $\mathsf{cost}(\mathsf{U}) = f \cdot |U| + \mathsf{cost}_{|U|}(U)$ holds for any nonempty subset $U \subseteq V$. Moreover, if we let OPT_k denote the total cost of an optimal solution to the k-median problem, then it follows that

$$\mathsf{OPT}_k = \min_{\substack{\varnothing\neq U\subseteq V\\ |U|=k}} \mathsf{cost}_k(U).$$

2.3 Three trivial heuristics

In this section we give a more precise description of the three solutions that will be investigated in this thesis, which are based on the following observations.
If the opening cost f for the (uncapacitated) facility location problem is (almost) zero, then it is easy to see that the optimal solution to the facility location problem will open a facility at (almost) every vertex $v \in V$. On the other hand, if this opening cost f is relatively large, the optimal solution to the facility location problem will open a facility at only one vertex $v \in V$. Somewhere in between these two extremes, the optimal solution will open a facility at some vertices $v \in V$.

Based on these observations, we define three different, but very simple, trivial solutions to the facility location problem. The first solution simply opens a facility at every vertex, i.e. we have

$U = V$. Throughout the remainder of this thesis the total cost corresponding to this solution will be referred to as ALL. Based on the observations above, this solution is likely to perform well whenever f is relatively small. The second solution opens only a facility at exactly one arbitrary vertex $v \in V$, i.e. we have $U = \{v\}$. Throughout the remainder of this thesis the total cost corresponding to this solution will be referred to as ONE. Based on the observations above, this solution is likely to perform well whenever f is relatively large.
The third solution opens a facility at $\lceil 1/f \rceil$ arbitrary vertices $v_1, \ldots, v_{\lceil 1/f \rceil} \in V$, i.e. we have $U = \{v_1, \ldots, v_{\lceil 1/f \rceil}\}$. Throughout the remainder of this thesis the total cost corresponding to this solution will be referred to as SOME. Based on the observations above, this solution is likely to perform well whenever f is neither relatively small nor relatively large.

The choice for opening a total of $\lceil 1/f \rceil$ facilities in the solution SOME is intuitively based on the following observation. Let OPT denote the optimal solution to the facility location problem and let OPT_k denote the optimal solution to the k-median problem for $k = 1, 2, \ldots, n$. Then it follows that

$$\mathsf{OPT} = \min_{\varnothing \neq U \subseteq V} \mathsf{cost}(U) = \min_{\varnothing \neq U \subseteq V} \left(f \cdot |U| + \mathsf{cost}_{|U|}(U) \right) = \min_{k \in [n]} \left(k \cdot f + \mathsf{OPT}_k \right).$$

Moreover, based on the results of Bringmann et al. [1, Sect. 5] we know that $U_k = \{v_1, \ldots, v_k\}$ is a good approximation for OPT_k (whenever k is not too large) and that $\mathbb{E}[k \cdot f + \mathsf{cost}_k(U_k)] = k \cdot f + \ln(n/k) + \Theta(1)$. Observe that the function $g(k) = k \cdot f + \ln(n/k)$ is minimal for $k = 1/f$. If we combine this, then we see that it is likely that opening approximately $1/f$ arbitrary facilities yields a good approximation for OPT. In this thesis we will show that this is indeed the case whenever f is neither too small nor too large.

3 Some technical results

In this section we present four technical lemmas that are being used for the proofs of our theorems in Section 6. These lemmas do not provide new structural insights, but are nonetheless very helpful.

First of all, several times throughout this thesis we will use the so-called Rényi's representation [14, 15], which links sums of exponentially distributed random variables and order statistics of exponentially distributed random variables. If $X_1, \ldots, X_m$ are m random variables, then $X_{(1)}, \ldots, X_{(m)}$ are the order statistics corresponding to $X_1, \ldots, X_m$ if $X_{(i)}$ is the ith smallest value among $X_1, \ldots, X_m$ for all $i \in [m]$ [16, p.326]. Now, Rényi's representation states the following.

Lemma 3.1. *Let $X_1, \ldots, X_m$ be independent identically distributed random variables with $X_i \sim \mathrm{Exp}(\lambda)$, and let $X_{(1)}, \ldots, X_{(m)}$ be the order statistics corresponding to $X_1, \ldots, X_m$. Then, for any $i \in [m]$,*

$$X_{(i)} = \frac{1}{\lambda} \sum_{j=1}^{i} \frac{Z_j}{m-j+1},$$

where $Z_j \sim \mathrm{Exp}(1)$ independently, and where "=" means equal distribution.

A special case of Rényi's representation is given by the following corollary.

Corollary 3.2. *Let $Y_1, \ldots, Y_{n-1}$ be independent identically distributed random variables with $Y_i \sim \mathrm{Exp}(1)$, and let $Y_{(1)}, \ldots, Y_{(n-1)}$ be the order statistics corresponding to $Y_1, \ldots, Y_{n-1}$. Then, for any $i \in [n-1]$,*

$$Y_{(n-i)} \sim \sum_{k=i}^{n-1} \mathrm{Exp}(k).$$

Proof. Let $Z_j \sim \mathrm{Exp}(1)$ independently. Using Lemma 3.1 (with $\lambda = 1$ and $m = n-1$) it follows that

$$Y_{(n-i)} = \sum_{j=1}^{n-i} \frac{Z_j}{n-j} = \sum_{k=i}^{n-1} \frac{Z_{n-k}}{k} \sim \sum_{k=i}^{n-1} \mathrm{Exp}(k),$$

since $\mathrm{Exp}(1)/k \sim \mathrm{Exp}(k)$. □

Secondly, we want to be able to bound the probability of two dependent events A and B occurring simultaneously. If A and B are independent, then we know that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. However, when A and B are not independent, this is not the case. Therefore the following lemma gives a bound for $\mathbb{P}(A \cap B)$ in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$, for the case in which A and B are dependent.

Lemma 3.3. *Let A and B be two arbitrary events (not necessarily independent). Then it follows that*

$$\mathbb{P}(A \cap B) \leq \mathbb{P}(A)\mathbb{P}(B) + \sqrt{\mathbb{P}(A)\mathbb{P}(B)}.$$

Proof. Define the indicator random variables X and Y by

$$X = \begin{cases} 1, & \text{if } A \text{ occurs,} \\ 0, & \text{if } A \text{ does not occur,} \end{cases} \qquad \text{and} \qquad Y = \begin{cases} 1, & \text{if } B \text{ occurs,} \\ 0, & \text{if } B \text{ does not occur.} \end{cases}$$

It follows from the definitions of X and Y that $\mathbb{E}[X] = \mathbb{E}[X^2] = \mathbb{P}(A)$ and $\mathbb{E}[Y] = \mathbb{E}[Y^2] = \mathbb{P}(B)$.

Furthermore, we have $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \mathbb{P}(A) - (\mathbb{P}(A))^2$ and $\mathrm{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \mathbb{P}(B) - (\mathbb{P}(B))^2$. Moreover, it can easily be seen that the random variable XY is given by

$$XY = \begin{cases} 1, & \text{if both } A \text{ and } B \text{ occur,} \\ 0, & \text{if either } A \text{ or } B \text{ does not occur.} \end{cases}$$

Thus, it follows that $\mathbb{E}[XY] = \mathbb{P}(A \cap B)$. Now, the probability $\mathbb{P}(A \cap B)$ can be computed as follows:

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{E}[XY] \\ &= \mathbb{E}[X]\mathbb{E}[Y] + \mathrm{Cov}(X, Y) \\ &= \mathbb{P}(A)\mathbb{P}(B) + \mathrm{Cov}(X, Y). \end{aligned}$$

Now we use the well-known variance-bound for the covariance, i.e. we use the inequality $\mathrm{Cov}(X, Y) \leq \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$. Then we obtain

$$\begin{aligned} \mathbb{P}(A \cap B) &\leq \mathbb{P}(A)\mathbb{P}(B) + \sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)} \\ &= \mathbb{P}(A)\mathbb{P}(B) + \sqrt{(\mathbb{P}(A) - (\mathbb{P}(A))^2)\,(\mathbb{P}(B) - (\mathbb{P}(B))^2)} \\ &\leq \mathbb{P}(A)\mathbb{P}(B) + \sqrt{\mathbb{P}(A)\mathbb{P}(B)}. \end{aligned}$$

Thus, indeed $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)\mathbb{P}(B) + \sqrt{\mathbb{P}(A)\mathbb{P}(B)}$ for any two arbitrary events A and B. □

Next, we want to be able to bound the expected value of the ratio X/Y for two dependent nonnegative variables X and Y conditioned on the event that Y is relatively small, i.e. $Y < y$ for some y. The next lemma gives such a bound.

Lemma 3.4. *Let X and Y be two arbitrary nonnegative random variables (not necessarily independent) and assume that $\mathbb{P}(Y \leq \beta) = 0$ for some $\beta > 0$. Then, for any y, we have*

$$\mathbb{P}(Y < y) \cdot \mathbb{E}\left[\frac{X}{Y} \,\middle|\, Y < y\right] \leq \frac{1}{\beta^2} \cdot \mathbb{P}(Y < y) + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(X \geq \sqrt{x}\right) dx.$$

Proof. The expected value on the left hand side can be computed and bounded as follows:

$$\begin{aligned} \mathbb{P}(Y < y) \cdot \mathbb{E}\left[\frac{X}{Y} \,\middle|\, Y < y\right] &= \mathbb{P}(Y < y) \cdot \int_0^{\infty} \mathbb{P}\left(\frac{X}{Y} \geq x \,\middle|\, Y < y\right) dx \\ &\leq \mathbb{P}(Y < y) \cdot \left(\frac{1}{\beta^2} + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(\frac{X}{Y} \geq x \,\middle|\, Y < y\right) dx\right) \\ &= \frac{1}{\beta^2} \cdot \mathbb{P}(Y < y) + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(\frac{X}{Y} \geq x \text{ and } Y < y\right) dx \\ &\leq \frac{1}{\beta^2} \cdot \mathbb{P}(Y < y) + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(\frac{X}{Y} \geq x\right) dx. \end{aligned}$$

Observe that $X/Y \geq x$ implies $X \geq \sqrt{x}$ or $Y \leq 1/\sqrt{x}$. Using this observation yields

$$\begin{aligned} \mathbb{P}(Y < y) \cdot \mathbb{E}\left[\frac{X}{Y} \,\middle|\, Y < y\right] &\leq \frac{1}{\beta^2} \cdot \mathbb{P}(Y < y) + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(X \geq \sqrt{x} \text{ or } Y \leq \frac{1}{\sqrt{x}}\right) dx \\ &\leq \frac{1}{\beta^2} \cdot \mathbb{P}(Y < y) + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(X \geq \sqrt{x}\right) dx + \int_{1/\beta^2}^{\infty} \mathbb{P}\left(Y \leq \frac{1}{\sqrt{x}}\right) dx. \end{aligned}$$

Since $\mathbb{P}(Y \leq \beta) = 0$, the second integral vanishes, which leaves us with the desired result. □

Lastly, we want to have an upper bound for the lower tail of the distribution of the sum of exponentially distributed independent random variables. The following lemma gives a bound for that distribution that will be useful.

Lemma 3.5. [11, Theorem 5.1(iii)] *Let* $X = \sum_{i=1}^{n} X_i$ *with* $X_i \sim \mathrm{Exp}(a_i)$ *independent, and define* $\mu := \mathbb{E}[X] = \sum_{i=1}^{n} 1/a_i$ *and* $a_* := \min_i a_i$*. Then, for any* $\lambda \leq 1$ *we have*

$$\mathbb{P}\left(X \leq \lambda\mu\right) \leq e^{-a_*\mu(\lambda - 1 - \ln(\lambda))}.$$

4 Properties of the trivial heuristics

In this section we will derive some properties for the heuristic solutions ALL, ONE and SOME, as introduced in Section 2.3. First we will take a look at the probability distributions of these solutions and derive (bounds for) their expected values. After that, we will derive a specific bound for the probability distribution of SOME.

4.1 Probability distributions

Recall from Section 2.3 that ALL represents the total cost of the solution which opens a facility at every vertex $v \in V$. Using equation (2.1) we can now deduce that

$$\mathsf{ALL} = \mathsf{cost}(V) = f \cdot |V| + \sum_{v \in V \setminus V} \min_{u \in V} d(v,u) = nf + \sum_{v \in \varnothing} \min_{u \in V} d(v,u) = nf.$$

In other words, ALL has in fact a degenerate probability distribution with $\mathbb{P}(\mathsf{ALL} = nf) = 1$.

The probability distribution for ONE and SOME is less trivial. In order to be able to derive them, we use a result from Bringmann et al. [1, Sect. 5] concerning the distribution of a trivial solution to the k-median problem which picks k arbitrary vertices. They show that for $U = \{v_1, \ldots, v_k\}$ it follows that

$$\mathsf{cost}_k(U) \sim \sum_{i=k}^{n-1} \mathrm{Exp}(i). \tag{4.1}$$

Now, recall from Section 2.3 that ONE represents the total cost of the solution which opens only a facility at exactly one arbitrary vertex $v_1 \in V$. Using equation (4.1) we can now deduce that

$$\mathsf{ONE} = \mathsf{cost}(\{v_1\}) = f \cdot |\{v_1\}| + \mathsf{cost}_1(\{v_1\}) \sim f + \sum_{i=1}^{n-1} \mathrm{Exp}(i).$$

From this probability distribution we can easily derive the expected value of ONE. It follows that

$$\mathbb{E}[\mathsf{ONE}] = \mathbb{E}\left[f + \sum_{i=1}^{n-1} \mathrm{Exp}(i)\right] = f + \sum_{i=1}^{n-1} \frac{1}{i} = f + H_{n-1} = f + \ln(n) + \Theta(1). \tag{4.2}$$

Before we derive the cumulative distribution function of ONE, we state a useful lemma which will help us derive it:

Lemma 4.1. [1, Lemma 3.2] *Let* $X \sim \sum_{i=1}^{n} \mathrm{Exp}(ci)$. *Then* $\mathbb{P}(X \le \alpha) = (1 - e^{-c\alpha})^n$ *(for any* $\alpha \ge 0$*).*

Using this lemma, it follows that for any $x \ge f$ we have

$$\mathbb{P}(\mathsf{ONE} \le x) = \mathbb{P}\left(f + \sum_{i=1}^{n-1} \mathrm{Exp}(i) \le x\right) = \mathbb{P}\left(\sum_{i=1}^{n-1} \mathrm{Exp}(i) \le x - f\right) = \left(1 - e^{-(x-f)}\right)^{n-1}.$$

Furthermore, it follows trivially that $\mathbb{P}(\mathsf{ONE} \le x) = 0$ whenever $x < f$. Summarizing these results, we have

$$\mathbb{P}(\mathsf{ONE} \le x) = \begin{cases} \left(1 - e^{-(x-f)}\right)^{n-1} & \text{if } x \ge f, \\ 0 & \text{if } x < f. \end{cases} \tag{4.3}$$

Recall from Section 2.3 that SOME represents the total cost of the solution which opens facilities at exactly $\lceil 1/f \rceil$ arbitrary vertices $v_1, \dots, v_{\lceil 1/f \rceil} \in V$. Using equation (4.1) we can now deduce that

$$\mathsf{SOME} = \mathsf{cost}(\{v_1, \dots, v_{\lceil 1/f \rceil}\}) = f \cdot \lceil 1/f \rceil + \mathsf{cost}_{\lceil 1/f \rceil}(\{v_1, \dots, v_{\lceil 1/f \rceil}\}) \sim f \cdot \lceil 1/f \rceil + \sum_{i=\lceil 1/f \rceil}^{n-1} \operatorname{Exp}(i). \tag{4.4}$$

From this probability distribution we can derive a bound for the expected value of SOME. It follows that

$$\begin{aligned}
\mathbb{E}[\mathsf{SOME}] &= \mathbb{E}\left[f \cdot \lceil 1/f \rceil + \sum_{i=\lceil 1/f \rceil}^{n-1} \operatorname{Exp}(i) \right] \\
&= f \cdot \lceil 1/f \rceil + H_{n-1} - H_{\lceil 1/f \rceil - 1} \\
&= f \cdot \lceil 1/f \rceil + \ln\left(\frac{n}{\lceil 1/f \rceil} \right) + \Theta(1) \\
&\leq f + 1 + \ln(nf) + \Theta(1) \\
&= f + \ln(nf) + \Theta(1). \qquad (4.5)
\end{aligned}$$

In the next subsection we will deal with a bound for the cumulative distribution function of SOME and derive an asymptotic bound that will be used for the proof of one of our theorems in Section 6.

4.2 A bound for the probability distribution of SOME

In this subsection we will derive an asymptotic bound for the expression $\int_{1/f^2}^{\infty} \mathbb{P}(\mathsf{SOME} \geq \sqrt{x})\, dx$ when f is neither too small nor too large (Lemma 4.4). This expression appears in the proof of one of our theorems in Section 6. In order to be able to derive the asymptotic bound, we first derive a bound for the cumulative distribution function of SOME.

Lemma 4.2. *Let $\varepsilon > 0$. For any $f \geq (1+\varepsilon)/n$ and any $t \in \mathbb{R}$, we can bound the probability $\mathbb{P}(\mathsf{SOME} \geq t)$ as follows (whenever n is sufficiently large, implying $\lceil 1/f \rceil \leq n-1$):*

$$\mathbb{P}(\mathsf{SOME} \geq t) \leq \binom{n-1}{\lceil 1/f \rceil} e^{-(t-2)\lceil 1/f \rceil}.$$

Proof. By equation (4.4) we know the distribution of SOME, from which we can derive that

$$\begin{aligned}
\mathbb{P}(\mathsf{SOME} \geq t) &= \mathbb{P}\left(f \cdot \lceil 1/f \rceil + \sum_{i=\lceil 1/f \rceil}^{n-1} \operatorname{Exp}(i) \geq t \right) \\
&= \mathbb{P}\left(\sum_{i=\lceil 1/f \rceil}^{n-1} \operatorname{Exp}(i) \geq t - f \cdot \lceil 1/f \rceil \right) \\
&\leq \mathbb{P}\left(\sum_{i=\lceil 1/f \rceil}^{n-1} \operatorname{Exp}(i) \geq t - 2 \right),
\end{aligned}$$

since $t - f \cdot \lceil 1/f \rceil \geq t - 2$.

Now let $Y_i \sim \text{Exp}(1)$ independently, $i = 1, 2, \ldots, n-1$, and let $Y_{(i)}$ denote the corresponding order statistics. Using Rényi's representation (see Corollary 3.2), we can now further rewrite the last probability as follows:

$$\begin{aligned}
\mathbb{P}(\mathsf{SOME} \geq t) &\leq \mathbb{P}\left(Y_{(n-\lceil 1/f \rceil)} \geq t-2\right) \\
&= \mathbb{P}\left(\exists\, L \subseteq [n-1], |L| = \lceil 1/f \rceil \ : \ \min_{j \in L} Y_j \geq t-2\right) \\
&\leq \binom{n-1}{\lceil 1/f \rceil} \cdot \mathbb{P}\left(\min_{j=1}^{\lceil 1/f \rceil} Y_j \geq t-2\right) \\
&= \binom{n-1}{\lceil 1/f \rceil} \cdot \prod_{j=1}^{\lceil 1/f \rceil} \mathbb{P}\left(Y_j \geq t-2\right) \\
&\leq \binom{n-1}{\lceil 1/f \rceil} e^{-(t-2)\lceil 1/f \rceil},
\end{aligned}$$

where we used the union bound. Note that the last inequality becomes an equality whenever $t-2 \geq 0$. □

Next, we provide a rather technical result that will be useful for the asymptotic bound we are deriving.

Lemma 4.3. *Let f satisfy $(1+\varepsilon)/n \leq f \leq M/n^{\varepsilon}$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$. Then, for n sufficiently large, we have*

$$(nf)^{2/f} e^{(4f-1)/f^2} \leq \frac{1}{n}.$$

Proof. Let n be sufficiently large. Then we have $f \leq 1$. Moreover, it follows that

$$2\ln(nf) + f\ln(n) \leq 3\ln(n) \leq \frac{n^{\varepsilon}}{M} - 4 \leq \frac{1}{f} - 4.$$

Upon dividing this inequality by f and rearranging, we obtain

$$\begin{aligned}
\frac{2}{f} \cdot \ln(nf) + \ln(n) &\leq \frac{1-4f}{f^2} \\
\frac{2}{f} \cdot \ln(nf) + \frac{4f-1}{f^2} &\leq -\ln(n).
\end{aligned}$$

Since e^x is an increasing function of x, we can now deduce that

$$e^{(2/f)\cdot\ln(nf) + (4f-1)/f^2} \leq e^{-\ln(n)},$$

or, equivalently,

$$(nf)^{2/f} e^{(4f-1)/f^2} \leq \frac{1}{n},$$

which shows our claim. □

Finally, we have sufficiently knowledge to prove our desired asymptotic bound.

Lemma 4.4. *Let f satisfy $(1+\varepsilon)/n \le f \le M/n^{\varepsilon}$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$. Then we have*

$$\int_{1/f^2}^{\infty} \mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)\, dx = O\left(\frac{1}{n}\right).$$

Proof. Let n be sufficiently large. Using Lemma 4.2 (with $t = \sqrt{x}$) we bound $\mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)$ as follows:

$$\begin{aligned}\int_{1/f^2}^{\infty} \mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)\, dx &\le \int_{1/2f}^{\infty} \binom{n-1}{\lceil 1/f \rceil} e^{-(\sqrt{x}-2)\lceil 1/f \rceil}\, dx \\ &= \binom{n-1}{\lceil 1/f \rceil} \int_{1/f^2}^{\infty} e^{-(\sqrt{x}-2)\lceil 1/f \rceil}\, dx \\ &= \binom{n-1}{\lceil 1/f \rceil} e^{2\lceil 1/f \rceil} \int_{1/f^2}^{\infty} e^{-\lceil 1/f \rceil \sqrt{x}}\, dx.\end{aligned}$$

Using $\int e^{-a\sqrt{x}}\, dx = -(2/a^2)(1+a\sqrt{x})e^{-a\sqrt{x}}$ (for any constant a), the integral can be evaluated and subsequently rewritten and bounded as follows:

$$\begin{aligned}\int_{1/f^2}^{\infty} \mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)\, dx &\le \binom{n-1}{\lceil 1/f \rceil} e^{2\lceil 1/f \rceil} \left(\frac{2}{\lceil 1/f \rceil^2} \left(1 + \lceil 1/f \rceil \cdot \frac{1}{f}\right) e^{-\lceil 1/f \rceil \cdot 1/f} \right) \\ &= 2\binom{n-1}{\lceil 1/f \rceil} \left(\frac{1}{\lceil 1/f \rceil^2} + \frac{1/f}{\lceil 1/f \rceil} \right) e^{-(1/f-2)\lceil 1/f \rceil} \\ &\le 2\left(\frac{e(n-1)}{\lceil 1/f \rceil} \right)^{\lceil 1/f \rceil} \left(f^2+1\right) e^{2/f - 1/f^2} \\ &\le 2\left(ef(n-1)\right)^{(f+1)/f} \left(f^2+1\right) e^{(2f-1)/f^2}.\end{aligned}$$

By our restrictions on f, it follows that $f \le 1$ (since n is sufficiently large). Using this inequality and $n-1 \le n$, we obtain

$$\begin{aligned}\int_{1/f^2}^{\infty} \mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)\, dx &\le 4(efn)^{2/f} e^{(2f-1)/f^2} \\ &= 4(nf)^{2/f} e^{(4f-1)/f^2}.\end{aligned}$$

Finally, using Lemma 4.3 we obtain that for n sufficiently large

$$\int_{1/f^2}^{\infty} \mathbb{P}\left(\mathsf{SOME} \ge \sqrt{x}\right)\, dx \le \frac{4}{n} = O\left(\frac{1}{n}\right),$$

which finishes the proof of this lemma. □

5 Properties of the optimal solution

In this section we will derive some properties for the optimal solution OPT to the facility location problem. First we will take a look at the different values that OPT can take. Then we will derive two different bounds for the cumulative distribution function of OPT. The first bound is (almost) tight for relative large values, whereas the second bound is (almost) tight for relative small values. At the end of this section we will derive two other results involving OPT which will be used in the proof of one of our theorems in Section 6.

5.1 The maximal value of OPT

Due to the nature of the facility location problem, we know that OPT is a bounded random variable. For the lower bound, we observe that any solution must open a facility at at least one vertex $v \in V$. Therefore we know that $\mathsf{OPT} \geq f$ for any instance. For the upper bound, we observe that $\mathsf{OPT} \leq \mathsf{ALL} = nf$ for any instance. So we know that $f \leq \mathsf{OPT} \leq nf$ for all instances. OPT is a mixed random variable with a continuous part and a discrete part. It can be seen that $\mathbb{P}(\mathsf{OPT} = x) > 0$ if and only if $x = nf$. The following lemma shows how large this probability is.

Lemma 5.1. *For any f, we have $\mathbb{P}\left(\mathsf{OPT} = nf\right) = e^{-n(n-1)f/2}$ and $\mathbb{P}\left(\mathsf{OPT} < nf\right) = 1 - e^{-n(n-1)f/2}$.*

Proof. Note that $\mathsf{OPT} = nf$ if and only if every edge has weight at least f. So, we can write

$$\begin{aligned}
\mathbb{P}\left(\mathsf{OPT} = nf\right) &= \mathbb{P}\left(\min_e w(e) \geq f\right) \\
&= \mathbb{P}\left(\forall e \; w(e) \geq f\right) \\
&= \prod_{j=1}^{\binom{n}{2}} \mathbb{P}\left(w(e_j) \geq f\right) \\
&= \left(\mathbb{P}(w(e) \geq f)\right)^{n(n-1)/2}.
\end{aligned}$$

By our definition of the shortest path metric (see Section 2.1), we know that $w(e)$ is distributed as Exp(1). So, we obtain that

$$\begin{aligned}
\mathbb{P}\left(\mathsf{OPT} = nf\right) &= \left(e^{-f}\right)^{n(n-1)/2} \\
&= e^{-n(n-1)f/2},
\end{aligned}$$

Since $f \leq \mathsf{OPT} \leq nf$ for all instances, it follows that $\mathbb{P}\left(\mathsf{OPT} < nf\right) = 1 - e^{-n(n-1)f/2}$. $\square$

5.2 Two bounds for the probability distribution of OPT

Since very little is known about the precise probability distribution of OPT, it is convenient to have some bounds for its cumulative distribution function. Lemma 5.2 states such a bound, which is (almost) tight for relative large values (i.e. values close to nf), but a lot weaker for relative small values (i.e. values close to f). The bound in Lemma 5.3 has the opposite property: it is almost tight for relative small values, but weak for relative large values.

Lemma 5.2. *Let $2 \leq k \leq n$. Then, for any f, we can bound the probability $\mathbb{P}(\mathsf{OPT} < kf)$ as follows:*

$$\mathbb{P}\left(\mathsf{OPT} < kf\right) \leq \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \leq f\right),$$

with $W_j \sim \text{Exp}(j)$ independently.

Proof. If $\mathsf{OPT} < kf$, then we know that at most $k-1$ facilities can be opened in the optimal solution OPT, or, equivalently, at least $n-k+1$ facilities are closed in the optimal solution OPT. These $n-k+1$ closed facilities are in the optimal solution connected to open facilities using edges with weight at most f (otherwise we can improve the optimal solution by opening such a facility). Therefore, we know that there need to be at least $n-k+1$ edges with weight at most f. In other words, the probability that $\mathsf{OPT} < kf$ is bounded by the probability that there are at least $n-k+1$ edges with weight at most f.
Now, let X_i, $i = 1, 2, \ldots, \binom{n}{2}$ denote the edge weights and note that $X_i \sim \text{Exp}(1)$ by definition. Furthermore let $X_{(i)}$ denote the corresponding order statistics. Now we can write

$$\mathbb{P}\left(\mathsf{OPT} < kf\right) \leq \mathbb{P}\left(X_{(n-k+1)} \leq f\right).$$

Using Rényi's representation (see Lemma 3.1), we can further rewrite this last probability in order to obtain the desired result:

$$\begin{aligned}
\mathbb{P}\left(X_{(n-k+1)} \leq f\right) &= \mathbb{P}\left(\sum_{i=1}^{n-k+1} \frac{Z_i}{\binom{n}{2} - i + 1} \leq f\right) \\
&= \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} \frac{Z_{\binom{n}{2}-j+1}}{j} \leq f\right) \\
&= \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \leq f\right),
\end{aligned}$$

where $Z_i \sim \text{Exp}(1)$ independently and $W_j \sim \text{Exp}(j)$ independently. ☐

Observe that the bound in this lemma is tight for $k = n$, since $\mathbb{P}(\mathsf{OPT} < nf) = 1 - e^{-n(n-1)f/2} = \mathbb{P}(\text{Exp}(n(n-1)/2) \leq f)$ by Lemma 5.1. For relative small values of k the bound in this lemma is relatively weak, but still strong enough in order to show our desired results in Section 6.

Lemma 5.3. *For any f and any $z \in [f, nf]$, we can bound the probability $\mathbb{P}(\mathsf{OPT} < z)$ as follows:*

$$\mathbb{P}(\mathsf{OPT} < z) \leq \sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1 - e^{-(z-if)}\right)^{n-i}.$$

Proof. Note that $\mathsf{OPT} < z$ if and only if exactly i facilities are opened in OPT for some $i \in \{1, 2, \ldots, \lfloor z/f \rfloor\}$. Since these cases are disjoint, we can condition as follows:

$$\begin{aligned}
\mathbb{P}(\mathsf{OPT} < z) &= \sum_{i=1}^{\lfloor z/f \rfloor} \mathbb{P}\left(\mathsf{OPT} < z \mid \mathsf{OPT} \text{ opens exactly } i \text{ facilities}\right) \mathbb{P}\left(\mathsf{OPT} \text{ opens exactly } i \text{ facilities}\right) \\
&\leq \sum_{i=1}^{\lfloor z/f \rfloor} \mathbb{P}\left(\mathsf{OPT} < z \mid \mathsf{OPT} \text{ opens exactly } i \text{ facilities}\right).
\end{aligned}$$

Now, we will bound the conditional probabilities $\mathbb{P}(\mathsf{OPT} < z \mid \mathsf{OPT}$ opens exactly i facilities). Note that

$$\begin{aligned}\mathbb{P}(\mathsf{OPT} < z \mid \mathsf{OPT} \text{ opens exactly } i \text{ facilities}) &= \mathbb{P}(\exists\, U \subseteq V, |U| = i \,:\, \mathsf{cost}_i(U) < z - if) \\ &\leq \binom{n}{i} \cdot \mathbb{P}(\mathsf{cost}_i(U) < z - if) \\ &= \binom{n}{i} \cdot \mathbb{P}\left(\sum_{k=i}^{n-1} \mathrm{Exp}(k) < z - if\right),\end{aligned}$$

where we used the union bound and the known result $\mathsf{cost}_i(U) \sim \sum_{k=i}^{n-1} \mathrm{Exp}(k)$ (see equation (4.1)). Now let $Y_i \sim \mathrm{Exp}(1)$ independently, $i = 1, 2, \ldots, n-1$, and let $Y_{(i)}$ denote the corresponding order statistics. Using Rényi's representation (see Corollary 3.2), we can now further rewrite the last probability as follows:

$$\begin{aligned}\mathbb{P}(\mathsf{OPT} < z \mid \mathsf{OPT} \text{ opens exactly } i \text{ facilities}) &\leq \binom{n}{i} \cdot \mathbb{P}\left(Y_{(n-i)} < z - if\right) \\ &= \binom{n}{i} \cdot \mathbb{P}\left(\exists\, L \subseteq [n-1], |L| = n - i \,:\, \max_{j \in L} Y_j < z - if\right) \\ &\leq \binom{n}{i}\binom{n-1}{n-i} \cdot \mathbb{P}\left(\max_{j=1}^{n-i} Y_j < z - if\right) \\ &= \binom{n}{i}\binom{n-1}{i-1} \cdot \prod_{j=1}^{n-i} \mathbb{P}(Y_j < z - if) \\ &= \binom{n}{i}\binom{n-1}{i-1} \left(1 - e^{-(z-if)}\right)^{n-i},\end{aligned}$$

where we used again the union bound. By combining the results above, the desired result follows now immediately. □

Observe that the bound in this lemma is (almost) tight for $z \in [f, 2f)$, since that implies $\lfloor z/f \rfloor = 1$ and all but one inequalities in the proof above become equalities for $i = 1$. For relative large values of z the bound in this lemma is relatively weak, but still strong enough in order to show our desired results in Section 6.

5.3 Two other results involving OPT

In this subsection we will derive two other results involving OPT, which will be used to prove one of our results in Section 6. First we provide a way to evaluate the conditional expectation $\mathbb{E}[1/\mathsf{OPT} \mid \mathsf{OPT} < nf]$.

Lemma 5.4. *For any f, we have*

$$nf \cdot \mathbb{E}\left[\frac{1}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < nf\right] = 1 + \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_f^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2}\, dy.$$

Proof. The expected value on the left hand side can be computed as follows:

$$\begin{aligned} nf \cdot \mathbb{E}\left[\frac{1}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < nf\right] &= nf \cdot \int_0^\infty \mathbb{P}\left(\frac{1}{\mathsf{OPT}} \geq x \,\middle|\, \mathsf{OPT} < nf\right) dx \\ &= nf \cdot \int_0^\infty \mathbb{P}\left(\mathsf{OPT} \leq \frac{1}{x} \,\middle|\, \mathsf{OPT} < nf\right) dx \\ &= nf \cdot \int_0^\infty \frac{\mathbb{P}\left(\mathsf{OPT} \leq \frac{1}{x} \text{ and } \mathsf{OPT} < nf\right)}{\mathbb{P}\left(\mathsf{OPT} < nf\right)} dx. \end{aligned}$$

Using Lemma 5.1, it follows now that

$$\begin{aligned} nf \cdot \mathbb{E}\left[\frac{1}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < nf\right] &= \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_0^\infty \mathbb{P}\left(\mathsf{OPT} \leq \frac{1}{x} \text{ and } \mathsf{OPT} < nf\right) dx \\ &= \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_0^\infty \mathbb{P}\left(\mathsf{OPT} < \min\left\{\frac{1}{x}, nf\right\}\right) dx. \end{aligned}$$

Now note that $\min\{1/x, nf\} = nf$ if and only if $x < 1/nf$. Using this property, we can split the integral in two parts as follows:

$$\begin{aligned} nf \cdot \mathbb{E}\left[\frac{1}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < nf\right] &= \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \left(\int_0^{1/nf} \mathbb{P}\left(\mathsf{OPT} < nf\right) dx + \int_{1/nf}^\infty \mathbb{P}\left(\mathsf{OPT} < \frac{1}{x}\right) dx\right) \\ &= \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \left(\frac{1 - e^{-n(n-1)f/2}}{nf} + \int_{1/nf}^\infty \mathbb{P}\left(\mathsf{OPT} < \frac{1}{x}\right) dx\right) \\ &= 1 + \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_{1/nf}^\infty \mathbb{P}\left(\mathsf{OPT} < \frac{1}{x}\right) dx, \end{aligned}$$

where we again applied the result of Lemma 5.1. Now, we use the substitution $x = 1/y$ to obtain the result of this lemma.

$$\begin{aligned} nf \cdot \mathbb{E}\left[\frac{1}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < nf\right] &= 1 + \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_{nf}^0 -\frac{1}{y^2}\mathbb{P}\left(\mathsf{OPT} < y\right) dy \\ &= 1 + \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_0^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2} dy \\ &= 1 + \frac{nf}{1 - e^{-n(n-1)f/2}} \cdot \int_f^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2} dy, \end{aligned}$$

since $\mathbb{P}(\mathsf{OPT} < f) = 0$. $\square$

Next we give a bound for the integral $\int_f^{nf} \mathbb{P}(\mathsf{OPT} < y)/y^2 \, dy$ which arose in the foregoing lemma.

Lemma 5.5. *For any f, we have the following bound:*

$$\int_f^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2} dy \leq \sum_{k=2}^{n} \frac{\mathbb{P}(\mathsf{OPT} < kf)}{k(k-1)f}.$$

Proof. We start by splitting the integral on the left hand side into $n-1$ parts. This yields

$$\int_f^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2} dy = \sum_{k=2}^{n} \int_{(k-1)f}^{kf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2} dy.$$

Now, since $\mathbb{P}(\mathsf{OPT} < y)$ is an increasing function of y, we know that $\mathbb{P}(\mathsf{OPT} < y) \leq \mathbb{P}(\mathsf{OPT} < kf)$ for any $y \in [(k-1)f, kf]$. Using this, it follows that

$$\begin{aligned}
\int_{f}^{nf} \frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2}\, dy &\leq \sum_{k=2}^{n} \int_{(k-1)f}^{kf} \frac{\mathbb{P}(\mathsf{OPT} < kf)}{y^2}\, dy \\
&= \sum_{k=2}^{n} \mathbb{P}(\mathsf{OPT} < kf) \int_{(k-1)f}^{kf} \frac{1}{y^2}\, dy \\
&= \sum_{k=2}^{n} \mathbb{P}(\mathsf{OPT} < kf) \cdot \frac{1}{k(k-1)f} \\
&= \sum_{k=2}^{n} \frac{\mathbb{P}(\mathsf{OPT} < kf)}{k(k-1)f},
\end{aligned}$$

which completes the proof. □

6 Probabilistic analysis of the trivial heuristics

In this section we analyze the approximation ratios of the three trivial heuristics introduced in Section 2.3. Depending on the (asymptotic growth) of the facility opening cost f, we will show that at least one of these three heuristics yields a constant approximation ratio in expectation or, even better, is asymptotically optimal (i.e. yields a $1 + o(1)$ approximation ratio in expectation).

The basic idea behind the proofs of the corresponding theorems will be conditioning the expected approximation ratio on the events OPT is larger respectively smaller than a certain threshold. The threshold will be chosen in such a way that for the case in which OPT is larger than the threshold, it is relatively easy to bound the conditional expected approximation ratio. On the other hand, the threshold will also be chosen in such a way that the probability of OPT being smaller than the threshold becomes sufficiently small. By doing so, we will be able to show that the (relatively) large conditional expected approximation ratio in this case becomes negligible when multiplied with that probability. For each heuristic, we will use a different technique in order to show that the product of the probability of OPT being smaller than the threshold and the conditional expected approximation ratio in that case becomes sufficiently small.

Section 6.1 will deal with the heuristic that opens a facility at every vertex $v \in V$, which performs well when the facility opening cost f is relatively small. Then, Section 6.2 will deal with the heuristic that opens a facility at exactly one arbitrary vertex $v \in V$, which performs well when the facility opening cost f is relatively large. Lastly, Section 6.3 will deal with the heuristic that opens a facility at exactly $\lceil 1/f \rceil$ arbitrary vertices $v_1, \ldots, v_{\lceil 1/f \rceil} \in V$, which performs well when the facility opening cost f is neither relatively small nor relatively large.

6.1 Opening all facilities

In this subsection we examine the approximation ratio of the trivial heuristic which opens a facility at every vertex $v \in V$. We show that ALL yields an $O(1)$ approximation for $f < 2/n$ and a $1+o(1)$ approximation when $f = 1/(n\alpha(n))$ for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$ (Theorem 6.5). In order to be able to prove this theorem, we first need another four lemmas.

The idea behind the proof of this theorem is to bound the expected approximation ratio (which contains a random variable with unknown distribution) with a function which contains only random variables with known distribution (in our case this will be sums of exponentially distributed random variables). This is done, using the foregoing Lemmas 5.1, 5.4, 5.5 and 5.2. Next, the idea is to bound the remaining probabilities using Lemma 3.5. Lemmas 6.1 and 6.2 ensure that the conditions of Lemma 3.5 hold. The last step is to show that the resulting explicit function (which is a bound for the expected approximation ratio) is bounded either by $O(1)$ or $1+o(1)$ (depending on the asymptotic value of f). Lemmas 6.3 and 6.4 provide us with the necessary tools in order to be able to do so.

Now, first of all, we need to show that the expression $\lambda_k := f/(H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k})$ is bounded by a positive constant, (slightly) smaller than 1, for almost every k with $2 \leq k \leq n$, when n is sufficiently large. This bound will later on allow us the use of Lemma 3.5, and will afterwards allow us to give an appropriate bound for the term $\lambda_k - 1 - \ln(\lambda_k)$ which arises when using this lemma. The following two lemmas provide this bound (Lemma 6.1 for the case where $f < 2/n$ and Lemma 6.2 for the case where $f = 1/(n\alpha(n))$).

Lemma 6.1. *Let $f \leq (2-\varepsilon)/n$ for n sufficiently large and some constant $\varepsilon > 0$. Furthermore, define $\rho = \varepsilon/3 > 0$ and $\varepsilon' = \varepsilon/(3+\varepsilon) > 0$. Note that $2\rho + \varepsilon' + \rho\varepsilon' = \varepsilon$ by these definitions. Let $c = \varepsilon'/4 \in (0,1)$. Then, for sufficiently large n, it follows that*

$$\lambda_k = \frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k}} \leq 1-\rho$$

for all k with $2 \leq k \leq \lfloor cn \rfloor$.

Proof. Let n be sufficiently large. Since $f \leq (2-\varepsilon)/n = (1-\rho)(2-\varepsilon')/n$ for such n, it follows that $f/(1-\rho) \leq (2-\varepsilon')/n$ for such n and thus

$$\frac{\binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1}{n} \geq \frac{\binom{n}{2}\left(e^{-(2-\varepsilon')/n}-1\right)+n-1}{n}.$$

Using elementary calculus it follows that

$$\lim_{n\to\infty} \frac{\binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1}{n} \geq \lim_{n\to\infty} \frac{\binom{n}{2}\left(e^{-(2-\varepsilon')/n}-1\right)+n-1}{n} = \frac{\varepsilon'}{2}.$$

Using the definition of the limit, we can now deduce that

$$\frac{\binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1}{n} \geq \frac{\varepsilon'}{4},$$

for sufficiently large n. Since $c = \varepsilon'/4$, it follows from the last inequality that for any k with $2 \leq k \leq \lfloor cn \rfloor$ and for sufficiently large n, we have

$$k \leq \binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1.$$

Upon rearranging and taking the logarithm on both sides, we obtain the following:

$$e^{f/(1-\rho)} \leq \frac{\binom{n}{2}}{\binom{n}{2}-n+k+1}$$

$$\frac{f}{1-\rho} \leq \ln\left(\frac{\binom{n}{2}}{\binom{n}{2}-n+k+1}\right).$$

Since $\ln(a/(b+1)) \leq H_a - H_b$ for any $a, b \in \mathbb{N}$, we finally obtain that

$$\frac{f}{1-\rho} \leq H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k}.$$

One last rearrangement finishes the proof of this lemma. □

Lemma 6.2. *Let $f = 1/(n\alpha(n))$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$ and let $c, \rho \in (0,1)$ be two arbitrary constants. Then, for sufficiently large n, it follows that*

$$\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k}} \leq 1-\rho$$

for all k with $2 \leq k \leq \lfloor cn \rfloor$.

Proof. Using elementary calculus it follows that

$$\lim_{n\to\infty} \frac{\binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1}{n} = \lim_{n\to\infty} \frac{\binom{n}{2}\left(e^{-1/((1-\rho)n\alpha(n))}-1\right)+n-1}{n} = 1.$$

Using the definition of the limit, we can now deduce that

$$\frac{\binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1}{n} \geq c,$$

for sufficiently large n. From this inequality it follows that for any k with $2 \leq k \leq \lfloor cn \rfloor$ and for sufficiently large n, we have

$$k \leq \binom{n}{2}\left(e^{-f/(1-\rho)}-1\right)+n-1.$$

Upon rearranging and taking the logarithm on both sides, we obtain the following:

$$\begin{aligned} e^{f/(1-\rho)} &\leq \frac{\binom{n}{2}}{\binom{n}{2}-n+k+1} \\ \frac{f}{1-\rho} &\leq \ln\left(\frac{\binom{n}{2}}{\binom{n}{2}-n+k+1}\right). \end{aligned}$$

Since $\ln(a/(b+1)) \leq H_a - H_b$ for any $a, b \in \mathbb{N}$, we finally obtain that

$$\frac{f}{1-\rho} \leq H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k}.$$

One last rearrangement finishes the proof of this lemma. □

Next, we need a bound for the summation of $1/k(k-1)$ for all those k for which the previous lemmas do not hold. This bound is given by the following lemma.

Lemma 6.3. *For any constant $c \in (0,1)$ it follows that*

$$n \cdot \sum_{k=\lfloor cn\rfloor+1}^{n} \frac{1}{k(k-1)} \leq \frac{1-c}{c}.$$

Proof. Using elementary calculus it follows that

$$n \cdot \sum_{k=\lfloor cn\rfloor+1}^{n} \frac{1}{k(k-1)} = n \cdot \sum_{k=\lfloor cn\rfloor+1}^{n} \left(\frac{1}{k-1} - \frac{1}{k}\right)$$

Now, by telescoping, we have

$$\begin{aligned} n \cdot \sum_{k=\lfloor cn\rfloor+1}^{n} \frac{1}{k(k-1)} &= n \cdot \left(\frac{1}{\lfloor cn\rfloor} - \frac{1}{n}\right) \\ &\leq n \cdot \left(\frac{1}{cn} - \frac{1}{n}\right) \\ &= \frac{1}{c} - 1 = \frac{1-c}{c}, \end{aligned}$$

which finishes the proof. □

Lastly, we need to show that the following expression, which appears in the exponent after the usage of Lemma 3.5, is relatively large. This is done by the following lemma.

Lemma 6.4. *Let f and c satisfy one of the following conditions:*

(i) $f \leq (2-\varepsilon)/n$ *for n sufficiently large and some constant $\varepsilon > 0$, and* $c = \varepsilon/(12+4\varepsilon)$,

(ii) $f = 1/(n\alpha(n))$ *for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$, and $c \in (0,1)$ an arbitrary constant.*

Then, for n sufficiently large, we have

$$\left(\binom{n}{2} - n + 2\right) \cdot \left(H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}\right) \cdot \left(\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}} - 1 - \ln\left(\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}}\right)\right) = \Omega(n).$$

Proof. We show that the three factors are bounded from below by $\Omega(n^2)$, $\Omega(1/n)$ and $\Omega(1)$, respectively. For the first factor, we can immediately see that $\binom{n}{2} - n + 2 = \frac{1}{2}n^2 - \frac{3}{2}n + 2 = \Omega(n^2)$. For the second factor, we have

$$\begin{aligned}
H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor} &= \sum_{i=\binom{n}{2}-n+\lfloor cn \rfloor+1}^{\binom{n}{2}} \frac{1}{i} \\
&\geq \sum_{i=\binom{n}{2}-n+\lfloor cn \rfloor+1}^{\binom{n}{2}} \frac{1}{\binom{n}{2} - n + \lfloor cn \rfloor + 1} \\
&= \frac{1}{\binom{n}{2} - n + \lfloor cn \rfloor + 1} \cdot \sum_{i=\binom{n}{2}-n+\lfloor cn \rfloor+1}^{\binom{n}{2}} 1 \\
&= \frac{n - \lfloor cn \rfloor}{\binom{n}{2} - n + \lfloor cn \rfloor + 1} = \Omega(1/n),
\end{aligned}$$

since $c \in (0,1)$ does not depend on n. Thus, indeed $H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor} = \Omega(1/n)$.
For the third factor, we have by either Lemma 6.1 (if f and c satisfy condition (i)) or Lemma 6.2 (if f and c satisfy condition (ii)) that

$$\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}} \leq 1 - \rho$$

for some constant $\rho > 0$ and n sufficiently large (if f and c satisfy condition (i), then $\rho = \varepsilon/3$; otherwise $\rho \in (0,1)$ is an arbitrary constant). Since the function $g(x) = x - 1 - \ln(x)$ is decreasing for $x \leq 1-\rho$, it follows that $g(x) \geq g(1-\rho)$ for any $x \leq 1-\rho$. Thus, it follows that

$$\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}} - 1 - \ln\left(\frac{f}{H_{\binom{n}{2}} - H_{\binom{n}{2}-n+\lfloor cn \rfloor}}\right) \geq (1-\rho) - 1 - \ln(1-\rho) = \Omega(1),$$

since ρ does not depend on n. □

Now we can finally state and prove the main result from this subsection.

Theorem 6.5. *Let $\delta > 0$. Consider the facility location problem with $f \le (2-\varepsilon)/n$ for n sufficiently large and some constant $\varepsilon > 0$. Furthermore let ALL denote the total cost of the solution in which a facility is opened at every vertex $v \in V$ and let OPT denote the total cost of the optimal solution to the problem. Then*

$$\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] = O(1).$$

Moreover, if $f = 1/(n\alpha(n))$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$, then

$$\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] = 1 + \delta + O\left(ne^{-n}\right).$$

Proof. From our observations in Section 4.1 we know that $\mathbb{P}(\mathsf{ALL} = nf) = 1$. Conditioning on the events $\mathsf{OPT} = nf$ and $\mathsf{OPT} < nf$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &= \mathbb{P}(\mathsf{OPT} = nf)\cdot\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\;\middle|\;\mathsf{OPT} = nf\right] + \mathbb{P}(\mathsf{OPT} < nf)\cdot\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\;\middle|\;\mathsf{OPT} < nf\right]\\
&= \mathbb{P}(\mathsf{OPT} = nf)\cdot\mathbb{E}\left[\frac{nf}{nf}\right] + \mathbb{P}(\mathsf{OPT} < nf)\cdot\mathbb{E}\left[\frac{nf}{\mathsf{OPT}}\;\middle|\;\mathsf{OPT} < nf\right]\\
&= \mathbb{P}(\mathsf{OPT} = nf) + \mathbb{P}(\mathsf{OPT} < nf)\cdot nf\cdot\mathbb{E}\left[\frac{1}{\mathsf{OPT}}\;\middle|\;\mathsf{OPT} < nf\right].
\end{aligned}$$

Now, by Lemma 5.1 and Lemma 5.4, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &= e^{-n(n-1)f/2} + \left(1 - e^{-n(n-1)f/2}\right)\cdot\left(1 + \frac{nf}{1 - e^{-n(n-1)f/2}}\cdot\int_f^{nf}\frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2}\,dy\right)\\
&= 1 + nf\cdot\int_f^{nf}\frac{\mathbb{P}(\mathsf{OPT} < y)}{y^2}\,dy.
\end{aligned}$$

Using Lemma 5.5, we can bound this expression as follows:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\le 1 + nf\cdot\sum_{k=2}^{n}\frac{\mathbb{P}(\mathsf{OPT} < kf)}{k(k-1)f}\\
&= 1 + n\cdot\sum_{k=2}^{n}\frac{\mathbb{P}(\mathsf{OPT} < kf)}{k(k-1)}.
\end{aligned}$$

Now we can bound the (unknown) distribution of OPT using Lemma 5.2, which gives:

$$\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] \le 1 + n\cdot\sum_{k=2}^{n}\frac{1}{k(k-1)}\cdot\mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \le f\right),$$

where $W_j \sim \mathrm{Exp}(j)$ independently. If $f = 1/(n\alpha(n))$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$, then we take $c = 1/(1+\delta) \in (0,1)$. Otherwise, we take $c = \varepsilon/(12+4\varepsilon)$. Note that this implies that f and c satisfy the conditions of Lemma 6.2 (if $f = 1/(n\alpha(n))$) or Lemma 6.1 (otherwise).

Now we split the summation into two parts and bound the probability in the second part by 1:

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \leq f\right) + n \cdot \sum_{k=\lfloor cn \rfloor+1}^{n} \frac{1}{k(k-1)} \cdot 1 \\
&\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \leq f\right) + \frac{1-c}{c},
\end{aligned}$$

where we used the result of Lemma 6.3 in the last inequality. Now note that for any term in the remaining summation we have $\lambda_k = f/\left(H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k}\right) \leq 1 - \rho < 1$ (by Lemma 6.1 or Lemma 6.2). Moreover, the expected values of the stochastic variables in each term of the remaining summations are given by

$$\mu_k := H_{\binom{n}{2}} - H_{\binom{n}{2}-n+k},$$

implying $\lambda_k \mu_k = f$.

Now we use the result of Lemma 3.5 to obtain

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot \mathbb{P}\left(\sum_{j=\binom{n}{2}-n+k}^{\binom{n}{2}} W_j \leq \lambda_k \mu_k\right) + \frac{1-c}{c} \\
&\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot e^{-\left(\binom{n}{2}-n+k\right)\mu_k(\lambda_k - 1 - \ln(\lambda_k))} + \frac{1-c}{c}.
\end{aligned}$$

Note that μ_k decreases and λ_k increases as k increases, i.e., $\mu_{k+1} < \mu_k$ and $\lambda_{k+1} > \lambda_k$. Furthermore, since $\lambda_k \leq 1$ for each k with $2 \leq k \leq \lfloor cn \rfloor$, it follows that $\zeta_k := \lambda_k - 1 - \ln(\lambda_k)$ decreases as k increases, i.e. $\zeta_{k+1} < \zeta_k$. Finally, since both $\mu_k \geq 0$ and $\zeta_k \geq 0$ for $2 \leq k \leq \lfloor cn \rfloor$, we obtain that $\xi_k := \mu_k \zeta_k$ decreases as k increases, i.e., $\xi_{k+1} < \xi_k$.
Using the results above, it is easy to see that we have $\mu_k(\lambda_k - 1 - \ln(\lambda_k)) \geq \mu_{\lfloor cn \rfloor}(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor}))$ for all k with $2 \leq k \leq \lfloor cn \rfloor$. Using this bound, and the obvious bound $\binom{n}{2} - n + k \geq \binom{n}{2} - n + 2$ for $k \geq 2$, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot e^{-\left(\binom{n}{2}-n+k\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} + \frac{1-c}{c} \\
&\leq 1 + n \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} \cdot e^{-\left(\binom{n}{2}-n+2\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} + \frac{1-c}{c} \\
&= 1 + n \cdot e^{-\left(\binom{n}{2}-n+2\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} \cdot \sum_{k=2}^{\lfloor cn \rfloor} \frac{1}{k(k-1)} + \frac{1-c}{c} \\
&\leq 1 + n \cdot e^{-\left(\binom{n}{2}-n+2\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} \cdot \sum_{k=2}^{n} \frac{1}{k(k-1)} + \frac{1-c}{c} \\
&= 1 + n \cdot e^{-\left(\binom{n}{2}-n+2\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} \cdot \left(1 - \frac{1}{n}\right) + \frac{1-c}{c} \\
&= 1 + (n-1) \cdot e^{-\left(\binom{n}{2}-n+2\right)\mu_{\lfloor cn \rfloor}\left(\lambda_{\lfloor cn \rfloor} - 1 - \ln(\lambda_{\lfloor cn \rfloor})\right)} + \frac{1-c}{c}.
\end{aligned}$$

Now we use the result of Lemma 6.4 to obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\le 1 + (n-1)\cdot e^{-\Omega(n)} + \frac{1-c}{c} \\ &= 1 + \frac{1-c}{c} + O\left(ne^{-n}\right) \\ &= O(1).\end{aligned}$$

Moreover, if $f = 1/(n\alpha(n)$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$, then we have $c = 1/(1+\delta)$. Using that, we obtain that

$$\begin{aligned}\mathbb{E}\left[\frac{\mathsf{ALL}}{\mathsf{OPT}}\right] &\le 1 + \frac{1-c}{c} + O\left(ne^{-n}\right) \\ &= 1 + \delta + O\left(ne^{-n}\right).\end{aligned}$$

which finally completes this proof. □

6.2 Opening one facility

In this subsection we examine the approximation ratio of the trivial heuristic which opens a facility at exactly one arbitrary vertex $v \in V$. We first show that ONE yields a $1 + o(1)$ approximation when $f = \alpha(n)\ln(n)$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$ (Theorem 6.6). After that, we show that ONE yields an $O(1)$ approximation for $1/n^{1-\varepsilon} \le f \le \varepsilon\ln(n)$ and also a $1 + o(1)$ approximation when $1/n^{\varepsilon} \le f \le M\ln(n)$ (Theorem 6.11).

Theorem 6.6. *Consider the facility location problem with $f = \alpha(n)\ln(n)$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$. Furthermore let* ONE *denote the total cost of the solution in which a facility is opened at only one arbitrary vertex $v \in V$ and let* OPT *denote the total cost of the optimal solution to the problem. Then*

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] = 1 + O\left(\frac{1}{\alpha(n)}\right).$$

Proof. By equation (4.2) we know that the expected value of ONE is given by

$$\mathbb{E}[\mathsf{ONE}] = f + \ln(n) + \Theta(1).$$

Furthermore, from our observations in Section 5.1 we know that OPT is bounded from below by f. Therefore, we can now deduce that

$$\begin{aligned}\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] &\le \mathbb{E}\left[\frac{\mathsf{ONE}}{f}\right] \\ &= \frac{f + \ln(n) + \Theta(1)}{f} \\ &= 1 + \frac{\ln(n) + \Theta(1)}{f}.\end{aligned}$$

Now, since $f = \alpha(n)\ln(n)$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$, it follows that

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] \le 1 + \frac{1}{\alpha(n)} + \frac{\Theta(1)}{\alpha(n)\ln(n)} = 1 + O\left(\frac{1}{\alpha(n)}\right),$$

which finishes this proof. □

Theorem 6.11 is more difficult to prove. The idea behind the proof of this theorem is to bound the product of the probability of OPT being smaller than a well-chosen threshold and the conditional expected approximation ratio in that case (Lemma 6.10). First, we rewrite this product to an integral containing the probability that both ONE is large and OPT is small. Since these events are dependent, we use Lemma 3.3. Next, we use Lemma 5.3 and equation (4.3) to bound the result with an explicit function. The last step is to show that this explicit function (which is a bound for the expected approximation ratio) is bounded either by $O(1)$ or $1 + o(1)$ (depending on the asymptotic value of f). Lemmas 6.7, 6.8 and 6.9 provide us with the necessary tools in order to be able to do so.

For the proof of Theorem 6.11 as sketched above, we thus first need another four lemmas. First, we need to bound the summation that arises when bounding the cumulative distribution function of OPT using Lemma 5.3.

Lemma 6.7. *Set $m := 2n - 1$ to shorten notation. Then. for any f and any $z \in [f, nf]$, we have*

$$\sqrt{\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i}} \leq \sqrt[4]{(em)^2 \cdot \frac{(em)^{4z/f}}{(em)^4-1} \cdot \frac{\left(1-e^{-(z-f)}\right)^{2n-2z/f}}{e^{-(z-f)}}}.$$

Proof. Observe that it suffices to show that

$$\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i} \leq \sqrt{(em)^2 \cdot \frac{(em)^{4z/f}}{(em)^4-1} \cdot \frac{\left(1-e^{-(z-f)}\right)^{2n-2z/f}}{e^{-(z-f)}}}.$$

We start by bounding the product of the binomials using the inequalities $\binom{n_1}{k_1}\binom{n_2}{k_2} \leq \binom{n_1+n_2}{k_1+k_2}$ and $\binom{n}{k} \leq (en/k)^k$. This results in

$$\begin{aligned}
\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i} &\leq \sum_{i=1}^{\lfloor z/f \rfloor} \binom{m}{2i-1}\left(1-e^{-(z-if)}\right)^{n-i} \\
&\leq \sum_{i=1}^{\lfloor z/f \rfloor} \left(\frac{em}{2i-1}\right)^{2i-1}\left(1-e^{-(z-if)}\right)^{n-i} \\
&\leq \sum_{i=1}^{\lfloor z/f \rfloor} (em)^{2i-1}\left(1-e^{-(z-if)}\right)^{n-i},
\end{aligned}$$

where we used the inequality $2i-1 \geq 1$ for all i in the summation. Next we use the Cauchy-Schwarz inequality to obtain that

$$\begin{aligned}
\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i} &\leq \sqrt{\sum_{i=1}^{\lfloor z/f \rfloor} (em)^{4i-2} \cdot \sum_{i=1}^{\lfloor z/f \rfloor} \left(1-e^{-(z-if)}\right)^{2n-2i}} \\
&\leq \sqrt{\sum_{i=1}^{\lfloor z/f \rfloor} (em)^{4i-2} \cdot \sum_{i=1}^{\lfloor z/f \rfloor} \left(1-e^{-(z-f)}\right)^{2n-2i}},
\end{aligned}$$

since $(1 - e^{-(z-if)}) \leq (1 - e^{-(z-f)})$ for all i in the summation.

Now we can compute both summations using the known results $\sum_{i=1}^{y} x^{ai+b} = x^{a+b} \cdot (x^{ay}-1)/(x^a - 1)$ and $\sum_{i=1}^{y} x^{b-ai} = x^{b-ay} \cdot (1-x^{ay})/(1-x^a)$ for $a, b, x > 0$ and $y \geq 1$. This yields

$$\begin{aligned}
&\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i} \\
&\qquad \leq \sqrt{(em)^2 \cdot \frac{(em)^{4\lfloor z/f \rfloor}-1}{(em)^4-1} \cdot \left(1-e^{-(z-f)}\right)^{2n-2\lfloor z/f \rfloor} \cdot \frac{1-\left(1-e^{-(z-f)}\right)^{2\lfloor z/f \rfloor}}{1-\left(1-e^{-(z-f)}\right)^2}} \\
&\qquad \leq \sqrt{(em)^2 \cdot \frac{(em)^{4z/f}-1}{(em)^4-1} \cdot \left(1-e^{-(z-f)}\right)^{2n-2z/f} \cdot \frac{1-\left(1-e^{-(z-f)}\right)^{2z/f}}{1-\left(1-e^{-(z-f)}\right)^2}},
\end{aligned}$$

since $(1-e^{-(z-f)}) \leq 1$. By using this inequality again, it follows that $1-\left(1-e^{-(z-f)}\right)^{2z/f} \leq 1$. Using this observation, we obtain

$$\begin{aligned}
\sum_{i=1}^{\lfloor z/f \rfloor} \binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(z-if)}\right)^{n-i} &\leq \sqrt{(em)^2 \cdot \frac{(em)^{4z/f}-1}{(em)^4-1} \cdot \frac{\left(1-e^{-(z-f)}\right)^{2n-2z/f}}{1-\left(1-e^{-(z-f)}\right)^2}} \\
&= \sqrt{(em)^2 \cdot \frac{(em)^{4z/f}-1}{(em)^4-1} \cdot \frac{\left(1-e^{-(z-f)}\right)^{2n-2z/f}}{e^{-(z-f)}\left(2-e^{-(z-f)}\right)}} \\
&\leq \sqrt{(em)^2 \cdot \frac{(em)^{4z/f}}{(em)^4-1} \cdot \frac{\left(1-e^{-(z-f)}\right)^{2n-2z/f}}{e^{-(z-f)}}},
\end{aligned}$$

where we used the inequalities $0 \leq (em)^{4z/f} - 1 \leq (em)^{4z/f}$ and $2 - e^{-(z-f)} \geq 1$ in the last step. □

Next, we give two rather technical bounds, which will be used to bound the approximation ratio of ONE in case OPT is small, weighted with the probability that OPT is small.

Lemma 6.8. *Let $f \leq \sqrt{n-1}$ for n sufficiently large. Furthermore, set $m := 2n-1$ to shorten notation. Then, for n sufficiently large, we have*

$$f + \int_f^\infty \sqrt{1-\left(1-e^{-(x-f)}\right)^{n-1}}\, dx \leq 2\sqrt{em}.$$

Proof. Let n be sufficiently large. We start by applying the change of variables $y = x - f$ to the integral. This yields

$$f + \int_f^\infty \sqrt{1-\left(1-e^{-(x-f)}\right)^{n-1}}\, dx = f + \int_0^\infty \sqrt{1-(1-e^{-y})^{n-1}}\, dy.$$

Next we use Bernoulli's inequality to obtain that

$$\begin{aligned}
f + \int_f^\infty \sqrt{1-\left(1-e^{-(x-f)}\right)^{n-1}}\, dx &\leq f + \int_0^\infty \sqrt{1-(1-(n-1)e^{-y})}\, dy \\
&= f + \int_0^\infty \sqrt{(n-1)e^{-y}}\, dy \\
&= f + \sqrt{n-1} \cdot \int_0^\infty e^{-\frac{1}{2}y}\, dy \\
&= f + 2\sqrt{n-1}.
\end{aligned}$$

Now we use the fact that $f \le \sqrt{n-1}$ (since n is sufficiently large) and the inequality $3 < 2\sqrt{e}$ to obtain the final result:

$$\begin{aligned} f + \int_f^\infty \sqrt{1 - \left(1 - e^{-(x-f)}\right)^{n-1}}\, dx &\le 3\sqrt{n-1} \\ &\le 3\sqrt{2n-1} \\ &\le 2\sqrt{e(2n-1)}. \end{aligned}$$

Setting $m = 2n - 1$ finishes the proof of this lemma. □

Lemma 6.9. *Let ε be a small constant such that $0 < \varepsilon < 1/2$ and let f and c satisfy one of the following conditions:*

(i) $1/n^{1-\varepsilon} \le f \le 1/n^{\varepsilon}$ for n sufficiently large, and $c \in (0, \varepsilon)$,

(ii) $1/n^{\varepsilon} \le f \le \varepsilon \ln(n)$ for n sufficiently large, and $c \in (\varepsilon, 1-\varepsilon)$,

(iii) $f = \varphi(n)\ln(n)$ for some function $\varphi(n)$ with $\varphi^* := \lim_{n\to\infty} \varphi(n) \in (0, \infty)$, and $c \in (\varphi^*, \varphi^* + 1)$.

Then, for n sufficiently large, we have

$$\left(2en^2\right)^{c\ln(n)/f} \le \frac{1}{n^2} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}c\ln(n)/f - \frac{1}{2}n}.$$

Proof. By looking at dominant terms, it follows for sufficiently large n that

$$\frac{c\ln(n)}{f} \cdot \ln\left(2en^2\right) \le -2\ln(n) + \left(\tfrac{1}{2}n - \frac{\frac{1}{2}c\ln(n)}{f}\right) \cdot \frac{e^f}{n^c}.$$

If f and c satisfy condition (i), then the dominant term on the left hand side is bounded from above by $n^{1-\varepsilon}\ln(n)^2$, whereas the dominant term on the right hand side is bounded from below by n^{1-c}, which becomes significantly larger as $n \to \infty$ since $c < \varepsilon$.
If f and c satisfy condition (ii), then the dominant term on the left hand side is bounded from above by $n^{\varepsilon}\ln(n)^2$, whereas the dominant term on the right hand side is bounded from below by n^{1-c}, which becomes significantly larger as $n \to \infty$ since $c < 1 - \varepsilon$.
If f and c satisfy condition (iii), then the dominant term on the left hand side is bounded from above by $\ln(n)$, whereas the dominant term on the right hand side is bounded from below by $n^{1+\varphi^*-c}$, which becomes significantly larger as $n \to \infty$ since $c < \varphi^* + 1$.
Now, rewriting the right hand side and then applying the well-known inequality $1 - x \le -\ln(x)$ for $x \ge 0$ yields

$$\begin{aligned} \frac{c\ln(n)}{f} \cdot \ln\left(2en^2\right) &\le -2\ln(n) + \left(\tfrac{1}{2}n - \frac{\frac{1}{2}c\ln(n)}{f}\right) \cdot \left(1 - \left(1 - e^{-(c\ln(n)-f)}\right)\right) \\ &\le -2\ln(n) + \left(\tfrac{1}{2}n - \frac{\frac{1}{2}c\ln(n)}{f}\right) \cdot -\ln\left(1 - e^{-(c\ln(n)-f)}\right) \\ &= -2\ln(n) + \left(\frac{\frac{1}{2}c\ln(n)}{f} - \tfrac{1}{2}n\right) \cdot \ln\left(1 - e^{-(c\ln(n)-f)}\right). \end{aligned}$$

Since e^x is an increasing function of x, we can now deduce that

$$e^{(c\ln(n)/f)\cdot\ln(2en^2)} \le e^{-2\ln(n) + \left(\frac{1}{2}c\ln(n)/f - \frac{1}{2}n\right)\cdot\ln\left(1-e^{-(c\ln(n)-f)}\right)},$$

or, equivalently,

$$\left(2en^2\right)^{c\ln(n)/f} \leq \frac{1}{n^2} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}c\ln(n)/f - \frac{1}{2}n},$$

which shows our claim. $\square$

Lastly, we need to bound the approximation ratio of ONE in case OPT is small, weighted with the probability that OPT is small. The following lemma shows that this can be bounded by $O(1/n)$.

Lemma 6.10. *Let ε be a small constant such that $0 < \varepsilon < 1/2$ and let f and c satisfy condition (i), (ii) or (iii) from Lemma 6.9. Then we have*

$$\mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] = O\left(\frac{1}{n}\right).$$

Proof. Observe that conditions (*i*), (*ii*) and (*iii*) from Lemma 6.9 all imply that $c\ln(n) > f$ for n sufficiently large. Now, since $\mathsf{OPT} > f$ by definition, we may bound and subsequently rewrite the conditional expectation as follows:

$$\begin{aligned}
&\mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad\leq \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{f} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad= \frac{1}{f} \cdot \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\mathsf{ONE} \mid \mathsf{OPT} < c\ln(n)\right] \\
&\quad= \frac{1}{f} \cdot \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\int_0^\infty \mathbb{P}\left(\mathsf{ONE} > x \mid \mathsf{OPT} < c\ln(n)\right)\,dx \\
&\quad= \frac{1}{f}\int_0^\infty \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{P}\left(\mathsf{ONE} > x \mid \mathsf{OPT} < c\ln(n)\right)\,dx \\
&\quad= \frac{1}{f}\int_0^\infty \mathbb{P}\left(\mathsf{ONE} > x \text{ and } \mathsf{OPT} < c\ln(n)\right)\,dx.
\end{aligned}$$

Since the events $\mathsf{ONE} > x$ and $\mathsf{OPT} < c\ln(n)$ are dependent, we use Lemma 3.3 to bound the probabilities inside the integral. This results in the following:

$$\begin{aligned}
&\mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad\leq \frac{1}{f}\int_0^\infty \mathbb{P}(\mathsf{ONE} > x)\mathbb{P}(\mathsf{OPT} < c\ln(n)) + \sqrt{\mathbb{P}(\mathsf{ONE} > x)\mathbb{P}(\mathsf{OPT} < c\ln(n))}\,dx \\
&\quad\leq \frac{2}{f}\int_0^\infty \sqrt{\mathbb{P}(\mathsf{ONE} > x)\mathbb{P}(\mathsf{OPT} < c\ln(n))}\,dx \\
&\quad= \frac{2}{f} \cdot \sqrt{\mathbb{P}(\mathsf{OPT} < c\ln(n))}\int_0^\infty \sqrt{\mathbb{P}(\mathsf{ONE} > x)}\,dx,
\end{aligned}$$

where we used the property that $y \leq \sqrt{y}$ for any $y \in [0,1]$. Now recall that we know the probability distribution of ONE (see equation (4.3)). So, it follows that $\mathbb{P}(\mathsf{ONE} > x) = 1$ for $x < f$ and $\mathbb{P}(\mathsf{ONE} > x) = 1 - (1 - e^{-(x-f)})^{n-1}$ for $x \geq f$.

Using this, it follows that:

$$\begin{aligned}
&\mathbb{P}(\mathsf{OPT} < c\ln(n))\, \mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad \le \frac{2}{f} \cdot \sqrt{\mathbb{P}(\mathsf{OPT} < c\ln(n))} \cdot \left(f + \int_f^\infty \sqrt{1 - \left(1 - e^{-(x-f)}\right)^{n-1}}\, dx\right) \\
&\quad \le \frac{2}{f} \cdot \sqrt{\mathbb{P}(\mathsf{OPT} < c\ln(n))} \cdot 2\sqrt{em},
\end{aligned}$$

where we used the bound in Lemma 6.8 for the last inequality (and set $m := 2n - 1$ to shorten notation). Now, for n sufficiently large, we can use Lemma 5.3 (with $z = c\ln(n)$) to bound the last remaining probability. This results in:

$$\begin{aligned}
&\mathbb{P}(\mathsf{OPT} < c\ln(n))\, \mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad \le \frac{4}{f} \cdot \sqrt{\sum_{i=1}^{\lfloor \frac{c\ln(n)}{f} \rfloor} \binom{n}{i}\binom{n-1}{i-1} \left(1 - e^{-(c\ln(n) - if)}\right)^{n-i}} \cdot \sqrt{em}.
\end{aligned}$$

Using the bound in Lemma 6.7 (again with $z = c\ln(n)$ and still using $m := 2n - 1$ to shorten notation), we obtain, for n sufficiently large, that

$$\begin{aligned}
&\mathbb{P}(\mathsf{OPT} < c\ln(n))\, \mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad \le \frac{4}{f} \cdot \sqrt[4]{(em)^2 \cdot \frac{(em)^{4c\ln(n)/f}}{(em)^4 - 1} \cdot \frac{\left(1 - e^{-(c\ln(n)-f)}\right)^{2n - 2c\ln(n)/f}}{e^{-(c\ln(n)-f)}}} \cdot \sqrt{em} \\
&\quad = \frac{4}{f} \cdot \sqrt[4]{\frac{(em)^4}{(em)^4 - 1} \cdot (em)^{4c\ln(n)/f} \cdot \frac{\left(1 - e^{-(c\ln(n)-f)}\right)^{2n - 2c\ln(n)/f}}{e^{-(c\ln(n)-f)}}} \\
&\quad \le \frac{5}{f} \cdot \sqrt[4]{(em)^{4c\ln(n)/f} \cdot \frac{\left(1 - e^{-(c\ln(n)-f)}\right)^{2n - 2c\ln(n)/f}}{e^{-(c\ln(n)-f)}}},
\end{aligned}$$

where we used the inequalities $x/(x-1) \le 2$ (for $x \ge 2$) and $4 \cdot \sqrt[4]{2} < 5$ for the last step. We can further rewrite and bound this as follows:

$$\begin{aligned}
&\mathbb{P}(\mathsf{OPT} < c\ln(n))\, \mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\quad \le \frac{5}{f} \cdot (em)^{c\ln(n)/f} \cdot e^{\frac{1}{4}(c\ln(n)-f)} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n - \frac{1}{2}c\ln(n)/f} \\
&\quad \le \frac{5}{f} \cdot (em)^{c\ln(n)/f} \cdot e^{c\ln(n)-f} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n - \frac{1}{2}c\ln(n)/f} \\
&\quad = \frac{5}{fe^f} \cdot (em)^{c\ln(n)/f} \cdot n^c \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n - \frac{1}{2}c\ln(n)/f} \\
&\quad \le \frac{5}{fe^f} \cdot (em)^{c\ln(n)/f} \cdot n^{c\ln(n)/f} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n - \frac{1}{2}c\ln(n)/f} \\
&\quad \le \frac{5}{fe^f} \cdot (2en^2)^{c\ln(n)/f} \cdot \left(1 - e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n - \frac{1}{2}c\ln(n)/f},
\end{aligned}$$

where we used $m = 2n - 1 \le 2n$ for the last inequality.

Now, by Lemma 6.9, for n sufficiently large, this can be further reduced to

$$\begin{aligned}\mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} < c\ln(n)\right] \\ &\le \frac{5}{fe^f}\cdot\frac{1}{n^2}\cdot\left(1-e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}c\ln(n)/f-\frac{1}{2}n}\cdot\left(1-e^{-(c\ln(n)-f)}\right)^{\frac{1}{2}n-\frac{1}{2}c\ln(n)/f} \\ &= \frac{5}{fe^f}\cdot\frac{1}{n^2}\cdot 1 \\ &= O\left(\frac{1}{n}\right),\end{aligned}$$

since $f \in \Omega(1/n)$ by our restrictions. □

Now we can finally state and prove the main result from this subsection.

Theorem 6.11. *Let $\delta > 0$ be sufficiently small. Consider the facility location problem with $1/n^{1-\varepsilon} \le f \le M\ln(n)$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < \delta/(2+\delta)$ and assume that $\delta < \min\{2, 1/\varphi^*\}$ where $\varphi^* := \lim_{n\to\infty} f/\ln(n)$. Furthermore let* ONE *denote the total cost of the solution in which a facility is opened at exactly one arbitrary vertex $v \in V$ and let* OPT *denote the total cost of the optimal solution to the problem. Then*

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] = O(1).$$

Moreover, if $1/n^{\varepsilon} \le f \le M\ln(n)$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < \delta/(2+\delta)$, then

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] = 1 + \delta + O\left(\frac{1}{\ln(n)}\right).$$

Proof. By equation (4.2) we know that the expected value of ONE is given by

$$\mathbb{E}\left[\mathsf{ONE}\right] = f + \ln(n) + \Theta(1).$$

Now take a constant c as follows:

$$c = \begin{cases} \dfrac{\varepsilon}{2} & \text{if } 1/n^{1-\varepsilon} \le f < 1/n^{\varepsilon} \text{ for } n \text{ sufficiently large,} \\[2ex] \dfrac{1+\varepsilon}{1+\delta} & \text{if } 1/n^{\varepsilon} \le f < \varepsilon\ln(n) \text{ for } n \text{ sufficiently large,} \\[2ex] \dfrac{1+\varphi^*+\varepsilon}{1+\delta} & \text{if } \varepsilon\ln(n) \le f \le M\ln(n) \text{ for } n \text{ sufficiently large.} \end{cases}$$

In the first case we have $c \in (0, \varepsilon)$ and thus do f and c satisfy condition (i) from Lemma 6.9 in that case. In the second case we have $c \in (\varepsilon, 1-\varepsilon)$ (since $\varepsilon < \delta/(2+\delta) < 1/2$ and $\delta < 2$) and thus do f and c satisfy condition (ii) from Lemma 6.9 in that case. In the third case we have $c \in (\varphi^*, 1+\varphi^*)$ (since $0 < \varepsilon < \delta < 1/\varphi^* < (1+\varepsilon)/\varphi$) and thus do f and c satisfy condition (iii) from Lemma 6.9 in that case.

Now, conditioning on the events $\mathsf{OPT} \geq c\ln(n)$ and $\mathsf{OPT} < c\ln(n)$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] &= \mathbb{P}\left(\mathsf{OPT} \geq c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} \geq c\ln(n)\right] \\
&\qquad + \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&\leq \mathbb{E}\left[\frac{\mathsf{ONE}}{c\ln(n)}\right] + \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&= \frac{f + \ln(n) + \Theta(1)}{c\ln(n)} + \mathbb{P}\left(\mathsf{OPT} < c\ln(n)\right)\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}} \,\middle|\, \mathsf{OPT} < c\ln(n)\right] \\
&= \frac{f + \ln(n) + \Theta(1)}{c\ln(n)} + O\left(\frac{1}{n}\right) \\
&= \frac{f + \ln(n)}{c\ln(n)} + O\left(\frac{1}{\ln(n)}\right),
\end{aligned}$$

where we used the result of Lemma 6.10 for the last equality. Since $f = O(\ln(n))$ by our restrictions, we can now deduce that

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] \leq \frac{O(\ln(n))}{c\ln(n)} + O\left(\frac{1}{\ln(n)}\right) = O(1).$$

Moreover, if $1/n^{\varepsilon} \leq f < \varepsilon\ln(n)$ for n sufficiently large, then we have $c = (1+\varepsilon)/(1+\delta)$, and it follows that

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] \leq \frac{\varepsilon\ln(n) + \ln(n)}{c\ln(n)} + O\left(\frac{1}{\ln(n)}\right) = 1 + \delta + O\left(\frac{1}{\ln(n)}\right),$$

and if $\varepsilon\ln(n) \leq f \leq M\ln(n)$ for n sufficiently large, then we have $c = (1+\varphi^* + \varepsilon)/(1+\delta)$; furthermore we know in this last case that $f \leq (\varphi^* + \varepsilon)\ln(n)$ for n sufficiently large, since $\lim_{n\to\infty} f/\ln(n) = \varphi^*$; thus it follows that

$$\mathbb{E}\left[\frac{\mathsf{ONE}}{\mathsf{OPT}}\right] \leq \frac{(\varphi^* + \varepsilon)\ln(n) + \ln(n)}{c\ln(n)} + O\left(\frac{1}{\ln(n)}\right) = 1 + \delta + O\left(\frac{1}{\ln(n)}\right),$$

which shows our claims. □

6.3 Opening some facilities

In this subsection we examine the approximation ratio of the trivial heuristic which opens a facility at exactly $\lceil 1/f \rceil$ arbitrary vertices $v_1, \ldots, v_{\lceil 1/f \rceil} \in V$. We show that SOME yields an $O(1)$ approximation for $1/n < f < 1$ and a $1 + o(1)$ approximation when $f = \alpha(n)/n$ for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$ and $\alpha(n) < n$ (Theorem 6.16).

The idea behind the proof of this theorem is to bound the product of the probability of OPT being smaller than a well-chosen threshold and the conditional expected approximation ratio in that case. First, we bound this expression using Lemma 3.4 and subsequently bound the remaining probabilities using Lemmas 4.4 and 5.3 in order to obtain an explicit function. The final step is to show that this explicit function (which is a bound for the expected approximation ratio) is bounded either by $O(1)$ or $1 + o(1)$ (depending on the asymptotic value of f). Lemmas 6.12, 6.13, 6.14 and 6.15 provide us with the necessary tools in order to be able to do so.

Now, first we need to show that the function $g(x)$, as defined in the following lemma, is negative in an open interval around the constant α^*.

Lemma 6.12. *Let $\alpha^* \in (1, \infty)$ and define $c = \ln(11/10)/\ln(\alpha^*) = \log_{\alpha^*}(11/10)$ and*

$$g(x) := \frac{2c\ln(x)}{x} \ln\left(\frac{ex}{c\ln(x)}\right) - \frac{x - c\ln(x)}{x^{c+1}}.$$

Then we have $g(\alpha^*) < -1/10$, and moreover there exists an $\varepsilon' > 0$ such that $g(x) \leq -1/10$ for all $x \in (\alpha^* - \varepsilon', \alpha^* + \varepsilon')$.

Proof. Using the definition of c, it follows that

$$\begin{aligned} g(\alpha^*) &= \frac{2\ln(11/10)}{\alpha^*} \ln\left(\frac{e\alpha^*}{\ln(11/10)}\right) - \frac{\alpha^* - \ln(11/10)}{\alpha^* \cdot 11/10} \\ &= \frac{1}{\alpha^*} \cdot \left(2\ln\left(\tfrac{11}{10}\right)\left(1 - \ln\left(\ln\left(\tfrac{11}{10}\right)\right) + \ln(\alpha^*)\right) - \tfrac{10}{11}\alpha^* + \tfrac{10}{11}\ln\left(\tfrac{11}{10}\right)\right). \end{aligned}$$

Now, since we have

$$\frac{d}{d\alpha^*}\left(g(\alpha^*)\right) = \frac{-2}{11(\alpha^*)^2} \cdot \ln\left(\tfrac{11}{10}\right)\left(11\ln(\alpha^*) + 5 - 11\ln\left(\ln\left(\tfrac{11}{10}\right)\right)\right) < 0$$

for all $\alpha^* \in (1, \infty)$ (since $5 - 11\ln(\ln(\frac{11}{10})) > 0$), it follows that $g(\alpha^*) < g(1) \approx -0.18$ for all $\alpha^* \in (1, \infty)$. So, we have indeed $g(\alpha^*) < -1/10$.
Moreover, since $g(x)$ is a continuous function, it follows by elementary calculus that there must be an open interval around α^* in which $g(x) \leq -1/10$. In other words, there must exist an $\varepsilon' > 0$ such that $g(x) \leq -1/10$ for all $x \in (\alpha^* - \varepsilon', \alpha^* + \varepsilon')$. ☐

Next, we need to show that the expression $2i\ln(en/i) - (n - c\ln(nf)/f) \cdot e^{if}(nf)^{-c}$ is bounded by a logarithmic function, which will ultimately enable us to bound the probability that OPT is small. The following two lemmas provide this bound (Lemma 6.13 for the case where $\alpha^* := \lim_{n\to\infty} nf < \infty$ and Lemma 6.14 for the case where $\alpha^* = \infty$).

Lemma 6.13. *Let f satisfy $(1+\varepsilon)/n \leq f \leq M/n^{\varepsilon}$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$ and assume that $\alpha^* := \lim_{n\to\infty} nf < \infty$. Define $c = \ln(11/10)/\ln(\alpha^*)$. Then, for n sufficiently large and for any i with $1 \leq i \leq c\ln(nf)/f$, we have*

$$2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} \leq -5\ln(n).$$

Proof. Let n be sufficiently large. Since $i\ln(en/i)$ is an increasing function of i whenever $0 < i < n$, it follows that

$$\begin{aligned} 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} &\leq \frac{2c\ln(nf)}{f} \ln\left(\frac{enf}{c\ln(nf)}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} \\ &= \frac{2c\ln(nf)}{f} \ln\left(\frac{enf}{c\ln(nf)}\right) - \left(1 - \frac{c\ln(nf)}{nf}\right) \cdot n \cdot \frac{e^{if}}{(nf)^c} \end{aligned}$$

Moreover, since $e^{if} \geq 1$ for all $i \geq 1$, we have

$$\begin{aligned} 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} &\leq \frac{2c\ln(nf)}{f} \ln\left(\frac{enf}{c\ln(nf)}\right) - \frac{nf - c\ln(nf)}{(nf)^{c+1}} \\ &= n \cdot \left(\frac{2c\ln(nf)}{nf} \ln\left(\frac{enf}{c\ln(nf)}\right) - \frac{nf - c\ln(nf)}{(nf)^{c+1}}\right). \end{aligned}$$

Now, by Lemma 6.12 we know that there exists an $\varepsilon' > 0$ such that $(2c\ln(x)/x) \cdot \ln(ex/c\ln(x)) - (x - c\ln(x))/x^{c+1} \leq -1/10$ for $c = \ln(11/10)/\ln(\alpha^*)$ and all $x \in (\alpha^* - \varepsilon', \alpha^* + \varepsilon')$. Moreover, since $\lim_{n\to\infty} nf = \alpha^*$, we know that $nf \in (\alpha^* - \varepsilon', \alpha^* + \varepsilon')$ for n sufficiently large. Therefore, we can conclude for n sufficiently large that

$$\begin{aligned} 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} &\leq n \cdot -\frac{1}{10} \\ &\leq -5\ln(n), \end{aligned}$$

where the last inequality holds since n is sufficiently large. □

Lemma 6.14. *Let f satisfy $(1+\varepsilon)/n \leq f \leq M/n^{\varepsilon}$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$ and assume that $\alpha^* := \lim_{n\to\infty} nf = \infty$. Let $c \in (0,1)$ be arbitrary. Then, for n sufficiently large and for any i with $1 \leq i \leq c\ln(nf)/f$, we have*

$$2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} \leq -5\ln(n).$$

Proof. Let n be sufficiently large. Using the same technique as in the proof of Lemma 6.13, we obtain that

$$\begin{aligned} 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} &\leq n \cdot \left(\frac{2c\ln(nf)}{nf}\ln\left(\frac{enf}{c\ln(nf)}\right) - \frac{nf - c\ln(nf)}{(nf)^{c+1}}\right) \\ &= \frac{n}{nf} \cdot \left(2c\ln(nf)\ln\left(\frac{enf}{c\ln(nf)}\right) - \frac{nf - c\ln(nf)}{(nf)^c}\right). \end{aligned}$$

Now, define $\alpha(n) = nf$ and note that $\lim_{n\to\infty} \alpha(n) = \infty$ by our assumptions, whereas $\lim_{n\to\infty} \alpha(n)/n^{1-\varepsilon/2} = 0$ by the restrictions on f. Substituting $nf = \alpha(n)$ results in:

$$2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} \leq \frac{n}{\alpha(n)} \cdot \left(2c\ln(\alpha(n))\ln\left(\frac{e\alpha(n)}{c\ln(\alpha(n))}\right) - (\alpha(n))^{1-c} + \frac{c\ln(\alpha(n))}{(\alpha(n))^c}\right).$$

Observe that the dominant term between the brackets on the right hand side is given by $-(\alpha(n))^{1-c}$ (since $c \in (0,1)$), implying that this factor becomes less than -1 whenever n is sufficiently large. Furthermore, observe that $n/\alpha(n) \geq n^{\varepsilon/2}$ when n is sufficiently large (since $\lim_{n\to\infty} \alpha(n)/n^{1-\varepsilon/2} = 0$). Combining these two observations, we obtain

$$\begin{aligned} 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right) \cdot \frac{e^{if}}{(nf)^c} &\leq \frac{n}{\alpha(n)} \cdot -1 \\ &\leq n^{\varepsilon/2} \cdot -1 \\ &\leq -5\ln(n), \end{aligned}$$

where the last inequality holds since n is sufficiently large. □

Lastly, we use the result from the previous two lemmas to show that each term of the summation that arises when using lemma 5.3 becomes sufficiently small when n gets large.

Lemma 6.15. *Let f satisfy $(1+\varepsilon)/n \leq f \leq M/n^{\varepsilon}$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$. Define $\alpha^* = \lim_{n\to\infty} nf$. If $\alpha^* < \infty$ define $c = \ln(11/10)/\ln(\alpha^*)$; otherwise let $c \in (0,1)$ be arbitrary. Then, for n sufficiently large and for any i with $1 \leq i \leq c\ln(nf)/f$, we have*

$$\binom{n}{i}\binom{n-1}{i-1}\left(1 - e^{-(c\ln(nf) - if)}\right)^{n-i} \leq \frac{1}{n^5}.$$

Proof. First note that the left hand side becomes zero for $i = c\ln(nf)/f$. So we may assume without loss of generality that $1 \le i < c\ln(nf)/f$. Now, since $\ln(1-x) \le -x$ for all $0 \le x < 1$, it follows that

$$2i\ln\left(\frac{en}{i}\right) + \left(n - \frac{c\ln(nf)}{f}\right)\ln\left(1 - e^{-(c\ln(nf)-if)}\right) \le 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right)\cdot e^{-(c\ln(nf)-if)}$$
$$= 2i\ln\left(\frac{en}{i}\right) - \left(n - \frac{c\ln(nf)}{f}\right)\cdot \frac{e^{if}}{(nf)^c}.$$

Using the result of Lemmas 6.13 (for the case $\alpha^* < \infty$) and 6.14 (for the case $\alpha^* = \infty$), it follows now that

$$2i\ln\left(\frac{en}{i}\right) + \left(n - \frac{c\ln(nf)}{f}\right)\ln\left(1 - e^{-(c\ln(nf)-if)}\right) \le -5\ln(n).$$

Since e^x is an increasing function of x, it follows now that

$$e^{2i\ln(en/i)+(n-c\ln(nf)/f)\ln(1-e^{-(c\ln(nf)-if)})} \le e^{-5\ln(n)},$$

or, equivalently,

$$\left(\frac{en}{i}\right)^{2i}\cdot\left(1 - e^{-(c\ln(nf)-if)}\right)^{n-c\ln(nf)/f} \le \frac{1}{n^5}.$$

On the other hand, since $\binom{n-1}{k-1} \le \binom{n}{k}$, $\binom{n_1}{k_1}\binom{n_2}{k_2} \le \binom{n_1+n_2}{k_1+k_2}$ and $\binom{n}{k} \le (en/k)^k$, it follows also that

$$\binom{n}{i}\binom{n-1}{i-1}\left(1 - e^{(c\ln(nf)-if)}\right)^{n-i} \le \binom{2n-1}{2i-1}\left(1 - e^{(c\ln(nf)-if)}\right)^{n-i}$$
$$\le \binom{2n}{2i}\left(1 - e^{(c\ln(nf)-if)}\right)^{n-i}$$
$$\le \left(\frac{en}{i}\right)^{2i}\cdot\left(1 - e^{(c\ln(nf)-if)}\right)^{n-i}.$$

Furthermore, since $1 - e^{(c\ln(nf)-if)} < 1$ and since $n - i \ge n - c\ln(nf)/f$, it follows that

$$\binom{n}{i}\binom{n-1}{i-1}\left(1 - e^{(c\ln(nf)-if)}\right)^{n-i} \le \left(\frac{en}{i}\right)^{2i}\cdot\left(1 - e^{-(c\ln(nf)-if)}\right)^{n-c\ln(nf)/f}.$$

Combining the two results above yields the desired inequality. ☐

Now we can finally state and prove the main result from this subsection.

Theorem 6.16. *Let $\delta > 0$. Consider the facility location problem with $(1+\varepsilon)/n \le f \le M/n^\varepsilon$ for n sufficiently large and some constants $M > 0$ and ε with $0 < \varepsilon < 1$. Furthermore let* SOME *denote the total cost of the solution in which a facility is opened at exactly $\lceil 1/f \rceil$ arbitrary vertices $v_1, \ldots, v_{\lceil 1/f \rceil} \in V$ and let* OPT *denote the total cost of the optimal solution to the problem. Then*

$$\mathbb{E}\left(\frac{\mathsf{SOME}}{\mathsf{OPT}}\right) = O(1).$$

Moreover, if $f = \alpha(n)/n$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$ and $\alpha(n) \le Mn^{1-\varepsilon}$, then

$$\mathbb{E}\left(\frac{\mathsf{SOME}}{\mathsf{OPT}}\right) = 1 + \delta + O\left(\frac{1}{\ln(\alpha(n))}\right).$$

Proof. By equation (4.5) we know that we can bound the expected value of SOME by

$$\mathbb{E}[\mathsf{SOME}] \leq f + \ln(nf) + \Theta(1) = \ln(nf) + \Theta(1),$$

where the equality holds since $f = O(1)$ by our restrictions. Observe that $\alpha^* := \lim_{n\to\infty} nf \geq 1+\varepsilon$ by our restrictions on f. We consider two cases: $\alpha^* < \infty$ and $\alpha^* = \infty$. Note that the second case occurs if and only if $f = \alpha(n)/n$ for some function $\alpha(n)$ with $\lim_{n\to\infty}\alpha(n) = \infty$ and $\alpha(n) \leq Mn^{1-\varepsilon}$ (due to the restrictions on f for n sufficiently large). Now we take a constant c as follows:

$$c = \begin{cases} \dfrac{\ln(11/10)}{\ln(\alpha^*)} & \text{if } \alpha^* < \infty, \\[2ex] \dfrac{1}{1+\delta} & \text{if } \alpha^* = \infty. \end{cases}$$

Conditioning on the events $\mathsf{OPT} \geq c\ln(nf)$ and $\mathsf{OPT} < c\ln(nf)$ yields

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\right] &= \mathbb{P}\left(\mathsf{OPT} \geq c\ln(nf)\right)\cdot\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} \geq c\ln(nf)\right] \\
&\qquad + \mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right)\cdot\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} < c\ln(nf)\right] \\
&\leq \mathbb{E}\left[\frac{\mathsf{SOME}}{c\ln(nf)}\right] + \mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right)\mathbb{E}\cdot\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} < c\ln(nf)\right] \\
&\leq \frac{\ln(nf)+\Theta(1)}{c\ln(nf)} + \mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right)\cdot\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} < c\ln(nf)\right] \\
&= \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right)\cdot\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\,\middle|\,\mathsf{OPT} < c\ln(nf)\right].
\end{aligned}$$

Now we use the result of Lemmas 3.4 (with $X = \mathsf{SOME}$, $Y = \mathsf{OPT}$, $y = c\ln(nf)$ and $\beta = f$) and 4.4 to obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\right] &\leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \frac{1}{f^2}\cdot\mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right) + \int_{1/f^2}^{\infty}\mathbb{P}\left(\mathsf{SOME} \geq \sqrt{x}\right)\,dx \\
&= \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \frac{1}{f^2}\cdot\mathbb{P}\left(\mathsf{OPT} < c\ln(nf)\right) + O\left(\frac{1}{n}\right).
\end{aligned}$$

Moreover, using Lemmas 5.3 (with $z = c\ln(nf)$) and 6.15, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\right] &\leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \frac{1}{f^2}\sum_{i=1}^{\left\lfloor\frac{c\ln(nf)}{f}\right\rfloor}\binom{n}{i}\binom{n-1}{i-1}\left(1-e^{-(c\ln(nf)-if)}\right)^{n-i} + O\left(\frac{1}{n}\right) \\
&\leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \frac{1}{f^2}\sum_{i=1}^{\left\lfloor\frac{c\ln(nf)}{f}\right\rfloor}\frac{1}{n^5} + O\left(\frac{1}{n}\right) \\
&\leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + \frac{c\ln(nf)}{f^3}\cdot\frac{1}{n^5} + O\left(\frac{1}{n}\right) \\
&= \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + O\left(n^4\right)\cdot\frac{1}{n^5} + O\left(\frac{1}{n}\right) \\
&= \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + O\left(\frac{1}{n}\right),
\end{aligned}$$

where we used $\ln(nf) \leq n$ and $1/f \leq n$ for n sufficiently large.

Now we consider our two cases:

Case 1 ($\alpha^* < \infty$): We have $c = \ln(11/10)/\ln(\alpha^*) = \Theta(1)$ and $\lim_{n\to\infty} nf = \alpha^*$, implying $\ln(nf) = \Theta(1)$. Using these observations, it follows immediately that

$$\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\right] \leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + O\left(\frac{1}{n}\right) = O(1) + O\left(\frac{1}{n}\right) = O(1).$$

Case 2 ($\alpha^* = \infty$): We have $c = 1/(1+\delta)$ and $\ln(nf) = \ln(\alpha(n))$ with $\lim_{n\to\infty} \alpha(n) = \infty$ and $\alpha(n) \leq Mn^{1-\varepsilon}$. Using these observations, we can now deduce that

$$\mathbb{E}\left[\frac{\mathsf{SOME}}{\mathsf{OPT}}\right] \leq \frac{1}{c} + \frac{\Theta(1)}{c\ln(nf)} + O\left(\frac{1}{n}\right) = 1 + \delta + O\left(\frac{1}{\ln(\alpha(n))}\right) + O\left(\frac{1}{n}\right) = 1 + \delta + O\left(\frac{1}{\ln(\alpha(n))}\right).$$

This shows our claims. □

7 Overview

In this section we give a brief overview of the results from the previous section. The approximation ratios which are achieved by the trivial heuristics, denoted by ALL, SOME and ONE, are shown in Table 1 for different ranges of facility opening cost f (as $n \to \infty$).

From this table it can be seen that for any facility opening cost f at least one these three trivial heuristics yields either an $O(1)$ approximation or a $1 + o(1)$ approximation as $n \to \infty$. Moreover, whenever $f \notin \Theta(1/n)$, at least one of these trivial heuristics is asymptotically optimal. These results suggest that the facility location problem using random shortest path metrics is relatively difficult for $f \in \Theta(1/n)$.

Note that for $f \approx 1/n$ we have the situation that SOME denotes a solution in which $\lceil 1/f \rceil \approx n$ facilities are opened, i.e. at (almost) every vertex $v \in V$ a facility is being opened, implying that SOME denotes (approximately) the same solution as ALL. In Table 1 it can be seen that both heuristics yield an $O(1)$ approximation in this case.
If $f < 1/n$ (as $n \to \infty$), then it follows that SOME denotes the same solution as ALL (as $n \to \infty$) (where we implicitly make the trivial assumption that SOME does not open more than n facilities). So, basically the obtained results for ALL when $f < 1/n$ (as $n \to \infty$) also hold for SOME. For clarity reasons this is not stated in Table 1.

Furthermore, note that for $f \approx 1$ we have the situation that SOME denotes a solution in which $\lceil 1/f \rceil \approx 1$ facility is opened, implying that SOME denotes (approximately) the same solution as ONE. In Table 1 it can be seen that for $f \in \Theta(1/n^{\varepsilon})$ (with ε an arbitrarily small constant) both heuristics are asymptotically optimal. This makes sense, since we have $\mathbb{E}[\mathsf{SOME}]/\mathbb{E}[\mathsf{ONE}] \approx 1 - \varepsilon$ as $n \to \infty$ when $f \in \Theta(1/n^{\varepsilon})$ (recall equations (4.2) and (4.5)). So, when ε is arbitrarily small, ONE and SOME have (asymptotically) the same expected value.
If $f > 1$ (as $n \to \infty$), then it follows that SOME denotes the same solution as ONE (as $n \to \infty$). So, basically the obtained results for ONE when $f > 1$ (as $n \to \infty$) also hold for SOME. For clarity reasons this is not stated in Table 1.

Facility opening cost	ALL	SOME	ONE
$f = 0$	Optimal	-	-
$0 < f < \varepsilon/n$ †	$1 + o(1)$ (Th. 6.5)	-	-
$\varepsilon/n \leq f < (1+\varepsilon)/n$	$O(1)$ (Th. 6.5)	-	-
$(1+\varepsilon)/n \leq f \leq (2-\varepsilon)/n$	$O(1)$ (Th. 6.5)	$O(1)$ (Th. 6.16)	-
$(2-\varepsilon)/n < f \leq M/n$	-	$O(1)$ (Th. 6.16)	-
$M/n < f < 1/n^{1-\varepsilon}$ ‡	-	$1 + o(1)$ (Th. 6.16)	-
$1/n^{1-\varepsilon} \leq f < 1/n^{\varepsilon}$	-	$1 + o(1)$ (Th. 6.16)	$O(1)$ (Th. 6.11)
$1/n^{\varepsilon} \leq f \leq M/n^{\varepsilon}$	-	$1 + o(1)$ (Th. 6.16)	$1 + o(1)$ (Th. 6.11)
$M/n^{\varepsilon} < f \leq M \ln(n)$	-	-	$1 + o(1)$ (Th. 6.11)
$M \ln(n) < f < \infty$ §	-	-	$1 + o(1)$ (Th. 6.6)
$f = \infty$	-	-	Optimal

Table **1**: Overview of the approximation ratios achieved by the three trivial heuristics for different ranges of the facility opening cost f (as $n \to \infty$). (In this table $\varepsilon > 0$ is an arbitrarily small constant, whereas M is an arbitrarily large constant.)

† $f = 1/(n\alpha(n))$ for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$
‡ $f = \alpha(n)/n$ for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$ and $\alpha(n) < n^{\varepsilon}$
§ $f = \alpha(n)\ln(n)$ for some function $\alpha(n)$ with $\lim_{n\to\infty} \alpha(n) = \infty$

8 Discussion and final remarks

In this thesis we have investigated the probabilistic properties of three rather trivial procedures which give a solution to the facility location problem, when using random shortest path metrics: opening all facilities, opening one arbitrary facility, and opening a certain number of arbitrary facilities (with that certain number only depending on the facility opening cost). Using random shortest path metrics, we have shown that these rather trivial procedures do produce solutions which are surprisingly close to the optimal solutions as the number of vertices gets large.

Table 1 in Section 7 gives an overview of the approximation ratios which are achieved by the three trivial procedures for any possible facility opening cost. As expected, the procedure which opens every facility performs well when the facility opening cost is relatively low, whereas the procedure which opens only one arbitrary facility performs well when the facility opening cost is relatively high. The procedure which opens a certain number of arbitrary facilities fills the gap between the other two procedures.

The results presented in this thesis form only a second step into the research of the behavior of (combinatorial) optimization problems using random shortest path metrics (the first step being the results by Bringmann et al. [1]). Even though random shortest path instances are more difficult to analyze than Euclidean instances or instances with independent random edge lengths, we were able to derive some good results when analyzing the facility location problem on it.
It would be interesting to see whether it is possible to prove similar results when using more sophisticated heuristics for solving the facility location problem. Furthermore, there are many other $\mathcal{NP}$-hard (combinatorial) optimization problems involving metric spaces for which it would be interesting to know how they behave on random short path metrics.

Acknowledgement

I would like to thank my supervisor Bodo Manthey for coming up with the research topic of this thesis and for guiding me through this final project.

References

[1] K. Bringmann, C. Engels, B. Manthey, and B.V.R. Rao. Random shortest paths: Non-euclidean instances for metric optimization problems. *Algorithmica*, 73(1):42–62, 2015. DOI: 10.1007/s00453-014-9901-9.

[2] G. Cornuejols, G.L. Nemhauser, and L.A. Wolsey. The uncapacitated facility location problem. In Pitu B. Mirchandani and Richard L. Francis, editors, *Discrete Location Theory*, chapter 3, pages 119–171. Wiley-Interscience, New York, 1990. ISBN 978-0-471-89233-5.

[3] R. Davis and A. Prieditis. The expected length of a shortest path. *Information Processing Letters*, 46(3):135–141, 1993. DOI: 10.1016/0020-0190(93)90059-I.

[4] A.D. Flaxman, A.M. Frieze, and J.C. Vera. On the average case performance of some greedy approximation algorithms for the uncapacitated facility location problem. *Combinatorics, Probability and Computing*, 16(5):713–732, 2007. DOI: 10.1017/S096354830600798X.

[5] S. Guha and S. Khuller. Greedy strikes back: Improved facility location algorithms. *Journal of Algorithms*, 31(1):228–248, 1999. DOI: 10.1006/jagm.1998.0993.

[6] J.M. Hammersley and D.J.A. Welsh. First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In Jerzy Neyman and Lucien M. Le Cam, editors, *Bernoulli 1713 Bayes 1763 Laplace 1813, Anniversary Volume, Proceedings of an International Research Seminar Statistical Laboratory, University of California, Berkeley 1963*, pages 61–110. Springer Berlin Heidelberg, 1965. DOI: 10.1007/978-3-642-49750-6_7.

[7] R. van der Hofstad, G. Hooghiemstra, and P. Van Mieghem. First-passage percolation on the random graph. *Probability in the Engineering and Informational Science*, 15(2):225–237, 2001. DOI: 10.1017/S026996480115206X.

[8] C.D. Howard. Models of first-passage percolation. In Harry Kesten, editor, *Probability on Discrete Structures*, pages 125–173. Springer Berlin Heidelberg, 2004. DOI: 10.1007/978-3-662-09444-0_3.

[9] K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V.V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing LP. *Journal of the ACM*, 50(6):795–824, 2003. DOI: 10.1145/950620.950621.

[10] S. Janson. One, two and three times log n/n for paths in a complete graph with random weights. *Combinatorics, Probability and Computing*, 8(4):347–361, 1999. DOI: 10.1017/S0963548399003892.

[11] S. Janson. Tail bounds for sums of geometric and exponential variables. 2014. URL http://www2.math.uu.se/~svante/papers/sjN14.pdf.

[12] S. Li. A 1.488 approximation algorithm for the uncapacitated facility location problem. *Information and Computation*, 222:45–58, 2013. DOI: 10.1016/j.ic.2012.01.007.

[13] M. Mahdian, Y. Ye, and J. Zhang. Improved approximation algorithms for metric facility location problems. In Klaus Jansen, Stefano Leonardi, and Vijay Vazirani, editors, *Approximation Algorithms for Combinatorial Optimization, 5th International Workshop, APPROX 2002, Rome, Italy, September 17 - 21, 2002, Proceedings*, pages 229–242. Springer Berlin Heidelberg, 2002. DOI: 10.1007/3-540-45753-4_20.

[14] H.N. Nagaraja. Order statistics from independent exponential random variables and the sum of the top order statistics. In N. Balakrishnan, José María Sarabia, and Enrique Castillo, editors, *Advances in Distribution Theory, Order Statistics, and Inference*, chapter 11, pages 173–185. Birkhäuser Boston, 2006. DOI: 10.1007/0-8176-4487-3_11.

[15] A. Rényi. On the theory of order statistics. *Acta Mathematica Academiae Scientiarum Hungarica*, 4(3-4):191–231, 1953. DOI: 10.1007/BF02127580.

[16] S.M. Ross. *Introduction to Probability Models*. Elsevier, 10th edition, 2010. ISBN 978-0-12-375686-2.

[17] J. Vygen. Approximation algorithms for facility location problems (lecture notes). Technical Report No. 05950, Research Institute for Discrete Mathematics, University of Bonn, 2005. URL http://www.or.uni-bonn.de/~vygen/files/fl.pdf.