Facility Location in Evolving Metrics *

David Eisenstat[1], Claire Mathieu[2], and Nicolas Schabanel[3,4]

[1] Brown University (USA) - http://www.davideisenstat.com/
[2] CNRS, École normale supérieure UMR 8548 (France)
[3] CNRS, Université Paris Diderot (France) - http://www.liafa.univ-paris-diderot.fr/~nschaban/
[4] IXXI, École normale supérieure de Lyon (France)

Abstract. Understanding the dynamics of evolving social or infrastructure networks is a challenge in applied areas such as epidemiology, viral marketing, and urban planning. During the past decade, data has been collected on such networks but has yet to be analyzed fully. We propose to use information on the dynamics of the data to find stable partitions of the network into groups. For that purpose, we introduce a time-dependent, dynamic version of the facility location problem, which includes a switching cost when a client's assignment changes from one facility to another. This might provide a better representation of an evolving network, emphasizing the abrupt change of relationships between subjects rather than the continuous evolution of the underlying network. We show for some realistic examples that this model yields better hypotheses than its counterpart without switching costs, where each snapshot can be optimized independently. For our model, we present an $O(\log nT)$-approximation algorithm and a matching hardness result, where n is the number of clients and T is the number of timesteps. We also give another algorithm with approximation ratio $O(\log nT)$ for a variant model where the decision to open a facility is made independently at each timestep.

1 Introduction

During the past decade, a massive amount of data has been collected on diverse networks such as the web (pages and links), social networks (e.g., Facebook, Twitter, and LinkedIn), and social encounters in hospitals, schools, companies, and conferences [18,21]. These networks evolve over time, and their dynamics have a considerable impact on their structure and effectiveness [19,14]. Understanding the dynamics of evolving networks is a central question in many applied areas such as epidemiology, vaccination planning, anti-virus design, management of human resources, and viral marketing. A relevant clustering of the data often is needed to design informative representations of massive data sets. Algorithmic approaches have yielded useful insights on real networks such as the social interaction networks of zebras [22].

* This work was partially supported by the ANR-2010-BLAN-0204 Magnum and ANR-12-BS02-005 RDAM grants.

The dynamics of real-life evolving networks, however, are not yet well understood, partly because it is difficult to observe and analyze such large, sparsely connected networks over time. Some basic mechanisms such as preferential attachment and copy/paste have been observed, but more specific structures remain to be discovered. In this article, we propose a new formulation of the facility location problem adapted to these evolving networks. We show that, in many realistic situations, solutions that are stable over time match the ground truth more closely than those obtained by independent optimization with respect to each snapshot of the network.

The problem. We focus on a generalized facility location problem where clients are moving in some metric space over time. We look for a set of *open* facilities (also called centers) and a dynamic many-to-one assignment of clients to open facilities that minimizes the sum of three costs, of which the first two are inherited from the classical facility location problem. The *distance cost* is the sum over each (client,timestep) pair of the distance from the client to its assigned facility at that timestep. This cost tends to ensure that assigned facilities are representative with respect to position. The *opening cost* is linear in the number of facilities. This cost tends to ensure that only the most meaningful facilities are open. The new cost, *switching*, is linear in the number of (client,timestep) pairs where the client is assigned to a different facility at the next timestep. This cost tends to ensure that clients switch facilities only in response to significant and lasting changes in position. We argue that, in many realistic situations, the switching cost makes solutions close to the ground truth relatively more attractive (see Section 2.1).

Related work. The facility location problem has been studied extensively in the offline, online, and incremental settings [12]. The offline setting was a case study accompanying the development of approximation techniques: primal-dual and dual fitting methods and local search, for example. A series of papers [20,16,13,2,5,3,15] obtained almost matching upper and lower bounds on the polynomially achievable approximation ratio: $\Theta(\log n)$ in general and $[1.463, 1.488]$ in the metric case, where the specified distances satisfy the triangle inequality.

The online setting, where clients arrive over time and the algorithm gradually opens more and more facilities to serve them, was addressed first by [17], which obtained the asymptotically tight bound $\Theta(\log n/\log\log n)$ on the competitive ratio of the best online algorithm. Subsequent work considered the special case where the clients are drawn from some distribution [1] and other special cases [11]. Since many clustering applications benefit from the flexibility to change the solution over time, incremental settings also have been studied. Such variants may allow better (i.e., constant) competitive ratios, e.g., the metric case with streaming constraints [10] and the Euclidean metric setting where facilities may be moved as new clients arrive [8]. We also mention the related clustering problem in which clusters may be merged but not split [4].

Our setting differs from previous dynamic settings because the distances between clients and facilities may vary over time and because it is desirable to achieve a trade-off between the *stability* of the solution – the assignment should be modified slowly – and its *adaptability* – the assignment should be modified if the distances change significantly. Given the existence of experiments such as [21], we assume access to the whole evolution of the network ahead of time. We show that constructing an independent optimal solution for each snapshot of the network yields results that, in a large variety of realistic situations, are not only unstable (and thus arbitrarily bad according to our objective) but also undesirable with respect to network dynamics analysis.

As far as we know, settings where the distances between locations vary over time are still largely unexplored.

Our results. After defining the problem formally in Section 2.1 and giving examples showing the benefits that one can expect from solving this problem in the context of metrics evolving over time, we give in Section 2.3 an $O(\log nT)$-approximation algorithm for this problem, where n is the number of clients and T is the number of timesteps.

Theorem 1 (Fixed opening cost) *For the dynamic facility location problem with fixed opening cost, there exists a polynomial-time randomized algorithm that, on all inputs, with probability at least* $1/4$*, outputs a solution satisfying*

$$\mathit{cost} \leqslant 8\log(2nT)\cdot \mathrm{LP} \leqslant 8\log(2nT)\cdot \mathrm{OPT},$$

where OPT *is the cost of an optimal solution and* LP *is the value of LP* (1)*, defined at the end of Section 2.1.*

Through repetition, running the algorithm t times and taking the best of the t solutions constructed, the probability $1/4$ can be improved to $1-(3/4)^t$. The constant 8 can be improved as well.

We show in Section 2.4 that this approximation ratio is asymptotically optimal, even for a very special case.

Theorem 2 (Hardness for fixed opening cost) *Unless P = NP, for the dynamic facility location problem with fixed opening cost, there is no* $o(\log T)$*-approximation.*

The lower bound holds even for the metric case with one client and two locations. This new problem differs significantly from the classic facility location problem, which admits no $o(\log n)$-approximation for nonmetric distances but can be 1.488-approximated when the distances satisfy the triangle inequality [15]. In Section 3, we show how to extend our approximation algorithm to the setting where facilities can be opened and closed at each timestep. The opening cost in this setting is equal to f times the number of (facility,timestep) pairs such that the facility is open at that timestep.

Theorem 3 (Hourly opening cost) *For the dynamic facility location problem with hourly opening cost, there exists a polynomial-time randomized algorithm that, on all inputs, with probability at least* $1/4$*, outputs a solution satisfying*

$$\textit{cost} \leqslant 8\log(2nT) \cdot \mathrm{LP} \leqslant 8\log(2nT) \cdot \mathrm{OPT},$$

where OPT *is the cost of an optimal solution and* LP *is the value of LP* (2)*, defined at the end of Section 3.1.*

Again, through repetition, running the algorithm t times and taking the best of the t solutions constructed, the probability $1/4$ can be improved to $1-(3/4)^t$. The constant 8 can be improved as well. This article concludes with several open questions and possible extensions of this work.

2 Facility Location in Evolving Metrics

2.1 Definition

We denote by $[n] = \{1, \ldots, n\}$ the subset of integers from 1 to n inclusive.

Dynamic facility location problem with fixed opening cost. We are given a set F of m *facilities* and a set C of n *clients* together with a finite sequence of distances $(d_t)_{t\in[T]}$ over $F \times C$, a nonnegative *facility opening cost* f, and a nonnegative *client switching cost* g. The goal is to output a subset $A \subseteq F$ of *open* facilities and, for each timestep $t \in [T]$, an assignment $\phi_t : C \to A$ of clients to open facilities so as to minimize

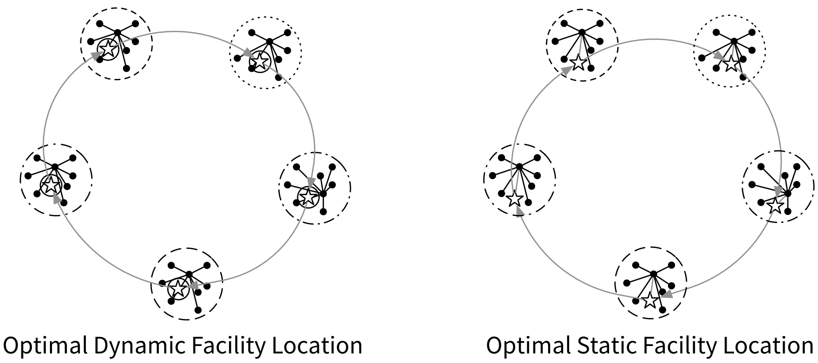
$$f \cdot |A| + \sum_{t\in[T], j\in C} d_t(\phi_t(j), j) + g \cdot \sum_{t\in[T-1], j\in C} \mathbb{1}\{\phi_t(j) \neq \phi_{t+1}(j)\},$$

namely, the sum of the *opening cost* (f for each open facility), the *distance cost* (the sum over each (client,timestep) pair from the client to its assigned facility at that timestep), and the *switching cost* (g for each (client,timestep) pair where the client is assigned to a different facility at the next timestep).

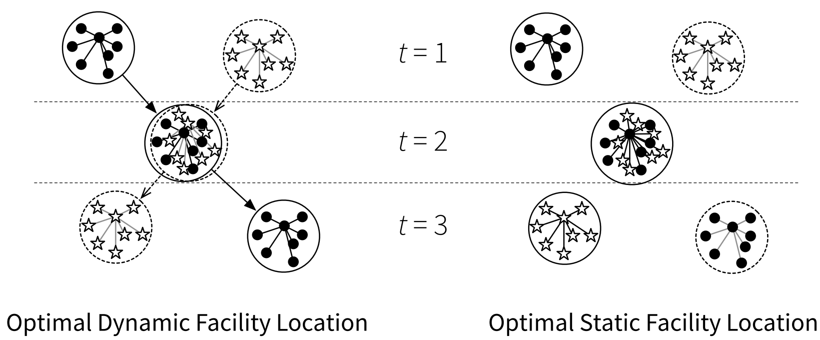
Examples. The two examples in Figure 1 show how facility location in the dynamic setting is quite different from facility location in the static setting and yields more desirable partitions of the clients. In both examples, a facility can be opened at every client (so that electing a facility consists of electing a representative for every significantly different behavior).

In example 1(a), we see a classroom with students split into five groups and a teacher moving from group to group in cyclic order. When the number of students is large, static facility location isolates the five groups and moves the teacher from one group to the next between snapshots. Dynamic facility location isolates every group of students and puts the teacher in a sixth group.

In example 1(b), we see two groups of people passing through each other, on a street for instance. Static facility location outputs first the two groups, then a



(a) The classroom: one teacher cycling between 5 groups of students.



(b) Two groups crossing.

Fig. 1. Dynamic versus static facility location.

single group, then two groups again. Dynamic facility location, however, keeps the same groups for the whole time period, with the same representatives.

Assuming in the first example that the distances between individuals are very small and in the second that they are very large, the ratio of the (dynamic) cost between the dynamic solution and the sequence of static solutions can be made arbitrarily large, because the switching cost grows for the sequence of static solutions as $\Omega(T)$ and $\Omega(n)$ respectively.

Fact 4 *The (dynamic) cost of a sequence of optimal static facility location solutions for each snapshot can be larger than the cost of an optimal dynamic facility location solution by a factor $\Omega(T+n)$.*

A linear relaxation. For an integer programming formulation, we define indicator 0-1 variables y_i, x_{ij}^t, z_{ij}^t for $t \in [T]$ and $i \in F$ and $j \in C$. We let $y_i = 1$ if and only if facility i is open; $x_{ij}^t = 1$ if and only if client j is assigned to facility i at timestep t; and $z_{ij}^t = 1$ if and only if client j is assigned to facility i at timestep t

but not at timestep $t+1$. The dynamic facility location problem is equivalent to finding an integer solution to the following linear programming relaxation.

$$\begin{cases} \text{Minimize } f \cdot \sum_{i \in F} y_i + \sum_{t \in [T], i \in F, j \in C} x_{ij}^t \cdot d_t(i,j) + g \cdot \sum_{t \in [T-1],\, i \in F,\, j \in C} z_{ij}^t \\ \text{subject to} \quad (\forall t \in [T],\ i \in F,\ j \in C) \ \ x_{ij}^t \leqslant y_i \\ \qquad (\forall t \in [T],\ j \in C) \ \ \sum_{i \in F} x_{ij}^t = 1 \\ \qquad (\forall t \in [T-1],\ i \in F,\ j \in C) \ \ z_{ij}^t \geqslant x_{ij}^t - x_{ij}^{t+1} \\ \qquad (\forall t \in [T],\ i \in F,\ j \in C) \ \ y_i, x_{ij}^t, z_{ij}^t \geqslant 0 \end{cases} \tag{1}$$

2.2 Facts about Probability

We use the following two facts and some properties of exponential distributions.

Fact 5 *Let $X \geqslant 0$ be a random variable and B be an event, not necessarily independent. We have $E[X \mid B] \leqslant E[X]/\Pr B$.*

Proof. Let $\overline{B}$ be the complement of B. We have

$$E[X \mid B] \leqslant E[X \mid B] + E[X \mid \overline{B}] \Pr \overline{B} / \Pr B = E[X]/\Pr B. \qquad \square$$

Fact 6 (Markov's inequality) *Let $X \geqslant 0$ be a random variable. For every $x > 0$, we have $\Pr\{X > x\} \leqslant E[X]/x$.*

A random variable X is *exponentially distributed* with rate λ if and only if, for every $x \geqslant 0$, it satisfies $\Pr\{X > x\} = e^{-\lambda x}$.

Fact 7 *If X is exponentially distributed with rate λ, then, for every $c > 0$, the distribution of X/c is exponential with rate $c\lambda$.*

Fact 8 *Let $(X_i)_{i \in F}$ be a sequence of independent random variables, where X_i is exponentially distributed with rate λ_i. Then $\min_{i \in F} X_i$ is exponentially distributed with rate $\sum_{i \in F} \lambda_i$, and the argument of the minimum is i with probability $\lambda_i / \sum_{k \in F} \lambda_k$.*

Proof. Indeed, $\Pr\{\min_{i \in F} X_i > x\} = \prod_{i \in F} \Pr\{X_i > x\} = e^{-\sum_{i \in F} \lambda_i x}$. As for the second claim,

$$\begin{aligned} \Pr\{\arg\min_{k \in F} X_k = i\} &= \int_{x=0}^{\infty} \Pr\{(\forall k \neq i)\ X_k > x\} \cdot \Pr\{X_i \in [x, x+dx]\} \\ &= \int_{x=0}^{\infty} e^{-\sum_{k \neq i} \lambda_k x} \cdot \lambda_i e^{-\lambda_i x} dx = \lambda_i \Big/ \sum_{k \in F} \lambda_k. \end{aligned} \qquad \square$$

2.3 Approximation Algorithm

In order to determine a solution, we need to (1) decide which facilities to open, (2) decide when each client switches from one facility to another, and (3) decide which facility to connect each client to between switches. After computing an optimal (fractional) solution (x, y, z) to LP (1), Algorithm 1 proceeds as follows. Decision (1) is made by sampling the facilities according to $(y_i)_i$ approximately $O(\log nT)$ times.As we will show, this ensures that every client selects a sampled facility with high probability.

Regarding decision (2), since $\sum_i x^t_{ij} = 1$, one can view $(x^t_{ij})_i$ as the desired distribution for the facility assigned to client j at timestep t. The subroutine Algorithm 2 partitions time, independently for each client j, into intervals during which the distribution $(x^t_{ij})_i$ remains stable enough, i.e., the distributions $(x^t_{ij})_i$ *share a large enough common probability mass* during each time interval of the partition. The common probability mass of the distributions $(x^t_{ij})_i$ during a time interval U is defined as the sum over all facilities i of the minimum probability $\hat{x}^U_{ij} = \min_{t\in U} x^t_{ij}$ of assigning client j to i over U. The rule defining the partition is that each interval (except the last one) is maximal subject to the constraint that the common probability mass is at least $1/2$. This ensures two key properties. First, the distributions $(x^t_{ij})_i$ for $t \in U$ are close enough to each other to be compatible and also, due to the first LP constraint, close enough to $(y_i)_i$ to match the sampling of the facilities. Second, the distributions are deemed to have changed too much when the x^t_{ij}s have had a combined decrease of at least $1/2$, which implies by the third LP constraint that the corresponding z^t_{ij}s sum to at least $1/2$, covering the cost of switching to another facility.

Decision (3) is made simply by assigning each client to the most likely of its preferred facilities to be open.

We propose two versions of the algorithm. The first assigns clients to open facilities via an optimal dynamic program, while the second uses the intuitive strategy described in Algorithm 2. We analyze the latter, as its approximation ratio is no worse than that of the former.

Theorem 1 states that Algorithm 1 outputs an $O(\log nT)$-approximation with positive constant probability. In the next section, we will show that this is asymptotically optimal (unless $P = NP$).

Proof (Theorem 1). Note that Algorithm 2 may produce an assignment that is not feasible. We bound the expected cost without conditioning on feasibility, bound the probability of feasibility, and finish by applying Fact 5 and Markov's inequality.

Algorithm 1 Fixed opening cost

- Solve the linear program LP (1) to obtain an optimal (fractional) solution (x, y, z).
- Choose the open facilities A randomly as follows. For each facility i, choose Y_i having exponential distribution with rate $2\log(2nT)$. Let $A = \{i \in F : Y_i \leqslant y_i\}$.
- With a dynamic program, determine how to assign optimally clients to facilities in A. Alternatively, for the purposes of analysis, use Algorithm 2.

Algorithm 2 Intuition-driven assignment of clients to facilities

for each client j **do**

- Partition time greedily into ℓ_j intervals $[t_k^j, t_{k+1}^j)$ where ℓ_j and $(t_k^j)_{k\in[\ell_j+1]}$ are defined as follows: $t_1^j = 1$, and t_{k+1}^j is defined inductively as the greatest $t \in (t_k^j, T+1]$ such that $\sum_{i\in F}\Big(\min_{t_k^j \leqslant u < t} x_{ij}^u\Big) \geqslant 1/2$. Let $t_{\ell_j+1}^j = T+1$.
- For each time interval $U = [t_k^j, t_{k+1}^j)$, assign client j to argument of $\min_{i\in F}(Y_i/\hat{x}_{ij}^U)$, where $\hat{x}_{ij}^U = \min_{u\in U} x_{ij}^u$.

end for

The unconditional expected facility opening cost is

$$f \cdot \sum_{i\in F}(1 - e^{-2y_i \log(2nT)}) \leqslant (2\log(2nT))f \cdot \sum_{i\in F} y_i$$

by the well known inequality $1+x \leqslant e^x$. The right-hand side is $2\log(2nT)$ times the corresponding term in the LP objective.

To analyze the unconditional expected distance cost, we define, for each client j, all of its time intervals U, and all $t \in U$, a fictitious independent event B_j^t such that $\Pr B_j^t = \sum_{k\in F}\hat{x}_{kj}^U \in [1/2, 1]$ by the LP and the definition of U.[5] We use this fictitious event B_j^t to define a random variable $I_j^t \in F$ by letting $\Pr\{I_j^t = i \mid B_j^t\} = \hat{x}_{ij}^U / \sum_{k\in F}\hat{x}_{kj}^U$ and $\Pr\{I_j^t = i \mid \overline{B}_j^t\} = (x_{ij}^t - \hat{x}_{ij}^U)/\sum_{k\in F}(x_{kj}^t - \hat{x}_{kj}^U)$, where $\overline{B}_j^t$ is the complement of B_j^t. Note that $\Pr \overline{B}_j^t = \sum_{k\in F}(x_{kj}^t - \hat{x}_{kj}^U)$ since $\sum_{k\in F} x_{kj}^t = 1$ by LP (1). The unconditional distribution of I_j^t thus is described by $\Pr\{I_j^t = i\} = x_{ij}^t$, so the expected distance from j to I_j^t is $E[d_t(I_j^t, j)] = \sum_{i\in F} x_{ij}^t \cdot d_t(i,j)$. Since $\arg\min_{i\in F}(Y_i/\hat{x}_{ij}^U)$ is i with probability $\hat{x}_{ij}^U/\sum_{k\in F}\hat{x}_{kj}^U$ by Fact 8, the actual assignment of Algorithm 2 is made according to the conditional distribution of I_j^t given B_j^t, so by applying Fact 5 and summing, the total unconditional expected distance cost is at most

$$2 \cdot \sum_{t\in[T], i\in F, j\in C} x_{ij}^t \cdot d_t(i,j),$$

which is twice the corresponding term in the LP objective.

To bound the switching cost, which is deterministic, we prove that, for each client j and all of its time intervals U except the last one,

$$\sum_{t\in U, i\in F} z_{ij}^t > 1/2.$$

[5] More concretely, let B_j^t be an event corresponding to the outcome Heads of an independent biased coin flip that results a priori in Heads with probability $\sum_{i\in F}\hat{x}_{ij}^U$. This event represents our ability to sample from the common probability mass of the distributions $(x_{ij}^t)_i$ for $t \in U$.

Each client switches only after its non-last intervals. Since each variable z^t_{ij} appears in exactly one sum, the total switching cost is bounded above by

$$2g \cdot \sum_{t \in [T-1], i \in F, j \in C} z^t_{ij},$$

which is twice the corresponding term in the LP objective.

The z-variables measure decreases in the corresponding x-variables. Specifically, for every $t_1 \leqslant t_2$, the LP inequalities telescope to yield $x^{t_1}_{ij} - x^{t_2}_{ij} \leqslant \sum_{u \in [t_1, t_2)} z^u_{ij}$. By letting t_1 be the first time in $U = [t_1, t_3)$ and t_2 be the argument of the minimum $\min_{u \in [t_1, t_3]} x^u_{ij}$, whose domain is $U \cup \{t_3\}$, we sum to obtain the inequality

$$1/2 = 1 - 1/2 < 1 - \sum_{i \in F} \min_{u \in [t_1, t_3]} x^u_{ij} = \sum_{i \in F} (x^{t_1}_{ij} - \min_{u \in [t_1, t_3]} x^u_{ij}) \leqslant \sum_{u \in U, i \in F} z^u_{ij},$$

where the first inequality is a consequence of defining U maximally.

As the next to last step, we bound the probability that every client is assigned to an open facility. Recall that, at each timestep t, Algorithm 2 assigns each client j to the argument i^* of the minimum $\min_{i \in F}(Y_i / \hat{x}^U_{ij})$, where $U \ni t$ is the corresponding interval for j. This facility is open if and only if $Y_{i^*} \leqslant y_{i^*}$. Since $\hat{x}^U_{ij} \leqslant x^t_{ij} \leqslant y_i$,

$$\Pr\{Y_{i^*} \leqslant y_{i^*}\} \geqslant \Pr\{Y_{i^*} \leqslant \hat{x}^U_{i^* j}\} = \Pr\{\min_{i \in F}(Y_i / \hat{x}^U_{ij}) \leqslant 1\}.$$

The quantity $\min_{i \in F}(Y_i / \hat{x}^U_{ij})$ is exponentially distributed with rate $2 \log(2nT) \cdot \sum_{i \in F} \hat{x}^U_{ij} \geqslant \log(2nT)$ since $\sum_{i \in F} \hat{x}^U_{ij} \geqslant 1/2$, so $\Pr\{Y_{i^*} \leqslant y_{i^*}\} \geqslant 1 - 1/(2nT)$. By a union bound over all clients and timesteps, the probability of a feasible assignment is at least $1/2$.

In conclusion, we observe that the unconditional expected cost is a $2 \log(2nT)$-approximation of the LP objective, and the probability of a feasible assignment is at least $1/2$. By Fact 5, the conditional expected cost given feasibility is a $4 \log(2nT)$-approximation. By Markov's inequality, with probability at least $1/2 \cdot 1/2 = 1/4$, the output is a feasible $8 \log(2nT)$-approximation. □

2.4 Hardness of Approximation

Proof (Theorem 2). We exhibit an objective-preserving reduction from the set cover problem. Fix an instance of set cover with T elements and m sets. We define the following instance of dynamic facility location. There is one timestep t for each element of the set cover instance, one facility i for each set of the set cover instance, and a single client. We set $g = 0$ (i.e., g is small enough with respect to f and $1/n$ and $1/T$). Assume that the only possible positions for the client and facilities are two locations a and b at distance ∞ (i.e., large enough) from each other (note that this metric satisfies the triangle inequality). At every timestep t, the client's position is location a. For each set i of the set cover instance, the position of the corresponding facility is location a if set i contains element t and location b otherwise.

Since the distance between the two locations is infinite, a solution for our instance of dynamic facility location has finite cost if and only if, at every timestep, some open facility has position a, i.e., the set of open facilities corresponds to a cover. The cost of such a solution is f times the number of open facilities. We conclude that the $\Omega(\ln T)$-inapproximability result for set cover with T elements [7] implies the same inapproximability result for our problem. $\square$

3 Hourly Opening Cost

3.1 Dynamic Facility Location with Hourly Opening Cost

We now focus on a variant of the problem studied in the previous section, where each facility may be open or closed independently at each timestep and where the opening cost f is paid for each (facility,timestep) pair where the facility is open at that timestep. In other words, the cost of a facility is not its construction cost but its rental cost.

Dynamic facility location problem with hourly *opening cost.* We are given a set F of m facilities and a set C of n clients together with a finite sequence of distances $(d_t)_{t\in[T]}$ over $F \times C$ and two nonnegative values f and g. The goal is to output a sequence of subsets $A_t \subseteq F$ of facilities and, for each timestep $t \in [T]$, an assignment $\phi_t : C \to A_t$ of clients to facilities so as to minimize

$$f \cdot \sum_{t\in[T]} |A_t| + \sum_{t\in[T],j\in C} d_t(\phi_t(j), j) + g \cdot \sum_{t\in[T-1],j\in C} \mathbb{1}\{\phi_t(j) \neq \phi_{t+1}(j)\}.$$

Linear relaxation. LP (1) can easily be adapted to this variant, with new variables y_i^t replacing y_i. The interpretation of y_i^t is that it equals 1 if and only if facility i is open at timestep t.

$$\begin{cases} \text{Minimize } f \displaystyle\sum_{t\in[T],i\in A} y_i^t + \sum_{t\in[T],i\in F,j\in C} x_{ij}^t \cdot d_t(i,j) + g \sum_{t\in[T-1],i\in F,j\in C} z_{ij}^t \\ \text{subject to} \quad (\forall t \in [T],\ i \in F,\ j \in C) \ \ x_{ij}^t \leqslant y_i^t \\ \qquad (\forall t \in [T],\ j \in C) \ \ \displaystyle\sum_{i\in F} x_{ij}^t = 1 \\ \qquad (\forall t \in [T-1],\ i \in F,\ j \in C) \ \ z_{ij}^t \geqslant x_{ij}^t - x_{ij}^{t+1} \\ \qquad (\forall t \in [T],\ i \in F,\ j \in C) \ \ y_i^t, x_{ij}^t, z_{ij}^t \geqslant 0 \end{cases} \tag{2}$$

3.2 Approximation Algorithm

Our algorithm for hourly costs, Algorithm 3, is very similar to Algorithm 1, for fixed costs. The key idea is to choose the random variables Y_i only once to ensure

Algorithm 3 Hourly opening cost

- Solve the linear program LP (2) to obtain an optimal (fractional) solution (x, y, z).
- For each timestep t, choose the open facilities A_t randomly as follows. **Once**, for each facility i, choose Y_i having exponential distribution with rate $2\log(2nT)$. Let $A_t = \{i \in F : Y_i \leqslant y_i^t\}$.
- With a dynamic program, determine how optimally to assign clients to facilities in A_t. Alternatively, for the purposes of analysis, use Algorithm 2 (as done in Algorithm 1).

that the set of open facilities is stable. The statements of correctness, Theorems 1 and 3, are proved by exactly the same arguments. The only difference is that, in order for the facility $i^* = \arg\min_{i\in F} Y_i/\hat{x}_{ij}^U$ to be open to client j throughout its time interval U, we need $Y_{i^*} \leqslant y_{i^*}^t$ for all $t \in U$. For each choice of j and U, this family of inequalities is satisfied with probability at least $1 - 1/(2nT)$, the same bound as before, since the fact that $x_{ij}^U \leqslant x_{ij}^t \leqslant y_i^t$ for all $t \in U$ and all $i \in F$ implies as before that $\Pr\{(\forall t \in U)\ Y_{i^*} \leqslant y_{i^*}^t\} \geqslant \Pr\{Y_{i^*} \leqslant \hat{x}_{i^*j}^U\} = \Pr\{\min_{i\in F} Y_i/\hat{x}_{ij}^U \leqslant 1\} \geqslant 1 - 1/(2nT)$. The rest of the proof requires no change.

4 Conclusion and Open Questions

Algorithm 1 applies even if the distances between clients and facilities do not satisfy the triangle inequality, and it extends directly to nonuniform opening costs as well as arrival and departure dates for clients. It is striking that instances with distances satisfying the triangle inequality are not easier in the dynamic setting as opposed to the classic static setting (the approximation ratio $\Theta(\log nT)$ of Algorithm 1 is tight in both dynamic cases). Algorithm 3 also extends directly to the setting of opening costs that are nonuniform in time. The last section naturally raises the question of whether there exists an $\omega(1)$-hardness result / $O(1)$-approximation algorithm for the general hourly opening cost case.

We believe that our dynamic setting should be helpful in designing better *static* representations of dynamic graphs (e.g., two dimensional flowcharts of clients navigating between facilities over time). Another natural extension of our work is to study other objective functions for the distance cost, such as the sum of the diameters of the reported clusters over all timesteps (i.e., the sum of the distance of the farthest client assigned to each facility, see, e.g., [6] for a static formulation). As it turns out, the optimal dynamic solutions with respect to this objective tend to exhibit very intriguing behaviors, even in the simplest case of clients moving along a fixed line [9].

References

1. A. Anagnostopoulos, R. Bent, E. Upfal, and P. Van Hentenryck. A simple and deterministic competitive algorithm for online facility location. *Information and Computation*, 194:175–202, 2004.

2. V. Arya, N. Garg, R. Khandekar, A. Meyerson, K. Munagala, and V. Pandit. Local search heuristics for k-median and facility location problems. *SIAM J. on Computing*, 33(3):544–562, 2004.
3. J. Byrka and K. Aardal. An optimal bifactor approximation algorithm for the metric uncapacitated facility location problem. *SIAM J. on Computing*, 39(6):2212–2231, 2010.
4. M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In *STOC*, pages 626–635, 1997.
5. M. Charikar and S. Guha. Improved combinatorial algorithms for facility location problems. *SIAM J. on Computing*, 34(4):803–824, 2005.
6. M. Charikar and R. Panigrahy. Clustering to minimize the sum of cluster diameters. In *STOC*, pages 1–10, 2001.
7. Irit Dinur and David Steurer. Analytical approach to parallel repetition. In *STOC 2014, arXiv:1305.1979*, 2014.
8. G. Divéki and C. Imreh. Online facility location with facility movements. *Central European J. of Operations Research*, 19(2):191–200, 2010.
9. Cristina G. Fernandes, Marcio I. Oshiro, and Nicolas Schabanel. Dynamic clustering of evolving networks: some results on the line. In *AlgoTel*, 2013. Url: hal-00818985 (4 pages).
10. D. Fotakis. Incremental algorithms for facility location and k-median. *Theoretical Computer Science*, 361(2-3):275–313, 2006.
11. D. Fotakis. On the competitive ratio for online facility location. *Algorithmica*, 50(1):1–57, 2008.
12. D. Fotakis. Online and incremental algorithms for facility location. *SIGACT News*, 42(1):97–131, 2011.
13. K. Jain, M. Mahdian, E. Markakis, A. Saberi, and V. Vazirani. Greedy facility location algorithms analyzed using dual fitting with factor-revealing lp. *J. ACM*, 50(6):795–824, 2003.
14. J. M. Kleinberg. The small-world phenomenon and decentralized search. *SIAM News*, 37(3), 2004.
15. Shi Li. A 1.488-approximation algorithm for the uncapacitated facility location problem. In *ICALP*, pages 77–88, 2011.
16. M. Mahdian, Y. Ye, and J. Zhang. Improved approximation algorithms for metric facility location problems. In *APPROX*, volume LNCS 2462, pages 229–242, 2002.
17. A. Meyerson. Online facility location. In *FOCS*, volume 42, pages 426–431, 2001.
18. M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
19. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, 2001.
20. D. B. Shmoys, E. Tardos, , and K. I. Aardal. Approximation algorithms for facility location problems. In *STOC*, volume 29, pages 265–274, 1997.
21. J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, and P. Vanhems. High-resolution measurements of face-to-face contact patterns in a primary school. *PLoS ONE*, 6(8):e23176, 2011.
22. C. Tantipathananandh, T. Y. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD*, pages 717–726, 2007.