Connected Facility Location via Random Facility Sampling and Core Detouring[☆]

Friedrich Eisenbrand[a,*], Fabrizio Grandoni[b,**], Thomas Rothvoß[a,*], Guido Schäfer[c,*]

[a] *Institute of Mathematics, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland.*
[b] *Dipartimento di Informatica, Sistemi e Produzione, Università di Roma Tor Vergata, Via del Politecnico 1, 00133 Roma, Italy.*
[c] *Centrum Wiskunde & Informatica, P.O. Box 94079 1090 GB Amsterdam, The Netherlands.*

Abstract

We present a simple randomized algorithmic framework for connected facility location problems. The basic idea is as follows: We run a black-box approximation algorithm for the unconnected facility location problem, randomly sample the clients, and open the facilities serving sampled clients in the approximate solution. Via a novel analytical tool, which we term *core detouring*, we show that this approach significantly improves over the previously best known approximation ratios for several NP-hard network design problems. For example, we reduce the approximation ratio for the connected facility location problem from 8.55 to 4.00 and for the single-sink rent-or-buy problem from 3.55 to 2.92. The mentioned results can be derandomized at the expense of a slightly worse approximation ratio. The versatility of our framework is demonstrated by devising improved approximation algorithms also for other related problems.

Key words: connected facility location, approximation algorithm, randomized algorithm, network design.

1. Introduction

We consider network design problems that combine facility location and connectivity problems. These problems have a wide range of applications and have recently received considerable attention both in the theoretical computer

[☆] A preliminary version of this paper appeared in SODA'08 [11].

[*] Corresponding author.
[**] Principal corresponding author.
Email addresses: friedrich.eisenbrand@epfl.ch (Friedrich Eisenbrand), grandoni@disp.uniroma2.it (Fabrizio Grandoni), thomas.rothvoss@epfl.ch (Thomas Rothvoß), g.schaefer@cwi.nl (Guido Schäfer)

science literature (see, e.g., [17, 20, 28, 38]) and in the operations research literature (see, e.g., [32, 35]).

As an example (see also [2, 38]), consider the problem of installing a telecommunication network infrastructure. The network consists of a central high-bandwidth *core* with unlimited capacity on the links and individual connections from *endnodes* to nodes in the core. Among the potential core nodes, we need to select a subset that we connect with each other and then route the traffic from each endnode to a core node. Each core node comes with an installation cost and we assume that the cost of installing the high-bandwidth links in the core is larger than the (per unit) routing cost from the endnodes to the core.

We can model the scenario above as a *connected facility location problem* (*CFL*). We are given an undirected graph $G = (V, E)$ with edge costs $c : E \to \mathbb{Q}^+$, a set of facilities $\mathcal{F} \subseteq V$, a set of clients $\mathcal{D} \subseteq V$, and a parameter $M \geq 1$. Every facility $i \in \mathcal{F}$ has an opening cost $f(i) \in \mathbb{Q}^+$ and every client $j \in \mathcal{D}$ has a demand $d(j) \in \mathbb{Q}^+$. The goal is to determine a subset $F \subseteq \mathcal{F}$ of the facilities to be opened, assign each client $j \in \mathcal{D}$ to some open facility $\sigma(j) \in F$ and build a Steiner tree T connecting the open facilities such as to minimize the total cost

$$\sum_{i \in F} f(i) \;+\; M \sum_{e \in T} c(e) \;+\; \sum_{j \in \mathcal{D}} d(j)\, \ell(j, \sigma(j)), \tag{1}$$

where $\ell(v, w)$ is the shortest path distance between vertices $v, w \in V$ in G (with respect to c). We refer to the first, second and last term in (1) as the *opening cost*, *Steiner cost* and *connection cost*, respectively. Subsequently, we assume that every client $j \in \mathcal{D}$ has a unit demand $d(j) = 1$. This assumption is without loss of generality as we may replace j by several copies of co-located unit-demand clients. The algorithms presented in this paper can easily be adapted in order to run in polynomial time even if the original demands are not polynomially bounded in the number n of vertices; we refer the reader to [20] for additional details.

The special case where $\mathcal{F} = V$ and all opening costs are zero is known as the *single-sink rent-or-buy problem* (*SROB*). There are various natural extensions of *CFL* that differ with respect to the underlying facility location and core connectivity problem. For example, in the *connected k-facility location problem* (k-*CFL*) we can open at most k facilities. In the *connected soft-capacitated facility location problem* (*soft-CFL*), facility $i \in \mathcal{F}$ can serve at most $b(i) \in \mathbb{N}$ clients, but we are allowed to open several copies of i (paying its opening cost each time). In both problems, the core constitutes a Steiner tree. We may alternatively consider the variant of *CFL* where the open facilities are connected by a traveling salesman tour. We call the latter problem the *tour-connected facility location problem* (*tour-CFL*).

1.1. Our Results

We present an algorithmic framework to devise simple approximation algorithms for connected facility location problems. Via a novel analytical tool, which we term *core detouring*, we are able to show that this framework yields

approximation algorithms that significantly improve over the previous best approximation ratios for the problems mentioned above. From a high level point of view, our framework works as follows:

1. Compute an approximate solution for the (unconnected) facility location problem.
2. Randomly sample the clients and open the facilities serving sampled clients in the approximate solution.
3. Compute an approximate solution for the connectivity problem on the open facilities and assign clients to the open facilities.

We remark that in Steps 1 and 3, we can use any approximation algorithm for the (unconnected) facility location and core connectivity problem as a black box—this allows us to use the current best approximation algorithms for the respective subproblems.

Our framework yields a 4.00-approximation algorithm for *CFL*. The previous best approximation algorithm for *CFL* is the primal-dual 8.55-approximation algorithm by Swamy and Kumar [37, 38]. In the special case of *SROB*, our algorithm provides a 2.92-approximation, hence improving on the previous best 3.55-approximation algorithm by Gupta et al. [19, 20]. We show that our algorithms for *SROB* and *CFL* can be derandomized using a technique by van Zuylen and Williamson [41]; this way we obtain 4.23 and 3.28 worst-case approximation factors for *CFL* and *SROB*, respectively. We eventually demonstrate the versatility of our framework by applying it to the problems *k-CFL*, *tour-CFL*, and *soft-CFL* for which we improve the current best known approximation ratios.

A key ingredient in our analysis is that we use a novel *core detouring scheme* to bound the expected connection cost of random sampling algorithms. The basic idea is to construct a (possibly sub-optimal) connection scheme and to bound its cost in terms of the optimum cost. In this scheme, we reassign the clients to open facilities by detouring their connection paths through the core in the optimum solution. This construction is set up such that the reassignment is perfectly symmetric, which allows us to bound the expected cost of the detoured paths.

Our method leads to better approximation factors when $|\mathcal{D}|/M$ is large. For this reason, we developed ad-hoc improved approximation algorithms for the case $|\mathcal{D}|/M$ is a constant. For this special case, we designed polynomial-time approximation schemes (PTASs) for *CFL*, *k-CFL*, and *tour-CFL*, and a 2-approximation for *soft-CFL*. This might be of independent, practical interest.

1.2. Previous and Related Work

The network design problems considered here are NP-hard [13] and APX-complete [3, 5, 33], as they contain the Steiner tree problem or the metric traveling salesman problem as a special case. Researchers have therefore concentrated on obtaining good approximation algorithms for them.

CFL and *SROB* have recently received considerable attention in the computer science literature. Gupta et al. [17] obtain a 10.66-approximation algorithm for *CFL*, based on rounding an exponential size LP. Gupta, Srinivasan

and Tardos [22] describe a random facility sampling algorithm for *CFL* leading to a 9.01-approximation. Their algorithm randomly samples clients, and then runs an (unconnected) facility location approximation algorithm on the sampled clients: the corresponding open facilities form the set of open facilities in the final *CFL* solution. We remark that our approach is subtly but substantially different, since we solve an unconnected facility location problem on *all* the clients (not only on the sampled ones), and then randomly select a subset of the resulting (deterministic) pool of open facilities. The best algorithm for *CFL* prior to our work is a primal-dual 8.55-approximation algorithm by Swamy and Kumar [37, 38]. Gupta et al. [20] leave open the question whether a randomized sampling approach can be used to improve the primal-dual approximation algorithm of Swamy and Kumar [37, 38]. In this paper, we answer this question affirmatively.

Better results are known for *SROB*. The first constant approximation is given in [28]. Gupta et al. [17] give a 9.01-approximation algorithm. Swamy and Kumar [37, 38] describe a primal-dual 4.55-approximation algorithm for the same problem. Gupta, Kumar, and Roughgarden [20] propose a simple random sampling algorithm which gives a 3.55-approximation. Gupta, Srinivasan and Tardos [22] show that this algorithm can be derandomized to obtain a 4.2-approximation algorithm. In a recent unpublished work, van Zuylen and Williamson present a derandomization of the same random sampling algorithm that reduces the worst-case approximation factor to 4.

Swamy and Kumar [37, 38] give a 15.55-approximation algorithm for *k-CFL*, which is also the current best. Ravi and Salman [34] consider the special case of *tour-CFL*, where $\mathcal{F} = V$ and all opening costs are zero, and give a 5.83-approximation for it. To the best of our knowledge, *soft-CFL* has not been considered in the literature before, while the corresponding unconnected version is a well-studied problem (see [30] and the references therein).

The results presented in this paper and previous best results are summarized in Table 1. Table 1 refers to the state of the art at the time the conference version of this paper was submitted to SODA'08. In 2008 Hasan, Jung, and Chwa [23] independently found a primal-dual 8.29-approximation algorithm for *CFL* (which is worse than our result). The (unpublished) deterministic 4-approximation by van Zuylen and Williamson is obtained by applying a novel derandomization technique to the algorithm and analysis of Gupta et al. [19, 20]. Combining the same derandomization technique with our core-detouring scheme (as described in an unpublished manuscript that we sent to the authors), van Zuylen and Williamson independently achieved a 3.28-approximation for *SROB*, which now appears in [41]. Later on van Zuylen [40] generalized the derandomization technique in [41]. Using the new technique, the 4.12 randomized algorithm for *tour-CFL* can be turned into a deterministic algorithm with the same approximation guarantee [40].

Random sampling is at the heart of some of the best known approximation algorithms for several basic network design problems: besides *SROB* [20], multi-commodity rent-or-buy (*MROB*) [4, 12, 18], virtual private network design (*VPN*) [8, 9, 10, 20], and single-sink buy-at-bulk (*SSBB*) [15, 20, 27], to

Table 1: Improved approximation ratios obtained in this paper; expected approximation ratios are marked with a star.

Problem	**This paper**	**Previous best**	
CFL	4.00⋆	8.55	Swamy and Kumar [37, 38]
	4.23		
SROB	2.92⋆	3.55⋆	Gupta et al. [19, 20]
	3.28	4	van Zuylen and Williamson [manuscript]
k-CFL	6.85⋆	15.55⋆	Swamy and Kumar [37, 38]
	6.98		
tour-CFL	4.12⋆	5.83⋆	Ravi and Salman [34] (special case only)
soft-CFL	6.27⋆		

name a few. Random sampling is also a useful tool for the solution of facility location problems [31]. The analysis of most of these algorithms is, more or less explicitly, based on *strict cost shares*, a concept originating from game-theoretic cost sharing (see, e.g., the exposition in [19]). These cost shares are used to relate the expected connection cost of the approximate solution to the cost of the core (or cores) in the optimum solution. We remark that our core-detouring scheme provides a different way to analyze random sampling algorithms. In fact, we relate the connection cost of the approximate solution *both* to the optimal core cost *and* to the optimal connection cost.

1.3. Organization of Paper

In Section 2, we study core connection games, which form the basis of our core detouring scheme. Our random facility sampling framework for *CFL* and *SROB* and its analysis are given in Section 3. Refinements of the analysis are presented in Section 4. The derandomization of the algorithm is described in Section 5. In Section 6, we discuss extensions of our framework to other connected facility location problems. Finally, we give some conclusions in Section 7.

1.4. Preliminaries

Throughout this paper, we assume without loss of generality that the number $n = |V|$ of vertices of $G = (V, E)$ satisfies $n \gg 1$. For a given assignment σ of clients to facilities, we let $\sigma^{-1}(i)$ denote the set of clients assigned to facility i. Recall that $\ell(v, w)$ is the shortest path distance between vertices v and w in the graph $G = (V, E)$ with respect to c. We also define $\ell(v, W) = \min_{w \in W} \ell(v, w)$ for a given subset $W \subseteq V$. Finally, we let $c(S) = \sum_{e \in S} c(e)$ denote the total cost of all edges in a subset $S \subseteq E$.

2. Core Connection Games

In this section, we study some random games that we call *core connection games*. These games form the basis of our core detouring scheme introduced in Section 3.

Consider the following setting. We are given a set $\mathcal{N}$ of *core nodes* that are connected by an undirected cycle $\mathcal{C}$, which we call the *core*. Every core node $i \in \mathcal{N}$ has exactly one *client node* $j \in \mathcal{D}$ assigned to it, i.e., $|\mathcal{N}| = |\mathcal{D}|$. We use $\mu(j) \in \mathcal{N}$ to refer to the core node of $j \in \mathcal{D}$. Each client node $j \in \mathcal{D}$ has two oppositely directed edges (j, i) and (i, j) to its respective core node $i = \mu(j)$; see Figure 1. Let $\mathcal{H}_{in}$ be the set of all edges that are directed from client nodes to core nodes and $\mathcal{H}_{out}$ the set of all oppositely directed edges. Define $\mathcal{H} = \mathcal{H}_{in} \cup \mathcal{H}_{out}$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the resulting graph and $w : \mathcal{E} \to \mathbb{Q}^+$ a non-negative weight function on the edges of $\mathcal{G}$. We slightly abuse notation here by using $\mathcal{C} \subseteq \mathcal{E}$ to refer to the set of undirected edges in the cycle. By $w(\mathcal{S})$ we denote the total weight of all edges in $\mathcal{S} \subseteq \mathcal{E}$.

We now consider the following random *cycle-core connection game*: We mark (or sample) one client node uniformly at random and every other client node independently with probability $p \in (0, 1)$. Now, every client node $j \in \mathcal{D}$ sends one unit of (unsplittable) flow to the closest marked client node (with respect to the distances induced by w). We bound the cost of the total flow sent in this game in the following theorem.

Theorem 1. *The cost X of the flow in the cycle-core connection game satisfies $\mathbf{E}[X] \leq w(\mathcal{H}) + w(\mathcal{C})/(2p)$.*

Proof. We bound the cost of the following sub-optimal flow routing scheme: Every client $j \in \mathcal{D}$ sends its flow unit to a closest marked client, with respect to unit edge weights (breaking ties uniformly at random); see Figure 1. The symmetry properties of this routing scheme make it easier to bound its expected cost. Let $f(e)$ be the flow on edge $e \in \mathcal{E}$ and let Y denote the total cost of this flow (with respect to the original weights). Clearly, $\mathbf{E}[X] \leq \mathbf{E}[Y]$.

By linearity of expectation, the cost of this flow is

$$\mathbf{E}[Y] = \sum_{e \in \mathcal{H}} \mathbf{E}[f(e)] \cdot w(e) + \sum_{e \in \mathcal{C}} \mathbf{E}[f(e)] \cdot w(e).$$

Note that $f(e) \leq 1$ holds deterministically for every edge $e \in \mathcal{H}_{in}$. By symmetry reasons, $\mathbf{E}[f(e)] \leq 1$ for all edges $e \in \mathcal{H}_{out}$.

It remains to bound the expected flow on the edges of the cycle. Again exploiting the symmetry of the routing scheme, it is sufficient to consider an arbitrary edge $e \in \mathcal{C}$. Let X_j be the number of edges of the cycle crossed by the flow-path of a given client node j. Clearly,

$$\sum_{e \in \mathcal{C}} f(e) = \sum_{j \in \mathcal{D}} X_j.$$

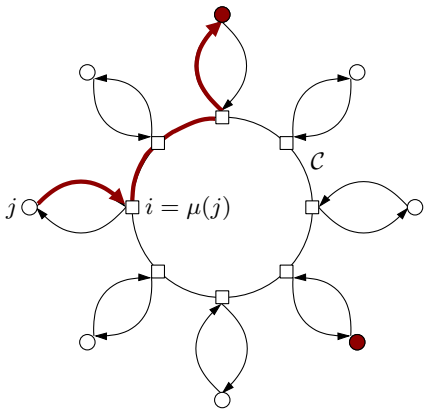

Figure 1: Core connection game instance. Marked client nodes are drawn in bold. The flow of j in the routing scheme is indicated by the bold path.

By symmetry, we can conclude that $\mathbf{E}[f(e)] = \mathbf{E}[X_j]$. Let us call a core node $i = \mu(j)$ *by-sampled* if j is sampled. We now observe that $X_j > k$ if and only if i and the first k nodes of $\mathcal{C}$ to the left and right of i are not by-sampled. As a consequence

$$\Pr(X_j > k) < (1-p)^{2k+1},$$

where the strict inequality is due to the fact that at least one core node is by-sampled by assumption. We conclude that

$$\mathbf{E}[f(e)] = \mathbf{E}[X_j] = \sum_{k \geq 0} \Pr(X_j > k) \leq \frac{1-p}{1-(1-p)^2} \leq \frac{1}{2\,p}.$$

The theorem follows. □

We can modify the cycle-core connection game in a way which is better suited for our purposes. Suppose the core is given by an (undirected) Steiner tree $\mathcal{T}$ on the core nodes in $\mathcal{N}$ instead of a cycle. The tree $\mathcal{T}$ may contain some other non-core nodes. As before, every client node $j \in \mathcal{D}$ is assigned to exactly one core node $\mu(j)$. Let $\mu^{-1}(i)$ be the set of client nodes assigned to a core node $i \in \mathcal{N}$. However, a core node $i \in \mathcal{N}$ might now have more than one client node assigned to it, i.e., we have $|\mu^{-1}(i)| \geq 1$ for every $i \in \mathcal{N}$. The rest of the construction remains the same as before. We define a *tree-core connection game* analogously to the cycle-core connection game.

Theorem 2. *The cost X of the flow in the tree-core connection game satisfies $\mathbf{E}[X] \leq w(\mathcal{H}) + w(\mathcal{T})/p$.*

Proof. We transform the Steiner tree $\mathcal{T}$ into a cycle $\mathcal{C}$ using the following standard arguments: We replace every edge of the tree by two oppositely directed edges and compute a Eulerian tour on the resulting graph. Starting from an

arbitrary core node in $\mathcal{N}$, we traverse this tour and shortcut all nodes that do not belong to $\mathcal{N}$ or have been visited before. Let the resulting cycle on the core nodes $\mathcal{N}$ be $\mathcal{C}'$. By triangle inequality, $w(\mathcal{C}') \leq 2w(\mathcal{T})$.

We now replace every core node i in $\mathcal{C}'$ by a path of $|\mu^{-1}(i)|$ copies of i and assign every client node j in $\mu^{-1}(i)$ to a unique random copy, i.e., compute a random matching between the client nodes and the copies. The weights of the edges in this replacement path are set to zero. Denote the cycle obtained in this way by $\mathcal{C}$. We finally add the two oppositely directed edges between every client node j and its unique copy of $\mu(j)$ in $\mathcal{C}$. Let Y be the cost of the flow in the cycle-core connection game. It is not difficult to see that $X \leq Y$ holds deterministically. The claim now follows from Theorem 1 and the fact that $w(\mathcal{C}) = w(\mathcal{C}') \leq 2w(\mathcal{T})$. □

3. Connected Facility Location

In this section we present our improved approximation algorithms for *CFL* and *SROB*. Let us assume that $M/|\mathcal{D}| \leq \epsilon$, for a sufficiently small constant $\epsilon > 0$. As we will see in Section 3.3, this is without loss of generality since otherwise the problem admits a PTAS.

3.1. Random Facility Sampling

Let $\alpha \in (0, 1]$ be a constant parameter which will be fixed later. Our algorithm `randCFL` for *CFL* works as follows:

1. Compute a ρ_{ft}-approximate solution $U = (F_U, \sigma_U)$ for the (unconnected) facility location instance induced by the input instance.
2. Choose a client $j^* \in \mathcal{D}$ uniformly at random and mark it. Mark every other client j independently with probability α/M. Let D be the set of marked clients.
3. Open facility $i \in F_U$ if there is at least one marked client in $\sigma_U^{-1}(i)$. Let F be the (non-empty) set of open facilities.
4. Compute a ρ_{st}-approximate Steiner tree on D. Augment this tree by adding the shortest path between every $j \in D$ and the corresponding open facility $\sigma_U(j) \in F$. Extract a tree T spanning F from the resulting multi-graph.
5. Output $APX = (F, T, \sigma)$, where σ assigns each client $j \in \mathcal{D}$ to a closest open facility in F.

In Step 4 we might alternatively construct a Steiner tree directly on the open facilities in F; however, this would lead to a worse approximation factor with our analysis. In the special case of *SROB*, we can assume without loss of generality that the facility location approximation algorithm used in Step 1 of `randCFL` opens all the facilities.

The main result of this section is the following theorem; its proof is given in the next subsection.

Theorem 3. *For a proper choice of* α, `randCFL` *is an expected* 4.55*-approximation algorithm for CFL. In the special case of SROB, the approximation ratio reduces to* 3.05*.*

3.2. Analysis

We introduce some more notation. An optimal solution is denoted by $OPT = (F^*, T^*, \sigma^*)$. We use Z^*, O^*, S^* and C^* to refer to its total, opening, Steiner, and connection cost, respectively. Similarly, we use Z, O, S and C to refer to the respective costs of the approximate solution APX computed by `randCFL`. We let O_U and C_U be the opening and connection cost, respectively, of the approximate solution $U = (F_U, \sigma_U)$ for the unconnected instance computed in Step 1.

We first bound the opening cost.

Lemma 1. *The opening cost of APX satisfies* $O \leq O_U$.

Proof. We open a subset of the facilities in F_U, whose total cost is O_U. □

The following bound on the expected Steiner cost is inspired by [20]. We recall that we assume $M/|\mathcal{D}| \leq \epsilon$.

Lemma 2. *The Steiner cost of APX satisfies* $\mathbf{E}[S] \leq \rho_{st}(S^* + (\alpha + \epsilon)\, C^*) + (\alpha + \epsilon)C_U$.

Proof. We obtain a feasible Steiner tree on the marked clients in D by augmenting the optimal Steiner tree T^* by the shortest paths from each client in D to T^*. This Steiner tree has expected cost at most

$$\sum_{e \in T^*} c(e) + \sum_{j \in \mathcal{D}} \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) \ell(j, F^*) = \frac{1}{M}\, S^* + \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) C^*.$$

Thus the expected cost of the ρ_{st}-approximate Steiner tree over D computed in Step 4 is at most

$$\frac{\rho_{st}}{M}\, S^* + \rho_{st} \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) C^*.$$

Additionally, the expected cost of adding the shortest paths from each client $j \in D$ to the corresponding open facility $\sigma_U(j) \in F_U$ is at most

$$\sum_{j \in \mathcal{D}} \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) \ell(j, F_U) = \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) C_U.$$

Altogether we obtain

$$\begin{aligned} \mathbf{E}[S] &\leq M \left(\frac{\rho_{st}}{M} S^* + \rho_{st} \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) C^* + \left(\frac{\alpha}{M} + \frac{1}{|\mathcal{D}|} \right) C_U \right) \\ &\leq \rho_{st}(S^* + (\alpha + \epsilon)\, C^*) + (\alpha + \epsilon) C_U. \end{aligned}$$

□

Core Detouring Scheme. We next introduce our new *core detouring scheme* to bound the expected connection cost of *APX*. Note that since the clients are assigned to their closest open facility in F, it suffices to bound the total cost of connecting every client $j \in \mathcal{D}$ to *some* open facility in F. To this aim, we use the tree-core connection game introduced in Section 2.

We let the tree-core $\mathcal{T}$ in the game be the tree T^* in the optimum solution and set $w(e) = c(e)$ for every edge e in the tree. The client nodes simply correspond to the clients in $\mathcal{D}$. We define the mapping μ as the assignment σ^* of *OPT*. For every client node $j \in \mathcal{D}$, the weight of the directed edge $(j, \mu(j)) \in \mathcal{H}_{in}$ is defined as the connection cost $\ell(j, \sigma^*(j))$; the weight of the directed edge $(\mu(j), j) \in \mathcal{H}_{out}$ is $\ell(\sigma^*(j), j) + \ell(j, \sigma_U(j))$. The sampling probability p is set to $p = \alpha/M$.

The key-insight now is the following: Fix an outcome of the random sampling. For every flow-path from a client node $j \in \mathcal{D}$ to a marked client $j' \in \mathcal{D}$ in $\mathcal{G}$, there is a corresponding path between j and the open facility $\sigma_U(j')$ in the original graph; moreover, the costs of these paths are equal. Thus, for every fixed outcome of the random sampling, the connection cost C is at most the cost X of the flow in the tree-core connection game. Since this holds true for every fixed outcome of the random sampling, it also holds true unconditionally. We can thus bound the expected connection cost by the expected cost of X; for the latter, we derived an upper bound in Section 2. The proof of the following lemma now follows easily.

Lemma 3. *The connection cost of APX satisfies* $\mathbf{E}[C] \leq 2C^* + C_U + S^*/\alpha$.

Proof. Note that the total weight of the tree-core $\mathcal{T}$ is S^*/M. From the discussion above and Theorem 2 it follows that

$$\begin{aligned}\mathbf{E}[C] \leq \mathbf{E}[X] &\leq w(\mathcal{H}) + \frac{1}{p} \cdot w(\mathcal{T}) \\ &= 2\sum_{j \in \mathcal{D}} \ell(j, \sigma^*(j)) + \sum_{j \in \mathcal{D}} \ell(j, \sigma_U(j)) + \frac{M}{\alpha} \cdot \frac{S^*}{M} \\ &= 2C^* + C_U + \frac{S^*}{\alpha}.\end{aligned}$$

□

Now we have all ingredients together to prove Theorem 3. The proof relies on the current best approximation factors for Steiner tree and facility location, which are $\rho_{st} < 1.55$ [36] and $\rho_{fl} < 1.52$ [29], respectively.

Proof. *(Theorem 3)* By Lemmas 1, 2, and 3,

$$\mathbf{E}[Z] \leq O_U + \rho_{st}(S^* + (\alpha + \epsilon)\, C^*) + (\alpha + \epsilon)C_U + 2\, C^* + C_U + \frac{S^*}{\alpha}.$$

The optimum solution to the facility location problem induced by the input instance is a lower bound on $C^* + O^*$. As a consequence, $C_U + O_U \leq \rho_{fl}\,(C^* + O^*)$.

We thus obtain

$$\begin{aligned}\mathbf{E}[Z] &\leq \rho_{st}(S^* + (\alpha+\epsilon)\,C^*) + 2C^* + \frac{S^*}{\alpha} + (1+\alpha+\epsilon)\rho_{fl}\,(C^*+O^*) \\ &\leq (C^*+O^*)(\rho_{st}(\alpha+\epsilon)+2+\rho_{fl}(1+\alpha+\epsilon)) + S^*\left(\rho_{st}+\frac{1}{\alpha}\right).\end{aligned}$$

Choosing ϵ sufficiently small and balancing the coefficients of C^*+O^* and S^*, we obtain the claimed approximation ratio for $\alpha = 0.334$.

Recall that in the special case of *SROB*, we can assume without loss of generality that the facility location approximation algorithm used in Step 1 of `randCFL` opens all the facilities. As a consequence, `randCFL` opens a facility at every marked client. By imposing $O_U = O^* = C_U = 0$ in the analysis above and choosing $\alpha = 0.671$, it follows that `randCFL` is an expected 3.05-approximation algorithm for *SROB*. □

3.3. PTAS for Constant $|\mathcal{D}|/\mathbf{M}$

In this subsection, we present our PTAS for *CFL* in the special case that $|\mathcal{D}|/M$ is upper bounded by a constant, hence justifying the assumption made at the beginning of this section. Besides helping to improve the analysis for the general case, this PTAS might also be of independent interest.

Theorem 4. *If* $|\mathcal{D}|/M = O(1)$*, then there is a PTAS for k-CFL.*

Proof. Let $OPT = (F^*, T^*, \sigma^*)$ be an optimal solution for *k-CFL*. We use Z^*, O^*, S^* and C^* to refer to its total, opening, Steiner, and connection cost, respectively. If k is a constant, we can trivially compute an optimum solution in polynomial time. Hence, let $m \geq 1$ be an arbitrary integral constant and assume $k \geq 2m$. Consider the following algorithm:

1. For all possible choices of $F \subseteq \mathcal{F}$ with $|F| \leq 2m$ do:
   (a) Compute an optimal Steiner tree T over F.
   (b) Assign every client $j \in \mathcal{D}$ to its closest facility $\sigma(j)$ in F.
2. Output a minimum cost solution (F, T, σ), among the solutions obtained.

In Step 1(a), we can use, for example, the algorithm by Dreyfus and Wagner [7]. Note that the algorithm outputs a feasible solution, since $2m \leq k$, and runs in polynomial time.

It is sufficient to show that there is a proper choice of F which satisfies the claim. Let us construct F as follows: Initially, set $F := \{i^*\}$, where i^* is an arbitrary facility in F^*. Then, while there exists a facility $i \in F^*$ with $\ell(i, F) > c(T^*)/m$, add i to F. Note that this way, we ensure that the following two properties hold for the final set F:

1. For any two facilities $i, i' \in F$, $\ell(i, i') > c(T^*)/m$.
2. For every facility $i \in F^*$, there is a facility i' in F such that $\ell(i, i') \leq c(T^*)/m$.

We first show that $|F| \leq 2m$. To see this, double the edges of T^*, compute an Eulerian tour E^* on the resulting graph, and shortcut the vertices not in F. The cost of the resulting tour on F is at least $|F| \cdot c(T^*)/m$ due to Property 1. Moreover, the cost of the Eulerian tour is $c(E^*) \leq 2c(T^*)$. Thus, $|F| \cdot c(T^*)/m \leq 2\, c(T^*)$, which implies that $|F| \leq 2m$.

We next bound the cost Z of the solution $APX = (F, T, \sigma)$ for our particular choice of F. Clearly, $c(T) \leq c(T^*)$, since $F \subseteq F^*$ and we compute an optimum Steiner tree T over F. Therefore,

$$\begin{aligned} Z &= \sum_{i \in F} f(i) + Mc(T) + \sum_{j \in \mathcal{D}} \ell(j, \sigma(j)) \\ &\leq \sum_{i \in F^*} f(i) + Mc(T^*) + \sum_{j \in \mathcal{D}} \ell(j, \sigma^*(j)) + \sum_{j \in \mathcal{D}} \ell(\sigma^*(j), F) \\ &\leq O^* + S^* + C^* + |\mathcal{D}| \cdot \frac{c(T^*)}{m} = Z^* + \frac{|\mathcal{D}|}{M} \cdot \frac{Mc(T^*)}{m} \\ &= Z^* + O(1) \cdot \frac{S^*}{m} \leq \left(1 + \frac{O(1)}{m}\right) Z^*. \end{aligned}$$

For the second inequality, we exploit the fact that $\ell(\sigma^*(j), F) \leq c(T^*)/m$ by Property 2. Since we can choose m arbitrarily large, the claim follows. □

Corollary 1. *If $|\mathcal{D}|/M = O(1)$, then there is a PTAS for CFL.*

Proof. It follows from Theorem 4 observing that *CFL* is equivalent to $|\mathcal{F}|$-*CFL*. □

4. Refinements

We can improve the approximation ratio of `randCFL` given in Section 3 by combining the following techniques.

(a) Bifactor facility location. We obtain a better approximation ratio if we run a (proper) bifactor approximation algorithm on the induced facility location instance in Step 1. An algorithm for the facility location problem is a (ρ_O, ρ_C)-approximation algorithm if for every feasible solution with opening cost O and connection cost C, the cost of the solution computed by the algorithm is at most $\rho_O\, O + \rho_C\, C$. Mahdian, Ye, and Zhang [29] give a $(1.11, 1.78)$-approximation algorithm. Moreover, they (essentially) show that any (ρ_O, ρ_C)-approximation algorithm can be converted into a $(\rho_O + \ln \delta, 1 + (\rho_C - 1)/\delta)$-approximation algorithm for any $\delta \geq 1$.

Note that an optimum solution *OPT* for *CFL* induces a feasible solution for the underlying facility location problem with opening cost O^* and connection cost C^*. Exploiting this, we obtain

$$C_U + O_U \leq (1.11 + \ln \delta) O^* + \left(1 + \frac{0.78}{\delta}\right) C^*.$$

We can now optimize the parameter δ so as to balance the coefficients of the connection and opening costs; while the parameter α is used to balance the Steiner and connection costs.

(b) Flow canceling. We can refine Theorem 2, and hence the bound on the connection cost given in Lemma 3, by means of flow canceling. Consider a given edge e of $\mathcal{T}$ in the tree-core connection game and let e_1 and e_2 be the two edges of $\mathcal{C}$ associated to e (because of shortcutting, it might be $e_1 = e_2$). If the flows along e_1 and e_2 in $\mathcal{C}$ are both directed clockwise or counterclockwise (and $e_1 \neq e_2$), this means that we are sending two oppositely directed flows along e in $\mathcal{T}$. In this case, it is possible to cancel the difference of the two flows (independently for each $e \in \mathcal{T}$) by redirecting the flow paths in a proper way. The somewhat technical proof of the following theorem is given in the appendix.

Theorem 5. *For $|\mathcal{D}| \gg 1/p$, the cost X of the flow in the tree-core connection game satisfies $\mathbf{E}[X] \leq w(\mathcal{H}) + 0.807\, w(\mathcal{T})/p$.*

In particular, since by assumption $|\mathcal{D}|/M \gg 1$ and α is a constant, this implies the following refined bound on the connection cost:

$$\mathbf{E}[C] \leq 2C^* + C_U + 0.807\,\frac{S^*}{\alpha}.$$

Combining Techniques (a) and (b), we obtain the following theorem.

Theorem 6. *There is an expected* 4.00*-approximation algorithm for CFL. In the special case of SROB, the expected approximation ratio reduces to* 2.92*.*

Proof. Let us adapt the proof of Theorem 3. Combining (a) and (b), we obtain

$$\begin{aligned}
\mathbf{E}[Z] &\leq O_U + \rho_{st}(S^* + (\alpha+\epsilon)\,C^*) + (\alpha+\epsilon)\,C_U + 2\,C^* + C_U + 0.807\,\frac{S^*}{\alpha} \\
&\leq \rho_{st}(S^* + (\alpha+\epsilon)\,C^*) + 2\,C^* + 0.807\,\frac{S^*}{\alpha} \\
&\quad + (1+\alpha+\epsilon)\left((1.11 + \ln\delta)O^* + \left(1 + \frac{0.78}{\delta}\right)C^*\right) \\
&= C^*\left(\rho_{st}(\alpha+\epsilon) + 2 + (1+\alpha+\epsilon)\left(1 + \frac{0.78}{\delta}\right)\right) \\
&\quad + S^*\left(\rho_{st} + \frac{0.807}{\alpha}\right) + O^*((1+\alpha+\epsilon)(1.11+\ln\delta)) \\
&\overset{\alpha=0.330,\ \delta=6.657}{<} 4.00\,Z^*.
\end{aligned}$$

The analysis above can be adapted to *SROB* by imposing $C_U = O_U = O^* = 0$. For $\alpha = 0.591$, this yields

$$\mathbf{E}[Z] \leq \rho_{st}(S^* + (\alpha+\epsilon)\,C^*) + 2\,C^* + 0.807\,\frac{S^*}{\alpha} < 2.92\,Z^*.$$

□

5. Derandomization

We can derandomize our algorithm for *CFL* using an idea by van Zuylen and Williamson [41]. In order to use the result by van Zuylen and Williamson as a black box, we adapt our randomized algorithm `randCFL` as follows:

- In Step 2, we first mark each client independently with probability α/M. Let D' be the set of marked clients. Then we sample one client j^* uniformly at random, and let $D = D' \cup \{j^*\}$. Eventually, we guess the facility $r^* = \sigma^*(j^*)$ in the Steiner tree of an optimal solution which is closest to j^*.
- In Step 4, we compute a 2-approximate Steiner tree on $D' \cup \{r^*\}$, using a primal-dual algorithm [1, 14]. Then we augment this tree by adding the shortest path from j^* to r^*, and from each $j \in D$ to the corresponding open facility $\sigma_U(j)$.

The guessing of r^* is implemented by considering all possible choices for $r^* \in \mathcal{F}$. We assume that $M/|\mathcal{D}| \leq \epsilon$; this is without loss of generality since our PTAS for *CFL* is deterministic.

Let $Z = O + S_{st} + S_{aug} + C$ be the expected total cost of the (modified) randomized algorithm, where O is the opening cost, S_{st} the cost of the Steiner tree on $D' \cup \{r^*\}$ computed in Step 4, S_{aug} the augmentation cost in Step 4, and C the connection cost of the modified algorithm. Following essentially the same line of arguments as in the proof of Theorem 6, we obtain for $\alpha = 0.361885$, $\delta = 7.359457$, and ϵ small enough

$$\mathbf{E}[Z] \leq O_U + 2(S^* + (\alpha+\epsilon)C^*) + (\alpha+\epsilon)\, C_U + 2\, C^* + C_U + 0.807\, \frac{S^*}{\alpha} < 4.23\, Z^*. \quad (2)$$

That is, the modified algorithm is a 4.23-approximation algorithm for *CFL*. In the special case of *SROB*, the approximation ratio reduces to 3.28 by letting $C_U = O_U = 0$ and choosing $\alpha = 0.635$.

We next show how to derandomize the algorithm above, without increasing its approximation ratio. For $x \in \mathcal{D}$, $y \in \mathcal{F}$, $A \subseteq \mathcal{D}$ with $x \in A$, and $\bar{A} \subseteq \mathcal{D}$ with $A \cap \bar{A} = \emptyset$, we denote by $(x, y, A, \bar{A})$ the event $\{j^* = x, r^* = y, A \subseteq D, \bar{A} \cap D = \emptyset\}$. Intuitively, $(x, y, A, \bar{A})$ refers to the event that we have sampled client $j^* = x$, guessed facility $r^* = y$, decided to mark the clients in A and to unmark the clients in $\bar{A}$. Suppose we were able to compute the conditional expected cost $\mathbf{E}[Z \mid (x, y, A, \bar{A})]$. We could then run the following (deterministic) algorithm for every possible choice $(x, y) \in \mathcal{D} \times \mathcal{F}$:

1. $D_{x,y} = \{x\}$, $\bar{D}_{x,y} = \emptyset$;
2. While $\exists j \in \mathcal{D} \setminus (D_{x,y} \cup \bar{D}_{x,y})$
   (a) If $\mathbf{E}[Z \mid (x, y, D_{x,y} \cup \{j\}, \bar{D}_{x,y})] \leq \mathbf{E}[Z \mid (x, y, D_{x,y}, \bar{D}_{x,y} \cup \{j\})]$, set $D_{x,y} \leftarrow D_{x,y} \cup \{j\}$.
   (b) Otherwise set $\bar{D}_{x,y} \leftarrow \bar{D}_{x,y} \cup \{j\}$.

We can interpret $D_{x,y}$ and $\bar{D}_{x,y}$ as the sets of clients that we have decided to mark and unmark, respectively. Initially, we mark client x. In each iteration of Step 2, we decide for some (yet undecided) client j whether to mark or unmark it; this process continues until eventually all clients are either marked or unmarked. It is easy to see that, for fixed x and y, at the end of the process we have

$$Z_{x,y} := \mathbf{E}[Z \mid (x, y, D_{x,y}, \bar{D}_{x,y})] \leq \mathbf{E}[Z \mid j^* = x, r^* = y].$$

In fact, the choice of whether to mark or unmark client j in each iteration of Step 2 is made so as to guarantee that the conditional expected cost does not increase. As a consequence, the cost of the cheapest solution obtained by the deterministic algorithm above satisfies

$$\min_{(x,y)\in\mathcal{D}\times\mathcal{F}} Z_{x,y} \leq \min_{(x,y)\in\mathcal{D}\times\mathcal{F}} \mathbf{E}[Z \mid j^* = x, r^* = y] \leq \mathbf{E}[Z],$$

and we thus would obtain the desired approximation ratio.

From the discussion above it follows that we would be able to derandomize our algorithm if we could efficiently compute the conditional expected cost $\mathbf{E}[Z \mid (x, y, A, \bar{A})]$. It is not difficult to see that we can indeed efficiently compute the conditional expected connection cost $\mathbf{E}[C \mid (x, y, A, \bar{A})]$ and opening cost $\mathbf{E}[O \mid (x, y, A, \bar{A})]$ of our algorithm. The same holds for the conditional expected augmentation cost $\mathbf{E}[S_{aug} \mid (x, y, A, \bar{A})]$. However, the problem is that we do not know how to compute the conditional expected Steiner cost $\mathbf{E}[S_{st} \mid (x, y, A, \bar{A})]$.

We circumvent this problem by using an idea of van Zuylen and Williamson [41]. A straightforward adaptation of their analysis (cf. [41, Lemma 2.4]) shows that, when y belongs to the optimal Steiner tree, there is a random variable X satisfying the following properties: (i) $\mathbf{E}[X \mid j^* = x, r^* = y] \leq S^* + \alpha C^*$, (ii) $\mathbf{E}[S_{st} \mid (x, y, A, \bar{A})] \leq 2\mathbf{E}[X \mid (x, y, A, \bar{A})]$, and (iii) $\mathbf{E}[X \mid (x, y, A, \bar{A})]$ can be computed in polynomial time. We remark that in order to prove this claim, the authors crucially exploit the fact that a primal-dual 2-approximate Steiner tree algorithm [1, 14] is used.

Observe that $2X$ is sandwiched between S_{st} and the upper bound $2(S^* + \alpha\, C^*)$ on S_{st} used in the analysis of the randomized algorithm. The idea is then to replace the random variable S_{st} by $2X$. In other words, the deterministic algorithm makes its decisions in Step 2 according to the new cost function $Z' = O + 2X + S_{aug} + C$ (for which we are able to compute $\mathbf{E}[Z' \mid (x, y, A, \bar{A})]$ efficiently by Property (iii)). Eventually, the algorithm still outputs the solution whose Z-cost is minimal among all pairs $(x, y) \in \mathcal{D} \times \mathcal{F}$. In the following, we denote by $Z'_{x,y}$ the Z'-cost of the solution returned for the pair (x, y).

Theorem 7. *There is a deterministic* 4.23*-approximation algorithm for CFL. In the special case of SROB, the approximation ratio reduces to* 3.28*.*

Proof. Consider the algorithm above. Let $E(x) = (x, \sigma^*(x), D_{x,\sigma^*(x)}, \bar{D}_{x,\sigma^*(x)})$. Since $\sigma^*(x)$ belongs to the optimal Steiner tree, the result of van Zuylen and

Williamson applies. In particular,

$$\begin{aligned} Z'_{x,\sigma^*(x)} &= \mathbf{E}[Z' \mid E(x)] \le \mathbf{E}[Z' \mid j^* = x, r^* = \sigma^*(x)] \\ &\le O_U + 2(S^* + \alpha C^*) + M\ell(x, \sigma^*(x)) + M\ell(x, \sigma_U(x)) \\ &\quad + \frac{\alpha}{M} \sum_{v \in \mathcal{D}} M\ell(v, \sigma_U(v)) + \mathbf{E}[C \mid j^* = x], \end{aligned} \tag{3}$$

where we use the fact that the choices of the algorithm are made such that the expected Z'-cost never increases and Property (i). Moreover, by Property (ii),

$$Z_{x,\sigma^*(x)} = \mathbf{E}[Z \mid E(x)] \le \mathbf{E}[Z' \mid E(x)] = Z'_{x,\sigma^*(x)}. \tag{4}$$

Combining (2), (3), and (4), we can conclude that the cost $\tilde{Z}$ of the final solution returned by the algorithm satisfies

$$\begin{aligned} \tilde{Z} &= \min_{(x,y) \in \mathcal{D} \times \mathcal{F}} Z_{x,y} \le \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \min_{y \in \mathcal{F}} Z_{x,y} \le \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} Z_{x,\sigma^*(x)} \\ &\le \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} Z'_{x,\sigma^*(x)} \le O_U + 2(S^* + \alpha C^*) + \frac{M}{|\mathcal{D}|} C^* + \frac{M}{|\mathcal{D}|} C_U + \alpha C_U + \mathbf{E}[C] \\ &\le O_U + 2(S^* + \alpha C^*) + \epsilon C^* + \epsilon C_U + \alpha C_U + 2\,C^* + C_U + 0.807\,\frac{S^*}{\alpha} < 4.23\,Z^*. \end{aligned}$$

A similar argument gives a deterministic 3.28-approximation algorithm in the case of *SROB*. □

6. Extensions

Our approach is flexible enough to be adapted to several natural variants of *CFL*. In this section we sketch three such applications.

6.1. Connected k*-Facility Location*

By Theorem 4, we can assume that $M/|\mathcal{D}| \le \epsilon$, for a sufficiently small constant $\epsilon > 0$. An algorithm for *k-CFL* is obtained by modifying `randCFL` in the following way:

- In Step 1, compute a ρ_{kfl}-approximate solution $U = (F_U, \sigma_U)$ for the (unconnected) k-facility location instance induced by the input instance.

The analysis can be refined via Technique (b). The following theorem relies on the current best approximation ratio for the k-facility location problem, which is $\rho_{kfl} \le 4$ [25, 26] (see also [42]).

Theorem 8. *There is an expected* 6.85*-approximation algorithm for k-CFL.*

Proof. By adapting the proof of Theorem 6, we obtain

$$\begin{aligned}\mathbf{E}[Z] &\leq \rho_{st}(S^* + (\alpha+\epsilon)C^*) + 2C^* + 0.807\,\frac{S^*}{\alpha} + (1+\alpha+\epsilon)\rho_{kfl}\,(C^*+O^*)\\ &\leq (C^*+O^*)(\rho_{st}(\alpha+\epsilon) + 2 + \rho_{kfl}\,(1+\alpha+\epsilon)) + S^*\left(\rho_{st} + \frac{0.807}{\alpha}\right)\\ &\overset{\alpha=0.1524}{<} 6.85\,Z^*.\end{aligned}$$

□

Also in this case the algorithm can be derandomized by applying the technique by van Zuylen and Williamson [41].

Corollary 2. *There is a deterministic* 6.98*-approximation algorithm for k-CFL.*

6.2. Tour-Connected Facility Location

When $|\mathcal{D}|/M$ is small we can easily obtain a PTAS for *tour-CFL* by adapting the analysis of Theorem 4. The main difference is that now T^* and T denote optimal tours (instead of optimal Steiner trees) on F^* and F, respectively. Note that, for $|F| = O(1)$, T can be computed in polynomial time, for example via the algorithm by Held and Karp [24].

Theorem 9. *If $|\mathcal{D}|/M = O(1)$, then there is a PTAS for tour-CFL.*

Due to Theorem 9, we can assume also in this case that $M/|\mathcal{D}| \leq \epsilon$, for a sufficiently small constant $\epsilon > 0$. We obtain an algorithm for *tour-CFL* by adapting `randCFL` in the following way:

- In Step 4, compute a ρ_{tsp}-approximate TSP-tour on D. Then augment the tour by adding *two* shortest paths between every client in D and the corresponding open facility in F. Finally, compute an Eulerian tour on the resulting multi-graph and shortcut it to obtain a TSP-tour T of F.

The algorithm above can be improved by means of Technique (a). The following result relies on Christofides' 1.5-approximation algorithm for metric TSP [6].

Theorem 10. *There is an expected* 4.12*-approximation algorithm for tour-CFL.*

Proof. We adapt the analysis of Section 3. Trivially, $O \leq O_U$. Taking into account the duplication of the shortest paths from D to F and using a similar duplication to bound the cost of the optimum TSP-tour over D, we obtain

$$\mathbf{E}[S] \leq \rho_{tsp}(S^* + 2(\alpha+\epsilon)C^*) + 2(\alpha+\epsilon)\,C_U.$$

Theorem 2 can be easily adapted to this case and we thus obtain

$$\mathbf{E}[X] \leq w(\mathcal{H}) + \frac{w(\mathcal{T})}{2p}.$$

It follows that

$$\mathbf{E}[C] \leq 2C^* + C_U + \frac{S^*}{2\alpha}.$$

Altogether

$$\begin{aligned}
\mathbf{E}[Z] &\leq O_U + \rho_{tsp}(S^* + 2(\alpha+\epsilon)C^*) + 2(\alpha+\epsilon)C_U + 2C^* + C_U + \frac{S^*}{2\alpha} \\
&\leq \rho_{tsp}(S^* + 2(\alpha+\epsilon)C^*) + 2C^* + \frac{S^*}{2\alpha} \\
&\quad + (1 + 2(\alpha+\epsilon))\left(\left(1 + \frac{0.78}{\delta}\right)C^* + (1.11 + \ln\delta)O^*\right) \\
&= C^*\left(2\rho_{tsp}(\alpha+\epsilon) + 2 + (1+2(\alpha+\epsilon))\left(1 + \frac{0.78}{\delta}\right)\right) \\
&\quad + S^*\left(\rho_{tsp} + \frac{1}{2\alpha}\right) + O^*((1+2(\alpha+\epsilon))(1.11+\ln\delta)) \\
&\overset{\alpha=0.19084,\ \delta=6.5004}{\leq} 4.12\, Z^*.
\end{aligned}$$

□

6.3. Connected Soft-Capacitated Facility Location

Let us assume that $M/|\mathcal{D}| \leq \epsilon$, for a constant $\epsilon > 0$ small enough. We will later show how to obtain a 2-approximation for $|\mathcal{D}|/M = O(1)$. Our algorithm for *soft-CFL* is obtained by modifying `randCFL` as follows:

- In Step 1, compute a ρ_{sf}-approximate solution $U = (F_U, \sigma_U)$ for the (unconnected) soft-capacitated instance induced by the input instance.
- In Step 5, output $APX = (F, T, \sigma)$, where σ assigns each client $j \in \mathcal{D}$ to the open facility $i = \sigma(j)$ which minimizes the quantity $\ell(i,j) + f(i)/b(i)$.

The analysis can be refined via Technique (b). The following theorem relies on the current best approximation ratio for the (unconnected) soft-capacitated facility location problem, which is $\rho_{sf} \leq 2$ [30].

Theorem 11. *There is an expected* 6.27*-approximation algorithm for soft-CFL.*

Proof. We adapt the analysis of Section 3. Let $APX = (F, T, \sigma)$ be the solution computed by the algorithm. By essentially the same analysis as in Lemma 2, the Steiner cost S of APX satisfies:

$$\mathbf{E}[S] \leq \rho_{st}(S^* + (\alpha+\epsilon)\, C^*) + (\alpha+\epsilon)\, C_U.$$

In order to bound the connection cost C and opening cost O of APX, we consider the following reduction. Let APX' be the solution to *CFL* corresponding to APX (assignment included). We augment the connection cost of APX' in the following way: for each client j assigned to a facility $i = \sigma(j)$, we increase the

corresponding connection cost from $\ell(i,j)$ to $\ell(i,j) + f(i)/b(i)$. Let C' and O' be the new connection and opening costs of APX' after the augmentation. Then

$$\begin{aligned} O + C &= \sum_{i \in F} \left\lceil \frac{|\sigma^{-1}(i)|}{b(i)} \right\rceil f(i) + \sum_{j \in \mathcal{D}} \ell(j, \sigma(j)) \\ &\leq \sum_{i \in F} f(i) + \sum_{i \in F} \frac{|\sigma^{-1}(i)|}{b(i)} f(i) + \sum_{j \in \mathcal{D}} \ell(j, \sigma(j)) \\ &\leq \sum_{i \in F_U} f(i) + \sum_{j \in \mathcal{D}} \left(\ell(j, \sigma(j)) + \frac{f(\sigma(j))}{b(\sigma(j))} \right) \leq O_U + C'. \end{aligned}$$

We next observe that the assignment σ is chosen such as to minimize the augmented connection cost C'. Hence, in order to bound this cost, we can use the same approach as in Lemma 3. This analysis can be refined by means of Technique (b) (since flow-canceling does not change the number of clients assigned to each facility). Thus

$$\begin{aligned} \mathbf{E}[C'] &\leq 2C^* + 0.807\frac{S^*}{\alpha} + \sum_{j \in \mathcal{D}} \left(\ell(j, \sigma_U(j)) + \frac{f(\sigma_U(j))}{b(\sigma_U(j))} \right) \\ &= 2C^* + 0.807\frac{S^*}{\alpha} + C_U + \sum_{i \in F_U} \frac{|\sigma_U^{-1}(i)|}{b(i)} f(i) \\ &\leq 2C^* + 0.807\frac{S^*}{\alpha} + C_U + \sum_{i \in F_U} \left\lceil \frac{|\sigma_U^{-1}(i)|}{b(i)} \right\rceil f(i) = 2C^* + 0.807\frac{S^*}{\alpha} + C_U + O_U. \end{aligned}$$

Altogether,

$$\begin{aligned} \mathbf{E}[Z] &\leq \rho_{st}(S^* + (\alpha+\epsilon)C^*) + (\alpha+\epsilon)C_U + O_U + 2C^* + 0.807\,\frac{S^*}{\alpha} + C_U + O_U \\ &\leq \rho_{st}(S^* + (\alpha+\epsilon)C^*) + (2+\epsilon)(C_U + O_U) + 2C^* + 0.807\,\frac{S^*}{\alpha} \\ &\leq \rho_{st}(S^* + (\alpha+\epsilon)C^*) + (2+\epsilon)\rho_{sf}(C^* + O^*) + 2C^* + 0.807\,\frac{S^*}{\alpha} \\ &\leq (C^* + O^*)((\alpha+\epsilon)\rho_{st} + (2+\epsilon)\rho_{sf} + 2) + S^* \left(\rho_{st} + \frac{0.807}{\alpha} \right) \overset{\alpha=0.1712}{<} 6.27\, Z^*, \end{aligned}$$

for $\epsilon > 0$ small enough.

□

It remains to present our 2-approximation for constant $|\mathcal{D}|/M$.

Theorem 12. *If $|\mathcal{D}|/M = O(1)$, then there is a 2-approximation for soft-CFL.*

Proof. With the usual notation, let F^* be the optimal set of facilities, T^* the optimal Steiner tree over F^*, and σ^* the optimal assignment of clients to

facilities of F^*. The cost of the optimal solution is

$$Z^* = \underbrace{\sum_{i \in F^*} \left\lceil \frac{|\sigma^{*-1}(i)|}{b(i)} \right\rceil f(i)}_{O^*} + \overbrace{Mc(T^*)}^{S^*} + \underbrace{\sum_{j \in \mathcal{D}} \ell(j, \sigma^*(j))}_{C^*}.$$

Consider the following algorithm:

1. For all possible choices of $F \subseteq \mathcal{F}$ with $|F| \leq k := \lceil 2|\mathcal{D}|/M \rceil$ do:
   (a) Compute an optimal Steiner tree T over F.
   (b) Assign every client $j \in \mathcal{D}$ to the facility $\sigma(j)$ minimizing $\ell(j, \sigma(j)) + f(\sigma(j))/b(\sigma(j))$.
2. Output a minimum cost solution (F, T, σ) obtained.

Note that the algorithm above can be implemented in polynomial time, using, e.g., the algorithm by Dreyfus and Wagner [7] to compute optimal Steiner trees, since $k = O(1)$.

Since the algorithm considers all possible subsets $F \subseteq \mathcal{F}$ with $|F| \leq k$, it is sufficient to bound the cost of a specific choice of F. Derive from T^* an Eulerian tour E^* by the same construction as in the proof of Theorem 2. In particular, $|E^*| = |\mathcal{D}|$, $c(E^*) \leq 2c(T^*)$, and each $i \in F^*$ appears in $|\sigma^{*-1}(i)|$ copies in E^*. Split E^* in k intervals $I_1, I_2, \ldots, I_k$ of at most $\lceil M/2 \rceil$ vertices each. In each interval mark the facility i which minimizes the quantity $f(i)/b(i)$, and let F be the set of marked facilities. (Thereby it does not harm if the same facility is marked more than once.)

Trivially, since $F \subseteq F^*$, we have $c(T) \leq c(T^*) = S^*/M$, and hence the Steiner cost paid by the algorithm is $Mc(T) \leq S^*$. Next, consider the sum of the connection and opening cost paid by the algorithm:

$$\begin{aligned}\sum_{j \in \mathcal{D}} \ell(j, \sigma(j)) + \sum_{i \in F} \left\lceil \frac{|\sigma^{-1}(i)|}{b(i)} \right\rceil f(i) &\leq \sum_{j \in \mathcal{D}} \left(\ell(j, \sigma(j)) + \frac{f(\sigma(j))}{b(\sigma(j))} \right) + \sum_{i \in F} f_i \\ &\leq \underbrace{\sum_{j \in \mathcal{D}} \left(\ell(j, \sigma(j)) + \frac{f(\sigma(j))}{b(\sigma(j))} \right)}_{A} + O^*,\end{aligned}$$

where we used the fact that $F \subseteq F^*$, and hence $\sum_{i \in F} f_i \leq \sum_{i \in F^*} f_i = O^*$. It remains to bound A. Let $\sigma'(j)$ be the marked facility in the interval containing j's unique copy of $\sigma^*(j)$. For a given choice of F, our algorithm minimizes A. Thus replacing $\sigma(j)$ with $\sigma'(j)$ provides a feasible upper bound on A. Recall that $\sigma'(j)$ minimizes the ratio $f(i)/b(i)$ over the interval associated to j. In particular,

$$\frac{f(\sigma'(j))}{b(\sigma'(j))} \leq \frac{f(\sigma^*(j))}{b(\sigma^*(j))}$$

since $\sigma^*(j)$ lies in the same interval. By the observations above and triangle inequality,

$$\begin{aligned} A &\leq \sum_{j\in\mathcal{D}} \left(\ell(j,\sigma'(j)) + \frac{f(\sigma'(j))}{b(\sigma'(j))} \right) \\ &\leq \sum_{j\in\mathcal{D}} \ell(j,\sigma^*(j)) + \sum_{j\in\mathcal{D}} \ell(\sigma^*(j),\sigma'(j)) + \sum_{j\in\mathcal{D}} \frac{f(\sigma^*(j))}{b(\sigma^*(j))} \\ &\leq C^* + \underbrace{\sum_{j\in\mathcal{D}} \ell(\sigma^*(j),\sigma'(j))}_{B} + O^*. \end{aligned}$$

In order to upper bound B, replace each shortest path from $\sigma^*(j)$ to $\sigma'(j)$ with the shortest path *in the Eulerian tour* E^* between the same two endpoints, and let $\ell'(\sigma^*(j),\sigma'(j)) \geq \ell(\sigma^*(j),\sigma'(j))$ be the length of the latter path. Since $\sigma^*(j)$ and $\sigma'(j)$ lie in the same interval I, containing at most $\lceil M/2 \rceil$ vertices, each edge of E^* is used by at most $\lceil M/2 \rceil - 1 \leq M/2$ such (longer) paths. Hence

$$B \leq \sum_{j\in\mathcal{D}} \ell'(\sigma^*(j),\sigma'(j)) \leq \frac{M}{2} c(E^*) \leq M c(T^*) = S^*.$$

Altogether, the cost paid by the algorithm is at most

$$S^* + A + O^* \leq S^* + C^* + B + O^* + O^* \leq 2S^* + C^* + 2O^* \leq 2Z^*.$$

□

7. Conclusions

We described a simple algorithmic framework, based on random facility sampling, to solve connected facility location problems. By means of our novel core detouring scheme, we showed that this framework yields much better approximation algorithms for the problems considered in this paper.

We leave open the question whether core detouring can also be used to obtain significantly better approximation algorithms for *MROB* and the single-sink buy-at-bulk problem. The major difficulty here is that the optimum solution does not exhibit a single central core. While a small improvement seems nonetheless possible for the single-sink buy-at-bulk problem, the situation is less clear for *MROB*.

There is a strong relation between random sampling algorithms and the boosted sampling framework for two-stage stochastic optimization with recourse by Gupta et al. [21]. It is a very interesting open question whether our core detouring scheme also leads to improved approximation algorithms in that framework.

References

[1] A. Agrawal, P. Klein, and R. Ravi. When trees collide: an approximation algorithm for the generalized Steiner problem on networks. SIAM Journal on Computing, 24:440–456., 1995.

[2] M. Andrews and L. Zhang. The access network design problem. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 40–49, 1998.

[3] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998.

[4] L. Becchetti, J. Könemann, S. Leonardi, and M. Pál. Sharing the cost more efficiently: improved approximation for multicommodity rent-or-buy. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 375–384, 2005.

[5] M. Bern and P. Plassmann. The Steiner problem with edge lengths 1 and 2. *Information Processing Letters*, 32(4):171–176, 1989.

[6] N. Christofides. Worst-case analysis of a new heuristic for the travelling salesman problem. Technical report, Graduate School of Industrial Administration, Carnegie-Mellon University, 1976.

[7] S. E. Dreyfus and R. A. Wagner. The Steiner problem in graphs. *Networks*, 1:195–207, 1971/72.

[8] F. Eisenbrand and F. Grandoni. An improved approximation algorithm for virtual private network design. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 928–932, 2005.

[9] F. Eisenbrand, F. Grandoni, G. Oriolo, and M. Skutella. New Approaches for Virtual Private Network Design. In *International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1151–1162, 2005.

[10] F. Eisenbrand, F. Grandoni, G. Oriolo, and M. Skutella. New Approaches for Virtual Private Network Design. SIAM Journal on Computing, 37(3):706–721, 2007.

[11] F. Eisenbrand, F. Grandoni, T. Rothvoß, and G. Schäfer. Approximating connected facility location problems via random facility sampling and core detouring. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1174–1183, 2008.

[12] L. Fleischer, J. Könemann, S. Leonardi, and G. Schäfer. Simple cost sharing schemes for multicommodity rent-or-buy and stochastic Steiner tree. In *ACM Symposium on the Theory of Computing (STOC)*, pages 663–670, 2006.

[13] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Freeman, San Francisco, 1979.

[14] M. X. Goemans and D. P. Williamson. A general approximation technique for constrained forest problems. SIAM Journal on Computing, 24:296–317, 1995.

[15] F. Grandoni and G. F. Italiano. Improved Approximation for Single-Sink Buy-at-Bulk. In *International Symposium on Algorithms and Computation (ISAAC)*, pages 111–120, 2006.

[16] S. Guha, A. Meyerson, and K. Munagala. A constant factor approximation for the single sink edge installation problem. In *ACM Symposium on the Theory of Computing (STOC)*, pages 383–388, 2001.

[17] A. Gupta, J. Kleinberg, A. Kumar, R. Rastogi, and B. Yener. Provisioning a virtual private network: a network design problem for multicommodity flow. In *ACM Symposium on the Theory of Computing (STOC)*, pages 389–398, 2001.

[18] A. Gupta, A. Kumar, M. Pal, and T. Roughgarden. Approximation via cost-sharing: a simple approximation algorithm for the multicommodity rent-or-buy problem. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 606–617, 2003.

[19] A. Gupta, A. Kumar, M. Pal, and T. Roughgarden. Approximation via cost-sharing: simpler and better approximation algorithms for network design. *Journal of the ACM*, 54(3): 11, 2007.

[20] A. Gupta, A. Kumar, and T. Roughgarden. Simpler and better approximation algorithms for network design. In *ACM Symposium on the Theory of Computing (STOC)*, pages 365–372, 2003.

[21] A. Gupta, M. Pál, R. Ravi, and A. Sinha. Boosted sampling: approximation algorithms for stochastic optimization. In *ACM Symposium on the Theory of Computing (STOC)*, pages 417–426, 2004.

[22] A. Gupta, A. Srinivasan, and E. Tardos. Cost-sharing mechanisms for network design. In *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, pages 139–150, 2004.

[23] H. Jung, M. K. Hasan, and K. Chwa, Improved Primal-Dual Approximation Algorithm for the Connected Facility Location Problem. In *Conference on Combinatorial Optimization and Applications (COCOA)*, pages 265–277, 2008.

[24] M. Held and R. M. Karp, A dynamic programming approach to sequencing problems. *Journal of SIAM*, 10:196–210, 1962.

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

[26] K. Jain and V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual scheme and Lagrangian relaxation. *Journal of the ACM*, 48:274–296, 2001.

[27] R. Jothi and B. Raghavachari. Improved Approximation Algorithms for the Single-Sink Buy-at-Bulk Network Design Problems. In *Scandinavian Workshop on Algorithm Theory (SWAT)*, pages 336–348, 2004.

[28] D. R. Karger and M. Minkoff. Building Steiner trees with incomplete global knowledge. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 613–623, 2000.

[29] M. Mahdian, Y. Ye, and J. Zhang. Improved approximation algorithms for metric facility location problems. In *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, pages 229–242, 2002.

[30] M. Mahdian, Y. Ye, and J. Zhang. A 2-approximation algorithm for the soft-capacitated facility location problem. In *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, pages 129–242, 2003.

[31] A. Meyerson. Online facility location. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 426–431, 2001.

[32] P. B. Mirchandani and R. L. Francis. *Discrete Location Theory*. John Wiley and Sons, Inc., New York, 1990.

[33] C. Papadimitriou and M. Yannakakis. The traveling salesman problem with distances one and two. *Mathematics of Operations Research*, 18:1–11, 1993.

[34] R. Ravi and F. S. Salman. Approximation algorithms for the travelling purchaser problem and its variants in network design. In *European Symposium on Algorithms (ESA)*, pages 29–40, 1999.

[35] R. Ravi and A. Sinha. Approximation algorithms for problems combining facility location and network design. *Operations Research*, 54(1):73–81, 2006.

[36] G. Robins and A. Zelikovsky. Improved Steiner tree approximation in graphs. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 770–779, 2000.

[37] C. Swamy and A. Kumar. Primal-dual algorithms for connected facility location problems. In *International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, pages 256–269, 2002.

[38] C. Swamy and A. Kumar. Primal–dual algorithms for connected facility location problems. *Algorithmica*, 40(4):245–269, 2004.

[39] K. Talwar. The single-sink buy-at-bulk LP has constant integrality gap. In *International Conference on Integer Programming and Combinatorial Optimization (IPCO)*, pages 475–486, 2002.

[40] A. van Zuylen. Deterministic Sampling Algorithms for Network Design. In *European Symposium on Algorithms (ESA)*, pages 830–841, 2008.

[41] A. van Zuylen and D. Williamson. A simpler and better derandomization of an approximation algorithm for single source rent-or-buy. *Operations Research Letters*, 35(6):707–712, 2007.

[42] J. Vygen. Approximation algorithms for facility location problems (Lecture Notes). Report No. 05950-OR, Research Institute for Discrete Mathematics, University of Bonn, 2005.

Appendix

In this section we will prove Theorem 5. Before proving the theorem, we list a few useful equations, whose simple (and tedious) proof is left as an exercise for the reader. For any integer $h \geq 0$ and any real number $0 < q < 1$:

$$\sum_{j=0}^{h} q^j = \frac{1 - q^{h+1}}{1 - q} \tag{5}$$

$$\sum_{j=0}^{h} j\, q^j = \frac{q - (h+1)q^{h+1} + h\, q^{h+2}}{(1-q)^2} \tag{6}$$

$$\sum_{j=0}^{h} j^2\, q^j = \frac{q + q^2 - (h+1)^2 q^{h+1} + (2h^2 + 2h - 1)q^{h+2} - h^2 q^{h+3}}{(1-q)^3} \tag{7}$$

$$\begin{aligned}\sum_{j=0}^{h} j^3\, q^j = {} & \frac{q + 4q^2 + q^3 - (h+1)^3 q^{h+1} + (3h^3 + 6h^2 - 4)q^{h+2}}{(1-q)^4} \\ & + \frac{(-3h^3 - 3h^2 + 3h - 1)q^{h+3} + h^3 q^{h+4}}{(1-q)^4}\end{aligned} \tag{8}$$

$$\sum_{i=0}^{h}\sum_{j=0}^{i} i^2 j\, q^{i+j} = \frac{q}{(1-q)^2}\left(\sum_{j=0}^{h} j^2 q^j - \sum_{j=0}^{h} j^2 (q^2)^j\right) - \frac{q}{(1-q)}\sum_{j=0}^{h} j^3 (q^2)^j \quad (9)$$

$$\sum_{i=0}^{h}\sum_{j=0}^{i} i\, j^2\, q^{i+j} = \frac{q+q^2}{(1-q)^3}\left(\sum_{j=0}^{h} j\, q^j - \sum_{j=0}^{h} j\, (q^2)^j\right) - \frac{2q}{(1-q)^2}\sum_{j=0}^{h} j^2 (q^2)^j$$
$$- \frac{q}{1-q}\sum_{j=0}^{h} j^3 (q^2)^j \quad (10)$$

Proof. *(Theorem 5)* Our client sampling process is equivalent to:

(1) Mark each client independently with probability p.

(2) Choose a client j^* (either marked or not) uniformly at random, and mark it.

Consider the following modified sampling process:

(a) Run (1).

(b) If no client is marked in Step (a), run (2).

Let Y denote the cost of the flow in the tree-connection game with respect to the modified sampling scheme. By a simple coupling argument, it is easy to see that $\mathbf{E}[X] \le \mathbf{E}[Y]$: Intuitively, sampling fewer clients can only make the cost of the flow larger (in expectation). Hence it is sufficient to bound $\mathbf{E}[Y]$.

Let Q denote the event that in Step (b) of the modified game we run (2). By elementary probability theory,

$$\mathbf{E}[Y] = \Pr(Q)\mathbf{E}[Y \mid Q] + \Pr(\bar{Q})\mathbf{E}[Y \mid \bar{Q}].$$

Trivially, $\Pr(Q) = (1-p)^{|\mathcal{D}|}$. Moreover, $\mathbf{E}[Y \mid Q] \le w(\mathcal{H}) + |\mathcal{D}|\, w(\mathcal{T})$. We will next show that

$$\mathbf{E}[Y \mid \bar{Q}] \le w(\mathcal{H}) + w(\mathcal{T})\,\frac{0.8067}{p}. \quad (11)$$

The claim easily follows:

$$\begin{aligned}\mathbf{E}[Y] &\le w(\mathcal{H}) + w(\mathcal{T})\,\left((1-p)^{|\mathcal{D}|}|\mathcal{D}| + \frac{0.8067}{p}\right)\\ &\le w(\mathcal{H}) + w(\mathcal{T})\,\left(e^{-p|\mathcal{D}|}|\mathcal{D}| + \frac{0.8067}{p}\right)\\ &\le w(\mathcal{H}) + w(\mathcal{T})\,\frac{0.807}{p},\end{aligned}$$

where we used the assumption $|\mathcal{D}| \gg 1/p$ which implies $e^{-p|\mathcal{D}|}|\mathcal{D}| \ll 1/p$.

It remains to prove (11). Subsequently, we assume that the event $\bar{Q}$ holds. It is clear that $\mathbf{E}[f(e)] \leq 1$ holds for every $e \in \mathcal{H}$. Thus it is sufficient to show that $\mathbf{E}[f(e)] \leq 0.8067/p$ for any given $e \in \mathcal{T}$. Let e_1 and e_2 be the two edges of $\mathcal{C}$ associated with e. We define the flow $f(e_i)$ along e_i in $\mathcal{C}$ to be positive if it goes clockwise and negative otherwise.

If $e_1 = e_2$, $\mathbf{E}[f(e)] = \mathbf{E}[|f(e_1)|] \leq 1/(2\,p)$ by essentially the standard analysis. Hence, let us assume $e_1 \neq e_2$. In this case, $F := f(e) = |f(e_1) - f(e_2)|$ by flow canceling. We next introduce some notation. Let $m = |\mathcal{D}|$. Let moreover I and I' be the two paths obtained by removing e_1 and e_2 from $\mathcal{C}$. Without loss of generality, we assume that I is the shortest of the two paths (in terms of number of edges), and we denote its length by $k := |I|$. Observe that $0 \leq k \leq m/2 - 1$ and $k \leq k' := |I'| = m - k - 2$. We also assume, still without loss of generality, that e_1 is incident to the left endpoint of I.

The value of $\mathbf{E}[F]$ is a (complicated) function of p, m, and k. Recall that each node of $\mathcal{C}$ is by-sampled with probability p, but under the event $\bar{Q}$ that at least one (random) node is by-sampled. Let $q = 1 - p$. We distinguish three events A, B, and C, which partition the considered probability space:

(A) No node selected in I, at least one node selected in I'. The value of F is deterministically $k+1$. In fact, if h flow-paths along I are directed to the left and the other $k+1-h$ to the right (event A'), then $f(e_1) = -h$, $f(e_2) = k+1-h$, and altogether $\mathbf{E}[F \,|\, A'] = \mathbf{E}[|(-h) - (k+1-h)|] = k+1$. Otherwise (event A''), the flow on e_1 and e_2 must go in the same direction, say from left to right, and it must be $f(e_2) = f(e_1) + k + 1$ (e_2 collects the same flow as e_1 plus the flow along I). Then $\mathbf{E}[F \,|\, A''] = \mathbf{E}[|f(e_1) - (f(e_1) + k + 1)|] = k+1$. Since event A happens with probability $q^{k+1}(1 - q^{k'+1})/(1 - q^m)$, the overall contribution of this case to the total expected flow is

$$F_A = \Pr(A)\mathbf{E}[F \,|\, A] = \frac{q^{k+1}(1 - q^{k'+1})}{1 - q^m}(k+1).$$

(B) No node selected in I', at least one node selected in I. By essentially the same argument as in case (A), we obtain

$$F_B = \Pr(B)\mathbf{E}[F \,|\, B] = \frac{q^{k'+1}(1 - q^{k+1})}{1 - q^m}(k'+1).$$

(C) At least one node selected in both I and I'. Let us denote by L_i (resp., R_i) the distance between the left (resp., right) endpoint of e_i and the first by-sampled node to its left (resp., right). It is not hard to see that $\mathbf{E}[f(e_i)] = (L_i - R_i)/2$. Define $X := L_2 + R_1 \leq k$ and $X' = L_1 + R_2 \leq k'$. Note that $\mathbf{E}[F \,|\, C] = \frac{1}{2}\mathbf{E}[|X' - X|]$. Moreover, X and X' are independent. Let us study the probability distribution of X. For $0 \leq a + b < k$, $\Pr(L_2 = a, R_1 = b) = \frac{pq^a \cdot pq^b}{1 - q^{k+1}}$. For $a + b = k$, $\Pr(L_2 = a, R_1 = b) = \frac{q^a \cdot q^b \cdot p}{1 - q^{k+1}}$. We can conclude that

$$\Pr(X = i) = \sum_{a=0}^{i} \Pr(L_2 = a, R_1 = i - a) = \begin{cases} (i+1)\frac{p^2 q^i}{1 - q^{k+1}} & \text{if } i \in [0, k-1]; \\ (k+1)\frac{p\,q^k}{1 - q^{k+1}} & \text{if } i = k. \end{cases}$$

Analogously,

$$\Pr(X' = j) = \begin{cases} (j+1)\frac{p^2 q^j}{1-q^{k'+1}} & \text{if } j \in [0, k'-1]; \\ (k'+1)\frac{p\, q^{k'}}{1-q^{k'+1}} & \text{if } j = k'. \end{cases}$$

Note that, as expected, $\sum_{i=0}^{k} \Pr(X = i) = \sum_{j=0}^{k'} \Pr(X' = j) = 1$. The contribution of this case to the overall flow is

$$\begin{aligned} F_C &= \Pr(C)\mathbf{E}[F \mid C] = \frac{(1-q^{k+1})(1-q^{k'+1})}{2(1-q^m)} \sum_{i=0}^{k}\sum_{j=0}^{k'} |i-j| \Pr(X=i)\Pr(X'=j) \\ &= \frac{(k+1)p^3q^k}{2(1-q^m)} \left(\sum_{j=0}^{k-1}(k-j)(j+1)q^j + \sum_{j=k}^{k'-1}(j-k)(j+1)q^j \right) \\ &+ \frac{(k+1)(k'+1)(k'-k)p^2q^{k'+k}}{2(1-q^m)} + \frac{(k'+1)p^3q^{k'}}{2(1-q^m)} \sum_{i=0}^{k-1}(k'-i)(i+1)q^i \\ &+ \frac{p^4}{2(1-q^m)} \left(\sum_{i=0}^{k-1}\sum_{j=0}^{i-1}(i-j)(i+1)(j+1)q^{i+j} + \sum_{i=0}^{k-1}\sum_{j=i}^{k'-1}(j-i)(i+1)(j+1)q^{i+j} \right) \\ &= \frac{(k+1)p^3q^k}{2(1-q^m)} \left(2\sum_{j=0}^{k-1}(k-j)(j+1)q^j + \sum_{j=0}^{k'-1}(j-k)(j+1)q^j \right) \\ &+ \frac{(k+1)(k'+1)(k'-k)p^2q^{k'+k}}{2(1-q^m)} + \frac{(k'+1)p^3q^{k'}}{2(1-q^m)} \sum_{i=0}^{k-1}(k'-i)(i+1)q^i \\ &+ \frac{p^4}{2(1-q^m)} \left(2\sum_{i=0}^{k-1}\sum_{j=0}^{i-1}(i-j)(i+1)(j+1)q^{i+j} + \sum_{i=0}^{k-1}\sum_{j=0}^{k'-1}(j-i)(i+1)(j+1)q^{i+j} \right) \end{aligned}$$

By variable substitution,

$$\begin{aligned} F_C &= \frac{(k+1)p^3q^k}{2(1-q^m)} \left(2q^{-1}\sum_{j=0}^{k}(k+1-j)j\, q^j - q^{-1}\sum_{j=0}^{k'}(k+1-j)j\, q^j \right) \\ &+ \frac{(k+1)(k'+1)(k'-k)p^2q^{k'+k}}{2(1-q^m)} + \frac{(k'+1)p^3q^{k'}}{2(1-q^m)} q^{-1}\sum_{i=0}^{k}(k'+1-i)i\, q^i \\ &+ \frac{p^4}{2(1-q^m)} \left(2q^{-2}\sum_{i=0}^{k}\sum_{j=0}^{i}(i-j)i\, j\, q^{i+j} - q^{-2}\sum_{i=0}^{k}\sum_{j=0}^{k'}(i-j)i\, j\, q^{i+j} \right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{(k+1)^2 p^3 q^{k-1}}{(1-q^m)} \sum_{j=0}^{k} j\, q^j - \frac{(k+1) p^3 q^{k-1}}{(1-q^m)} \sum_{j=0}^{k} j^2\, q^j \\
&- \frac{(k+1)^2 p^3 q^{k-1}}{2(1-q^m)} \sum_{j=0}^{k'} j\, q^j + \frac{(k+1) p^3 q^{k-1}}{2(1-q^m)} \sum_{j=0}^{k'} j^2\, q^j \\
&+ \frac{(k+1)(k'+1)(k'-k) p^2 q^{k'+k}}{2(1-q^m)} \\
&+ \frac{(k'+1)^2 p^3 q^{k'-1}}{2(1-q^m)} \sum_{j=0}^{k} j\, q^j - \frac{(k'+1) p^3 q^{k'-1}}{2(1-q^m)} \sum_{j=0}^{k} j^2\, q^j \\
&+ \frac{p^4}{q^2(1-q^m)} \left(\sum_{i=0}^{k} \sum_{j=0}^{i} i^2\, j\, q^{i+j} - \sum_{i=0}^{k} \sum_{j=0}^{i} i\, j^2\, q^{i+j} \right) \\
&+ \frac{p^4}{2q^2(1-q^m)} \left(-\sum_{j=0}^{k} j^2 q^j \cdot \sum_{j=0}^{k'} j\, q^j + \sum_{j=0}^{k} j\, q^j \cdot \sum_{j=0}^{k'} j^2\, q^j \right)
\end{aligned}
$$

Substituting Equations (5)-(10), and simplifying the formula,

$$
\begin{aligned}
F_C &= \frac{q^m(k'-k) - q^{k+1}(k+1) - q^{k'+1}(k'+1)}{1-q^m} \\
&+ \frac{2q(1+q+q^2) + q^{2k+2}(k^2(1-q^2)^2 + k(1-q^2)(3-q^2) + (2-2q(1+q)^2))}{(1-q)(1+q)^3(1-q^m)}.
\end{aligned}
$$

Altogether we obtain

$$
\begin{aligned}
\mathbf{E}[F] &= \Pr(A)\mathbf{E}[F|A] + \Pr(B)\mathbf{E}[F|B] + \Pr(C)\mathbf{E}[F|C] = F_A + F_B + F_C \\
&= \frac{-2(k+1)q^m}{1-q^m} \\
&+ \frac{2q(1+q+q^2) + q^{2k+2}(k^2(1-q^2)^2 + k(1-q^2)(3-q^2) + (2-2q(1+q)^2))}{(1-q)(1+q)^3(1-q^m)}.
\end{aligned}
$$

Observe that $\frac{-2(k+1)q^m}{1-q^m} < 0$, and recall that $q^m = (1-p)^m \leq e^{-p|\mathcal{D}|} \leq \epsilon$, for an arbitrarily small constant $\epsilon > 0$. Then

$$
\mathbf{E}[F] \leq \frac{R(q,k)}{p(1-\epsilon)},
$$

where

$$
R(q,k) := \frac{2q(1+q+q^2) + q^{2k+2}(k^2(1-q^2)^2 + k(1-q^2)(3-q^2) + 2 - 2q(1+q)^2)}{(1+q)^3}.
$$

Our goal is showing that $\mathbf{E}[F] \leq 0.8067/p$: to that aim it is sufficient to show that $R(q,k) \leq 0.8066$ for any q and k. Trivially

$$\sup_{\substack{0<q<1\\0\leq k\leq k'}} \{R(q,k)\} \leq \sup_{\substack{0<q<1\\x\geq 0}} \{R(q,x)\}.$$

Differentiating function $R(q,x)$ with respect to x:

$$\frac{\partial R(q,x)}{\partial x} = \frac{x^2 q^{2x+2}(2\ln q(1-q^2)^2)}{(1+q)^3} + \frac{xq^{2x+2}(2\ln q(1-q^2)(3-q^2)+2(1-q^2)^2)}{(1+q)^3} + \frac{q^{2x+2}(2\ln q(2-2q(1+q)^2)+(1-q^2)(3-q^2))}{(1+q)^3}.$$

The two roots of $\frac{\partial R(q,x)}{\partial x}$ are

$$x_1(q) := \frac{(q^2-3)\ln q - (1-q^2) - \sqrt{(1+8q+10q^2+8q^3+q^4)\ln^2 q + (1-q^2)^2}}{2(1-q^2)\ln q}.$$

and

$$x_2(q) := \frac{(q^2-3)\ln q - (1-q^2) + \sqrt{(1+8q+10q^2+8q^3+q^4)\ln^2 q + (1-q^2)^2}}{2(1-q^2)\ln q}.$$

Recall that the domain of $R(q,x)$ is $q \in (0,1)$ and $x \geq 0$. Function $x_2(q)$ is always negative for $q \in (0,1)$, while $x_1(q)$ can be either positive or negative, depending on q. As a consequence, for any fixed $q \in (0,1)$, $R(q,x)$ is maximized either for $x = 0$, or for $x = x_1(q)$, or for $x \to +\infty$. One has

$$R(q,0) = \frac{2q+4q^2-4q^4-2q^5}{(1+q)^3} \leq 0.5$$

and

$$\lim_{x\to+\infty} R(q,x) = \frac{2q(1+q+q^2)}{(1+q)^3} \leq 0.75.$$

Function $R(q,x_1(q))$ is monotonically increasing in q. Hence

$$\begin{aligned} R(q,x_1(q)) &\leq \lim_{q\to 1^-} R(q,x_1(q)) \\ &= \frac{3}{4} + \lim_{q\to 1^-} \frac{q^{-2\frac{\sqrt{8}}{2\ln q}}}{8}\left(\frac{8}{4\ln^2 q}4(1-q)^2 - \frac{\sqrt{8}}{2\ln q}4(1-q) - 6\right) \\ &= \frac{3}{4} + \frac{e^{-\sqrt{8}}}{8}\left(8+2\sqrt{8}-6\right) = \frac{3+(1+\sqrt{8})e^{-\sqrt{8}}}{4} < 0.806571. \end{aligned}$$

This concludes the proof of the theorem. □