Approximating Connected Facility Location Problems via Random Facility Sampling and Core Detouring

Friedrich Eisenbrand[*] Fabrizio Grandoni[†] Thomas Rothvoß[‡] Guido Schäfer[§]

July 6, 2007

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. We show that our connected facility location algorithms 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.

[*] Institut für Mathematik, Universität Paderborn, Germany. Email: eisen@math.uni-paderborn.de.

[†] Dipartimento di Informatica, Università di Roma "La Sapienza", Via Salaria 113, 00198 Roma, Italy. Email: grandoni@di.uniroma1.it. On leave at TU Berlin.

[‡] Institut für Mathematik, Universität Paderborn, Germany. Email: rothvoss@math.uni-paderborn.de.

[§] Technische Universität Berlin, Institut für Mathematik, Straße des 17. Juni 136, 10623 Berlin, Germany. Email: schaefer@math.tu-berlin.de.

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 science literature (see, e.g., [9, 12, 17, 26]) and in the operations research literature (see, e.g., [19, 23]).

As an example (see also [1, 26]), 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 to 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 [12] 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. 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*).

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.

Problem	**This paper**	**Previous best**	
CFL	4.00^*	8.55	Swamy and Kumar [25, 26]
	4.23		
SROB	2.92^*	3.55^*	Gupta et al. [11, 12]
	3.28	4	van Zuylen and Williamson [27]
k-CFL	6.85^*	15.55^*	Swamy and Kumar [25, 26]
	6.98		
tour-CFL	4.12^*	5.83^*	Ravi and Salman [22] (special case only)

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

Our framework yields a 4.00-approximation algorithm for *CFL*, which improves over the current best primal-dual 8.55-approximation algorithm by Swamy and Kumar [25, 26]. In the special case of *SROB*, our algorithm provides a 2.92-approximation, hence improving on the current best 3.55-approximation algorithm by Gupta et al. [10, 11]. We show that our algorithms for *SROB* and *CFL* can be derandomized using the method of conditional expectations (see, e.g., [20]) and an idea that van Zuylen and Williamson [27] used to derandomize the *SROB* algorithm of Gupta et al. [10, 11]; thereby the approximation ratios degrade only slightly. We eventually demonstrate the versatility of our framework by applying it to the problems *k-CFL* and *tour-CFL*, for which we improve the current best known approximation ratios. The results presented in this paper are summarized in Table 1.

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 (ideally) a 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. As a by-product of our analysis, we obtain a polynomial-time approximation scheme (PTAS) for the above problems if $|\mathcal{D}|/M$ is a constant. This might be of independent interest.

Previous and Related Work. The network design problems considered here are NP-hard [8] and APX-complete [2, 4, 21], 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. [9] obtain a 10.66-approximation algorithm for *CFL*, based on rounding an exponential size LP. The current best algorithm for *CFL* is a primal-dual 8.55-approximation algorithm by Swamy and Kumar [25, 26]. Better results are known for *SROB*. Gupta et al. [9] give a 9.01-approximation algorithm. Swamy and Kumar [25, 26] describe a primal-dual 4.55-approximation algorithm for the same problem. Gupta, Kumar, and Roughgarden [12] propose a simple random sampling algorithm which gives a 3.55-approximation. Gupta, Srinivasan and Tardos [14] show that this algorithm can be derandomized to obtain a 4.2-approximation algorithm. In a recent work, van Zuylen and Williamson [27] present a derandomization of the random sampling algorithm that yields a 4-approximation.

Swamy and Kumar [25, 26] give a 15.55-approximation algorithm for *k-CFL*, which is also the current best. Ravi and Salman [22] 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.

Most of the existing random sampling algorithms for connected facility location problems are analyzed by means of *strict cost shares* (see, e.g., [10, 12] and in particular the exposition in [11]), a concept originating from game-theoretic cost sharing. Basically, these cost shares are used to relate the expected connection cost of the computed solution to the cost of the core in the optimum solution. This concept has been used successfully to obtain simple and good approximation algorithms for network design problems, such as *SROB* [11, 12] and *MROB* [3, 7, 10], the multi-commodity counterpart of *SROB*. However, its use failed to prove better bounds for more general connected facility location problems. In fact, in [12], Gupta et al. leave open the question whether a randomized sampling approach can be used to improve the primal-dual approximation algorithm of Swamy and Kumar [25, 26]. In this paper, we answer this question affirmatively.

Organization of Paper. In Section 2, we study core connection games, which form the basis of our core detouring scheme. We present the polynomial-time approximation scheme for constant $\mathcal{D}/M$ in Section 3. Our random facility sampling framework for *CFL* and *SROB* and its analysis are given in Section 4. The extensions of this framework to other connected facility location problems are outlined in Section 5. Finally, we give some conclusions in Section 6.

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 4.

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 in the Appendix. 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 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,

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

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

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

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.

Corollary 1. *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 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 Polynomial-time Approximation Schemes for Constant $|\mathcal{D}|/\mathbf{M}$

In this section, we present polynomial-time approximation schemes (PTAS) for the connected facility location problems considered in this paper if $|\mathcal{D}|/M$ is upper bounded by a constant. These PTAS will help to improve our analysis for the general case; but might also be of independent interest.

Recall that $\ell(v,w)$ denotes 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 $W \subseteq V$. Let $c(S) = \sum_{e\in S} c(e)$ denote the total cost of all edges in a subset $S \subseteq E$.

Theorem 2. *If $|\mathcal{D}|/M = O(1)$, 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. 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) obtained.

In Step 1(a), we use, for example, the algorithm by Dreyfus and Wagner [6]. 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,

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

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

Using essentially the same arguments as above, it is not hard to obtain a PTAS for *tour-CFL* under the same assumptions. We state the following theorem without proof.

Theorem 3. *If $|\mathcal{D}|/M = O(1)$, there is a PTAS for tour-CFL.*

4 Connected Facility Location

Due to the results obtained in the previous section, we can assume that $M/|\mathcal{D}| \leq \varepsilon$ for a sufficiently small constant $\varepsilon > 0$. We also assume without loss of generality that $n \gg 1$. For a given assignment σ of clients to facilities, we let $\sigma^{-1}(i)$ denote the set of clients assigned to facility i.

4.1 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_{fl}-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.

We use the following 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 APX. 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.

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

Proof. We open a subset of the facilities in F_U, which costs at most O_U. □

The following bound on the Steiner cost is inspired by [12]. We recall that we assume $M/|\mathcal{D}| \leq \varepsilon$.

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

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

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

□

Core Detouring Scheme. We next introduce our new *core detouring scheme* to bound the expected connection cost of *APX*. Notice 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 Corollary 1 it follows

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

□

Now we have all the ingredients to prove the main result of this paper. The following theorem relies on the current best approximation factors for Steiner tree and facility location, which are $\rho_{st} < 1.55$ [24] and $\rho_{fl} < 1.52$ [18], respectively.

Theorem 4. *For a proper choice of* α, `randCFL` *is an expected* 4.55*-approximation algorithm for CFL.*

Proof. By Lemmas 1, 2, and 3,

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

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

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

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

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 of Theorem 4, and choosing α accordingly, we obtain the following corollary.

Corollary 3. *For a proper choice of* α, `randCFL` *is an expected* 3.05*-approximation algorithm for SROB.*

4.2 Refinements

We can improve the approximation ratio of `randCFL` 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 [18] 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

$$C_U + O_U \leq (1.11 + \ln\delta)O^* + (1 + 0.78/\delta)C^*.$$

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 Corollary 1, 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 equally directed (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 lemma 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:

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

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 can be reduced to* 2.92*.*

Proof. Let us adapt the proof of Theorem 4. Combining (a) and (b), we obtain

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

The analysis above can be adapted to *SROB* by imposing $C_U = O_U = O^* = 0$. This yields

$$\mathbf{E}[Z] \leq \rho_{st}(S^* + (\alpha+\varepsilon)C^*) + 2C^* + 0.807\, S^*/\alpha \overset{\alpha=0.591}{<} 2.92\, Z^*.$$

□

4.3 Derandomization

We can derandomize our algorithm for *CFL* using the method of conditional expectation (see, e.g., [20]) and an idea by van Zuylen and Williamson [27]. Consider any possible choice of a client j_1. Intuitively, j_1 is the client j^* that we sample uniformly at random. Let $j_2, j_3, \ldots, j_{|\mathcal{D}|}$ be the remaining clients, in an arbitrary order. Initially, we mark j_1. In iteration k, $k \geq 2$, we decide whether to *mark* or *unmark* client j_k. Let D_{k-1} be the subset of clients in $\{j_1, j_2 \ldots, j_{k-1}\}$ that we already marked. Ideally, we would like to mark client j_k if and only if

$$\mathbf{E}[Z \,|\, D_k = D_{k-1} \cup \{j_k\}] \leq \mathbf{E}[Z \,|\, D_k = D_{k-1}].$$

This would ensure, for a proper choice of j_1, that the cost of the final solution is at most $4.00\,Z^*$.

It is not difficult to see that we can efficiently compute the expected opening cost and connection cost, given D_k. The same holds for the expected augmentation cost in Step 4. The problem is that we do not know how to compute the conditioned expected cost of the Steiner tree over D. However, as it is shown by van Zuylen and Williamson [27], we can compute an estimate of this cost if we use a primal-dual 2-approximation algorithm for the Steiner tree computation instead. In our analysis, we essentially only need to replace $\rho_{st} < 1.55$ by $\rho_{st} = 2$, which gives a slightly larger (but deterministic) approximation ratio.

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

5 Extensions

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

5.1 Connected k-Facility Location

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.

This algorithm can be refined using Technique (b). The following theorem relies on the current best approximation ratio for the k-facility location problem, which is $\rho_{kfl} \leq 4$ [15, 16] (see also [28]).

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

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

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

□

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

Corollary 4. *There is a deterministic* 6.98*-approximation algorithm for k-CFL.*

5.2 Tour-Connected Facility Location

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. Eventually, compute an Euler 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 [5].

Theorem 9. *There is an expected* 4.12*-approximation algorithm for tour-CFL.*

Proof (sketch). We adapt the analysis of Section 4. Trivially, $O \leq O_U$. Taking into account the duplication of the shortest paths from D to F, and using a similar duplication when bounding the cost of the optimum TSP-tour over D, we obtain

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

We can easily adapt Corollary 1 to this case, thus obtaining $\mathbf{E}[X] \leq w(\mathcal{H}) + w(\mathcal{T})/(2p)$. It follows that

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

Altogether

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

□

6 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 family of problems considered.

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. [13]. It is a very interesting open question whether our core detouring scheme also leads to improved approximation algorithms in that framework.

References

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

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

[3] 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. Society for Industrial and Applied Mathematics, 2005.

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

[5] 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.

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

[7] 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. ACM Press, 2006.

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

[9] 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.

[10] 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.

[11] A. Gupta, A. Kumar, M. Pal, and T. Roughgarden. Approximation via cost-sharing: simpler and better approximation algorithms for network design. To appear in Journal of the ACM, 2007.

[12] 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.

[13] 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. ACM Press, 2004.

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

[15] 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 Association for Computing Machinery*, 50:795–824, 2003.

[16] 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 Association for Computing Machinery*, 48:274–296, 2001.

[17] 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.

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

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

[20] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, Cambridge, first edition, 1995.

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

[22] 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.

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

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

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

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

[27] A. van Zuylen and D. Williamson. A simpler and better derandomization of an approximation algorithm for single source rent-or-buy. 2007. To appear in Operations Research Letters.

[28] J. Vygen. Approximation algorithms for facility location problems.

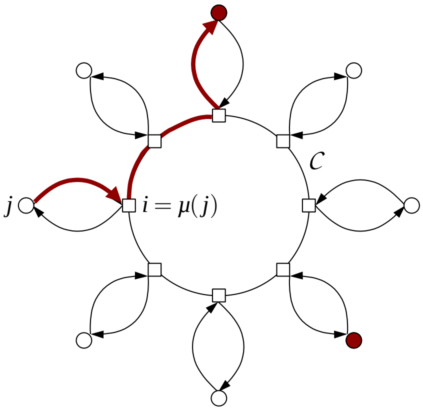

Figure 1: Core connection game instance. Marked client nodes are drawn in bold. The flow of j in the routing scheme is indicated by the bold path.

Appendix

Proof of 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 $\mathbf{E}[X] \leq \mathbf{E}[Y]$. Intuitively, sampling less 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,

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

Trivially, $\Pr(Q) = (1-p)^{|\mathcal{D}|}$. Moreover,

$$\mathbf{E}[Y \mid Q] \leq w(\mathcal{H}) + |\mathcal{D}|\, w(\mathcal{T})$$

We will next show that

$$\mathbf{E}[Y \mid \overline{Q}] \leq w(\mathcal{H}) + 0.8067\, w(\mathcal{T})/p. \tag{2}$$

From (2) we can conclude that

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

where we used the assumption $|\mathcal{D}| \gg 1/p$.

It remains to prove (2). 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 to e. We assume by definition that the flow $f(e_i)$ along e_i in $\mathcal{C}$ is positive if it goes clockwise, and negative otherwise.

If $e_1 = e_2$, $\mathbf{E}[f(e)] = \mathbf{E}[|f(e_1)|] \leq 1/(2p)$ by essentially the standard analysis. Hence, let us assume $e_1 \neq e_2$. In that case $F := f(e) = |f(e_1) - f(e_2)|$ by flow canceling. The value of $\mathbf{E}[F]$ is a (complicated) function of p, of $m = |\mathcal{D}|$, and of the distance k, $0 \leq k \leq m/2 - 1$, between e_1 and e_2 in $\mathcal{C}$.

We first need some notation. Let I, $k = |I|$, be the shortest path (in terms of number of hops) between e_1 and e_2 along $\mathcal{C}$. Without loss of generality, we assume e_1 is on the left side of I. Let I' be the complement of $I \cup \{e_1, e_2\}$ with respect to $\mathcal{C}$, and $k' := |I'| = m - k - 2$.

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. We let $q = 1 - p$, and distinguish three events A, B, and C, which partition the probability space considered:

(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 $F_1 = -h$, $F_2 = k+1-h$, and altogether $\mathbf{E}[F \mid A'] = \mathbf{E}[|(-h) - (k+1-h)|] = k+1$. 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 $\mathbf{E}[F \mid A''] = \mathbf{E}[|f(e_1) - (f(e_1) + k + 1)|] = k+1$. Since event A happens with probability $\frac{q^{k+1}(1-q^{k'+1})}{1-q^m}$, the overall contribution of this case to the total expected flow is

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

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

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

(C) At least one node selected in both I and I'. If we denote by L_i (R_i) the distance between e_i and the first by-sampled node to its left (right), then $\mathbf{E}[f(e_i)] = (L_i - R_i)/2$. Variables L_1, R_1, L_2, and R_2 can be interpreted as random geometric variables of parameter p, under the constraint that $X = L_2 + R_1 \leq k$ and $X' = L_1 + R_2 \leq k'$. Let us study the random variables X and X'. Note that $\mathbf{E}[F \mid C] = \frac{1}{2}\mathbf{E}[|X' - X|]$. Moreover, X and X' are independent. It is not hard to show that

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

Analogously

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

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

$$F_C = \Pr(C)\mathbf{E}[F \mid C] = \frac{(1-q^{k+1})(1-q^{k'+1})}{2(1-q^m)} \sum_{i=0}^{k} \sum_{j=0}^{k'} |i-j| \Pr(X = i) \Pr(X' = j).$$

Recall that $\mathbf{E}[F] = \Pr(A)\mathbf{E}[F|A] + \Pr(B)\mathbf{E}[F|B] + \Pr(C)\mathbf{E}[F|C] = F_A + F_B + F_C$. After a simple, but very long and tedious computation, we obtained

$$\begin{aligned}\mathbf{E}[F] \quad &= \frac{-2(k+1)q^m}{1-q^m} + \frac{2q(1+q+q^2)+q^{2k+2}(k^2(1-q^2)^2+k(1-q^2)(3-q^2)+(2-2q(1+q)^2))}{p(1-q^m)(1+q)^3} \\ &\leq \frac{2q(1+q+q^2)+q^{2k+2}(k^2(1-q^2)^2+k(1-q^2)(3-q^2)+(2-2q(1+q)^2))}{p(1-\varepsilon)(1+q)^3},\end{aligned}$$

where $\varepsilon > 0$ is an arbitrarily small constant. In the last inequality we used the assumptions that α is a positive constant and $m = |\mathcal{D}| \gg 1/p$. Consider function

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

where

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

It is sufficient to show that $R(q,k) \leq 0.8066 < 0.8067$ for any q and k. Fixing q, and maximizing in k,

$$\begin{aligned}\max_{0\leq k\leq k'}\{R(q,k)\} \quad &\leq \quad \frac{2q(1+q+q^2)}{(1+q)^3} + \frac{1}{(1+q)^3}\max_{0\leq k\leq k'}\{R'(q,k)\} \\ &\leq \quad \frac{2q(1+q+q^2)}{(1+q)^3} + \frac{1}{(1+q)^3}\max_{x\geq 0}\{R'(q,x)\}.\end{aligned}$$

By an elementary analysis of function $R'(q,x)$, we found that it has a maximum (either feasible or not) for

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

Then, by the constraint $x \geq 0$, function $R'(q,x)$ is maximized for $x = 0$ if $x(q) < 0$, and for $x = x(q)$ otherwise. In other words,

$$\max_{x\geq 0}\{R'(q,x)\} = R'(q, \max\{0, x(q)\}).$$

It follows that

$$\max_{0\leq k\leq k'}\{R(q,k)\} \leq \frac{2q(1+q+q^2)}{(1+q)^3} + \frac{R'(q,\max\{1,x(q)\})}{(1+q)^3}.$$

We found numerically that the right-hand side is upper bounded by 0.8066 for any feasible value of q. This concludes the proof of the theorem. □